# Computational Analysis of Host/Pathogen Coevolution


Samuel Campbell

St. Catherine's College

University of Oxford

Supervisor: Professor Astrid Iversen

A thesis submitted for the degree of

*Doctor of Philosophy in Zoology*

Michaelmas 2019

# Acknowledgements

I am extremely grateful for the opportunity to work with my supervisor, Professor Astrid Iversen. Her support and guidance has been invaluable both in my personal and professional development, and I count myself lucky for the excellent work environment she fosters. Dr. Nico Kist, a former DPhil of Astrid's, was instrumental in my developing my coding abilities essentially from nothing and it is not an overstatement to say that I could not have completed this project without him. Our collaborators Professor Daniel Lunn and Dr. Ben Lambert were both enormously helpful. Dan was patient in teaching me many of the mathematical analyses used in this thesis. His passing is a tragedy and he will be missed. Ben was generous with his time and assistance, for which I am tremendously thankful.

The Oxford Zoology Department and Weatherall Institute of Molecular Medicine have been my home (in various buildings due to the closure of the Tinbergen!) for the past four years and I'm thankful for them welcoming me in. I cannot thank the Zoology Graduate Administrator, Heather Green, enough. She went above and beyond dealing with my often convoluted paperwork. No amount of chocolate or tennis will ever repay her. My friends here in Oxford have made my adventures here in the East an unforgettable experience, and helped ground me in the world outside the lab.

Holly Bunce has been a constant source of joy, support, motivation, and making sure I eat and sleep. I am profoundly indebted to her. My family has been beyond supportive of my far flung journey, and I cannot ever adequately express my thanks to them.

# Computational Analysis of Host/Pathogen Coevolution

Samuel Campbell

St. Catherine's College
University of Oxford

*A thesis submitted for the degree of*
*Doctor of Philosophy in Zoology*

Michaelmas 2019

**Abstract**

In this thesis, I present my research on host/pathogen coevolution in several different systems and on different time scales. The thesis comprises multiple parts and shows how advances in computational methods can be applied to a broad range of questions, from basic research to issues of immediate clinical relevance.

Chapter 1 examines this host pathogen coevolution in a nonhuman system, looking at giant viruses, satellite viruses, related transposable elements found in eukaryotes and giant viruses, and individual 'jumping' homing endonuclease genes (HEGs) found in all three. Unfortunately, due to enormous amounts of time over which some of the elements involved have diverged, it is difficult to draw reliable conclusions here. However, the novel identification of HEGs in satellite and giant viruses shows a likely solution to the question of the mechanism for the horizontal gene transfer of HEGs between dissimilar phyla.

Chapter 2 focuses on the coevolution of humans with pathogens. This can be seen in the influence of archaic hominid immune genes on the modern immune system, and what appears to be the introgression of archaic hominids' granular adaptation to local pathogens when modern humans expanded out of Africa. The fact that these archaic HLAs have retained this regional granular adaptation for tens of thousands of years suggests some sort of survival advantage. While previous work has shown that there is different selective pressures acting on different class I HLAs, this chapter shows that this relationship is retained when looking exclusively at archaic derived HLAs. Selective pressures acting on HLA-A but not HLA-B have not previously been described, and Chapter 2 identifies the suppression of endogenous retroviruses as possibly being one such selective pressure.

Chapter 3 examines the granular adaptation of HIV-1M to humans in Sub-Saharan Africa by looking at correlations between human genetic diversity and HIV diversity. Unfortunately, widespread genotyping of HIV patients in Sub-Saharan Africa is not available, but linguistic diversity has previously been specifically shown to correlate with human genetic diversity in Sub-Saharan Africa, providing a useful proxy with which to make comparisons. My results show that linguistic diversity, and by extension human genetic diversity, does indeed correlate with both HIV subtype and amino acid sequence diversity in the region. This runs contrary to the dominant paradigm that HIV subtypes are artefacts of founder effects and may explain some of the difficulties in creating an HIV vaccine despite over 30 years of research on the subject.

# Table of Contents

# Abbreviations

| | |
|---|---|
| AA | Amino Acid |
| AAV | *Adeno-Associated Virus* |
| ACMV | *Acanthamoeba Castellanii Mamavirus* |
| AFD | Allele Frequency Database |
| AIC | Akaike Information Criterion |
| APC | Antigen Presenting Cell |
| APMV | *Acanthamoeba Polyphaga Mimivirus* |
| bp | Base Pair |
| CDD | Conserved Domain Database |
| CMV | Cytomegalovirus |
| CroV | *Cafeteria roenbergensis Virus* |
| CTL | Cytotoxic T Lymphocyte |
| dbMEM | Distance-based Moran's Eigenvector Maps |
| EBV | Epstein-Barr Virus |
| GAM | Generalised Additive Model |
| HCV | Hepatitis C Virus |
| HEG | Homing Endonuclease Gene |
| HPV | Human Papillomavirus |
| HIV | Human Immunodeficiency Virus |
| HSV1 | Herpes Simplex Virus 1 |
| HTLV-1 | Human T-Lymphotropic Virus 1 |
| ICTV | International Committee on Taxonomy of Viruses |
| Ig | Immunoglobulin |
| IVA | Influenza A Virus |
| kbp | Kilobase Pair |
| KSHV | Kaposi's Sarcoma-Associated Herpesvirus |
| KIR | killer-cell immunoglobulin-like receptor |

| | |
|---|---|
| LD | Linkage Disequilibrium |
| Mbp | Megabase Pair |
| MCMC | Markov chain Monte Carlo |
| MDS | Multidimensional Scaling (PCoA) |
| MHC | Major Histocompatibility Complex |
| MLE | Maximum Likelihood Estimate |
| MNP | Modern Native Populations |
| NA | Nucleic Acid |
| NCBI | National Center For Biotechnology Information (US) |
| NCLDV | Nucleocytoplasmic Large DNA Virus |
| NK | Natural Killer |
| NMDS | Non-Metric Multi-Dimensional Scaling |
| OLPV | *Organic Lake Phycodnavirus* |
| ORF | Open Reading Frame |
| PBMC | Primary Blood Mononuclear Cells |
| PCA | Principle Component Analysis |
| PCoA | Principle Coordinate Analysis (MDS) |
| PLE | Polinton-Like Element |
| POLB | Polymerase B |
| RNAP | RNA Polymerase |
| SARS-CoV-2 | Sudden Acute Respiratory Syndrome Corona Virus-2 |
| STI | Sexually Transmitted Infection |
| SMPLE | *Symbiodinium minutum* PLE |
| SNP | Single Nucleotide Polymorphism |
| SVG | Scalable Vector Graphic |
| TCR | T-Cell Receptor |
| TIR | Terminal Inverted Repeat |
| TSA | Transcriptome Shotgun Assembly |
| VIF | Variance Inflation Factor |

| | |
|---|---|
| VIP | Virus Interacting Protein |
| VLE | Virophage Like Element |
| YBP | Years Before Present |
| YSLV | Yellowstone Lake Virophage |

Regional Abbreviations from Chapter 2

| | |
|---|---|
| AF | Africa |
| AU | Australia |
| EU | Europe |
| NAF | North Africa |
| NAM | North America |
| NEA | North East Asia, |
| OC | Oceania |
| SA | South Asia |
| SAM | South America |
| SEA | South East Asia |
| WA | Western Asia |

# Introduction

The scientific method involves extracting useful information from data and has for centuries allowed humanity to slowly chisel away at the scourge of disease with discoveries such as antibiotics, vaccines, and the aetiologies of cancers. Both the amount of data that we have access to and the rate at which we accumulate it have been steadily increasing to almost unmanageable proportions. As of October 2019, GenBank (and the European Nucleotide Archive and the DNA Data Bank of Japan via daily data exchanges) contains 386,197,018,538 bases in 216,763,706 sequences and an additional 5,985,250,251,028 bases and 1,097,629,174 sequences in Whole Genome Shotgun data. The number of sequences is growing at an accelerating rate, doubling in size roughly every 18 months (Benson et al., 2009, 2008; GenBank, 2019). On top of this is all the metadata associated with these sequences, and this is only one type of data repository. There exist many more such as UniProt and the allele frequency database (AFD). This huge amount of data is useful for both basic research and for understanding, modelling, and reacting effectively to ongoing pandemics such as those caused by human immunodeficiency virus (HIV) and Sudden Acute Respiratory Syndrome Corona Virus-2 (SARS-CoV-2). Extracting useful information from this deluge is not always a straightforward procedure. Computers allow us to process this flood of data and make sense of it, but only if we ask the right questions, and perform the correct analyses. This thesis comprises multiple parts, examining different computational applications to a variety of questions which at their heart all involve host and pathogen coevolution, that is the two reciprocally affecting each other's evolution.

# Part 1

Chapter 1 of this thesis examines the tangled phylogenetic relationships between a probably paraphyletic grouping of viruses called giant viruses, smaller satellite viruses known as virophages, a class of transposable elements (TEs) known as either polintons or mavericks, related but apparently distinct TEs referred to as 'polinton-like' elements, mobile homing endonuclease genes (HEGs), and the eukaryotic hosts harbouring this convoluted menagerie.

## *Viruses*

Viruses are by far the most abundant biological entities on Earth. Depending on the specific environments being sampled, they are estimated as anywhere from five to a thousand times more numerous than prokaryotes (Suttle, 2007; Williamson et al., 2013). Multiple analyses have indicated that we have only characterised a miniscule fraction of the global virome (Brum et al., 2015; Edwards and Rohwer, 2005; Paez-Espino et al., 2016). Viruses are found in every life supporting environment on Earth: pools of boiling acid (Diemer and Stedman, 2012; Schleper et al., 1992), alkaline hypersaline lakes (Jiang et al., 2004), pelagic marine waters (Suttle, 2007), deep sea hydrothermal vents (Wommack et al., 2004), temperate soils (Williamson et al., 2013), and 'live' virus recovered after 30,000 years in the Siberian permafrost (Legendre et al., 2014). In theory, and as far as we have ascertained to date, all cellular organisms are vulnerable to some kind of viral infection (Fuhrman, 1999).

Despite this diversity of environments and hosts, all known viruses are obligate intracellular parasites, and as such need to get their genetic material, in the form of single stranded (ss) or double stranded (ds) ribonucleic acid (RNA) or deoxyribonucleic acid (DNA), into a host cell in order to replicate. This genetic material must end up in a form functionally equivalent to the host cell's messenger RNA (mRNA), in order to be read by the host's ribosomes and translated into proteins. By convention, cellular mRNA is considered 'sense' or 'positive sense' (+), and genomes that are complimentary to this are considered 'negative sense' or 'antisense' (-), *e.g.* a (+)ssRNA genome can be directly translated by the host ribosome, whereas (-)ssRNA genome must be first converted into mRNA. Accordingly, viruses can be classified under the 'Baltimore classification system', which initially comprised six categories (Baltimore, 1971), with the seventh being added later after the identification of the hepadnavirus replication system (Summers and Mason, 1982; Temin, 1985). Some viruses use reverse transcriptase (RT) to retrotranscribe RNA to complementary DNA and are labelled as such under the system:

Class I: dsDNA viruses

Class II: ssDNA viruses

Class III: dsRNA viruses

Class IV: (+)ssRNA viruses

Class V: (−)ssRNA viruses

Class VI: (+)ssRNA-RT viruses

Class VII: dsDNA-RT viruses

When the system was initially created, ssDNA viruses were thought to be only positive sense. Baltimore was aware even then that the satellite virus adeno-associated virus (AAV) packages both (-)ssDNA and (+)ssDNA equally in separate virions (Crawford et al., 1969; Mayor et al., 1969; Rose et al., 1969). However, transcription was thought to only occur from the antisense strand of AAV (Carter et al., 1976, 1972; Carter and Rose, 1972; Rose and Koczot, 1971). Since then it has been shown that there are ssDNA viruses that are sense (M13 bacteriophage) (Specthrie et al., 1992), antisense (AAV), and ambisense (geminiviruses) (Wright et al., 1997). There is room for at least three other categories, (-)ssRNA-RT viruses, and both sense/antisense ssDNA-RT viruses, however, examples of these have yet to be found.

The type and structure of the viral genome has important implications for genome size and stability, affecting mutation rates and therefore viral evolution. RNA viruses, including class VI retroviruses, mutate at a rate of $10^{-6}$ to $10^{-4}$ substitutions per nucleotide per cell infection (s/n/c), substantially faster than the $10^{-8}$ to $10^{-6}$ s/n/c seen in their DNA virus counterparts (Sanjuan et al., 2010). There are several reasons for this. RNA is more reactive due to the presence of a hydroxyl group on the second 2' carbon of the ribose ring as opposed to a hydrogen atom in DNA (Voet and Voet, 2011). Cytosine, found in both RNA and DNA, undergoes spontaneous deamination to uracil, found only in RNA. In DNA this is recognised as an error and corrected by an array of enzymes, whereas in RNA the change is not as readily distinguishable (Alberts, 2002; Neddermann et al., 1996). Viral RNA polymerases (RNAP), excluding some host

related RNAP found in poxviruses, do not have the built in proofreading found in DNA polymerases (Sydow and Cramer, 2009). In DNA based microbes, including DNA viruses, there is a strong negative correlation between genome size and mutation rate (known as Drake's rule) (Drake, 1991). Although it is less clear cut with RNA viruses, there is some indication that Drake's rule applies to them as well (Bradwell et al., 2013; Sanjuan et al., 2010). The aforementioned hydroxyl group in RNA makes it susceptible to base-catalysed hydrolysis which cleaves RNA's sugar-phosphate backbone, breaking the molecule apart (Voet and Voet, 2011). This seems to limit the maximum size of RNA genomes to ~30 kilobase pairs (kbp), as seen in the coronaviruses including SARS-CoV-2 (Wu et al., 2020), which make use of non-RNAP proofreading enzymes to help maintain their genomes (Denison et al., 2011). Although cause and effect can be debated, the higher mutation rates seen in RNA viruses may be artefacts of their smaller genomes. Indeed, some studies have suggested that the distinction in mutation and evolutionary rates (the rate at which these mutations become fixed in a population) between DNA and RNA viruses is blurry at best and possibly non-existent (Biek et al., 2015; Duffy et al., 2008). However, there is some evidence that taking into account the apparent rate decay dynamics recovers this distinction as well as that between RT and non-RT viruses (Aiewsakun and Katzourakis, 2016).

Genome type is not the only useful classification system used to categorise viruses. Phylogenetic groupings can show the relatedness of viral families, and can be used to trace the origins of viruses, e.g. determining that SARS-CoV-2 is zoonotic in origin (Lu

et al., 2020). Host range is often extremely important and the level of specificity depends on what is being asked. A broad classification of host range might be used in examining the differences between viruses of prokaryotes and eukaryotes, while a more narrow classification might only be looking at viruses that infect humans. Morphological distinctions can also be used to classify viruses and often have important implications on treatment and spread e.g. enveloped vs. nonenveloped viruses (Firquet et al., 2015). With multiple classification systems, viruses will be grouped together by some but not others in a Venn diagram. For example, coronaviruses and HIV both infect eukaryotes, but coronaviruses are class IV with (+)ssRNA genomes while HIV is a class VI (+)ssRNA-RT virus. New discoveries add new methods of classification, such as the later addition to the Baltimore system of the class VII dsDNA-RT viruses or the discovery of viruses visible under a light microscope.

## *Giant Viruses*

Viruses have historically been classified as submicroscopic (unable to be seen with a light microscope) agents able to pass through filters that block larger microorganisms such as bacteria (Beijerinck, 1898; Ivanovski, 1892). Later discoveries would overturn this view. While most viruses are submicroscopic, some poxviruses as well as the 400 nm wide giant virus *Acanthamoeba polyphaga mimivirus* (APMV) identified in 2003 and its relatives are visible under light microscopes (Institute of Medicine (US) Committee on the Assessment of Future Scientific Needs for Live Variola Virus, 1999, p. 17; La Scola

et al., 2003). Prior to this 'large virus like particles' were reported in *Naegleria* amoeba in the mid-1970s (Schuster and Dunnebacke, 1976, 1974), various algae throughout the 1970s and 1980s (Allen Dodds and Cole, 1980; Pickett-Heaps, 1972; Sicko-Goad and Walker, 1979), and *Phaeodarea* amoeba in 1993 (Gowing, 1993), but these were not identified as viruses when they were seen. The mostly linear dsDNA genomes of giant viruses and their relatives are similarly large, ranging from 100 kbp to 2.5 megabase pairs (Mbp) in length (Philippe et al., 2013; Tidona and Darai, 1997; Van Etten et al., 2010). Now that giant viruses are being actively sought after and no longer filtered out of samples, the number of known specimens has expanded dramatically (Abergel et al., 2015). Giant viruses are found in a wide range of environments, from the Amazon River basin to the aforementioned Siberian permafrost (Katzourakis and Aswad, 2014).

Similar to some other nucleocytoplasmic large DNA viruses (NCLDVs), such as poxviruses, most known giant viruses reproduce in their hosts by creating dense structures in the cytoplasm called viral factories (Benamar et al., 2016; Colson et al., 2013, 2012; La Scola et al., 2008). Other reproductive strategies exist as well, *e.g.* rather than constructing a cytoplasmic viral factory, the giant virus *Pandoravirus* reorganizes the host nucleus to such an extent that it loses its spherical appearance and the nucleolus region disappears entirely (Philippe et al., 2013). The exact mechanisms for how these viral factories are created are unknown (Gaia et al., 2014; Kuznetsov et al., 2013; La Scola et al., 2008). Giant viruses possess hundreds to over two thousand open reading frames (ORFs) depending on the species (Aherfi et al., 2016). The large

majority of putative genes in giant viruses are ORFans, ORFs that do not have known orthologues in other lineages of viruses or cells. However, some of their ORFs include aminoacyl tRNA synthetases and translation factors previously thought to be the exclusive domain of cellular organisms, suggesting a degree of autonomy from their host cell for transcription and translation unprecedented in viruses (Aherfi et al., 2016). Because giant viral ORFans lack similar genes with known functions, we would need to determine the function of many if not most of the genes experimentally. However, the large number of genes involved would complicate this, especially as genes may be involved in multiple gene networks. Finally, some of the giant viruses are only known from metagenomic data, effectively precluding these sorts of experiments until a virion can be isolated (Aherfi et al., 2016; Claverie and Abergel, 2013; Colson et al., 2013). The differing replication strategies of various families and putative families within the giant viruses, as well as the phylogenetic placement of the large but submicroscopic faustoviruses within the giant viruses (Reteno et al., 2015; Santini et al., 2013) suggests that viral gigantism has evolved multiple times, and that the group is not strictly monophyletic.

## *Virophages*

Satellite viruses are viruses that make their own capsid proteins (as opposed to subviral satellites like hepatitis D which do not), but that cannot replicate on their own. They require co-infection by a helper virus in order to replicate, and can have positive,

negative, or neutral impacts on the helper virus reproduction and pathogenicity. In 2008, a satellite virus of the giant virus *Acanthamoeba castellanii mamavirus* (ACMV), dubbed sputnik virophage, was discovered (La Scola et al., 2008). Since the discovery of sputnik, other virophage families have been indentified, primarily in metagenomic data (Bellas et al., 2015; Fischer and Suttle, 2011; Gaia et al., 2014, 2013; Gong et al., 2016; Oh et al., 2016; Santini et al., 2013; Yau et al., 2011; Zhou et al., 2014, 2013). The term virophage was derived from bacteriophage, viruses which infect bacteria, since sputnik hijacks the replication machinery of the giant virus viral factory, rather than the native replication machinery of the host cell and has a negative impact on ACMV replication (La Scola et al., 2008). There is ongoing debate on the naming convention of 'virophage' since there are other satellite viruses that negatively impact their helper virus, but thus far all new satellite viruses of giant viruses have been called virophages (Desnues and Raoult, 2012; Fischer, 2011; Krupovic and Cvirkaite-Krupovic, 2011).

The presence of sputnik results in dysfunctional and abortive forms of the ACMV virions and interferes with proper capsid assembly. This reduces the yield of ACMV by 70% and causes a threefold decrease in host cell lysis (La Scola et al., 2008). Mavirus virophage similarly inhibits its host virus *Cafeteria roenbergensis virus* (CroV) and increase *Cafeteria roenbergensis* survival (Fischer and Suttle, 2011), whereas zamilon virophages, a close relative of sputnik, does not inhibit APMV reproduction or cell lysis (Gaia et al., 2014). The exact mechanisms behind how sputnik and other virophages hijack the viral factories are unknown, and since most virophages are only known from metagenomic data,

it is not certain how many of them rescue the eukaryotic host. It is certainly conceivable (and given the scale of the global virome, presumably very likely) that similar to the subviral satellite hepatitis D, a virophage could lead to increased morbidity and mortality in the host. Most virophages appear to have circular genomes ranging from ~17-29 kbp, but the order, orientation, and even presence of most genes are not well conserved between proposed virophage families (Bellas et al., 2015; Fischer and Suttle, 2011; Gaia et al., 2014, 2013; Gong et al., 2016; La Scola et al., 2008; Oh et al., 2016; Santini et al., 2013; Yau et al., 2011; Zhou et al., 2014, 2013). The reported size of the few directly observed virophage virions varies widely from 50-75 nm, with sputnik alone being reported as ranging from 50 to 75 nm in size (Fischer and Suttle, 2011; La Scola et al., 2008; Zhang et al., 2012).

### *EVEs, Virophages, & Polintons*

Endogenous viral elements (EVEs) are the result of heritable horizontal gene transfer (HGT) from a virus to a host's germline cells, which are subsequently inherited vertically (Katzourakis and Gifford, 2010). These can be deleterious, potentially leading to oncogenesis or viremia (Yu et al., 2012), and tend to be eliminated by natural selection within a few generations (Katzourakis and Gifford, 2010). However, not all EVEs are harmful enough to warrant eradication. Some do not have substantial selective consequences and can be lost by neutral genetic drift over longer intervals. Others can expand in copy number and/or become fixed within the host population, becoming regular vertically transmitted host alleles at that point (Holmes, 2011; Horie

et al., 2010). Integration and fixation is more common with retroviruses as they have an obligate integration step as part of their life cycle, and unsurprisingly, the first eukaryotic EVEs discovered were endogenous retroviruses (ERVs) (Weiss, 2006, 1969a, 1969b). While ERVs are more prevalent, EVEs derived from all seven Baltimore classes of virus have been identified across a broad range of hosts (Katzourakis and Gifford, 2010). Bacteriophages integrated into their hosts' genome, known as prophages, were identified in bacteria even earlier (Lederberg and Lederberg, 1953). Both prophages that are able to excise themselves from their host as well as permanently integrated 'disabled' ones are generally not referred to as EVEs, but they and their remnants can be thought of as such. Because viruses lack a mineralised fossil record, these EVEs serve as one of our main windows into viral evolution.

The large size of the giant virus genomes makes the integration of the entire viral genome into a host genome seem improbable as it could be highly disruptive, for example by substantially distorting the relative position of regulatory elements. However, large integrations can and do occur, such as chromosomally integrated endogenous human herpes virus 6 (162 kbp), which has been found in the germline of approximately 1% of the human population (Luppi et al., 1993; Morissette and Flamand, 2010). Small and medium sized (150 kbp to 320 kbp) phycodnaviruses, relatives of the giant viruses, have been found integrated into algae (Delaroque et al., 1999; Meints et al., 2007). Although these demonstrate that large-scale viral integration and endogenization is possible, the genome size of these herpes and phycodnaviruses

are anywhere from two to twenty-five times smaller than those of the giant viruses (Honess, 1984; Sandaa et al., 2001). Smaller segments can also transfer separately from the whole genome. For example, recently endogenized segments of phycodnavirus genomes have been reported in unicellular green algae, brown algae, and haptophytes (Blanc et al., 2010; Delaroque et al., 1999; Meints et al., 2007; Read et al., 2013). Genes of likely NCLDV origin have also been discovered in land plants, despite there being no known exogenous NCLDVs that infect them, suggesting previous contact (Maumus et al., 2014; Mushegian et al., 2016). Although HGT is less well examined in plants than other systems, it does occur (Gao et al., 2013), offering a possible mechanism for the presence of the segments. At least some of the virophages, like their giant virus hosts, have genes seemingly derived from all three domains of life, as well as ORFans, and virophages may facilitate HGT between giant viruses and their host eukaryotes. Genes taken from a giant virus by a virophage and subsequently donated to a eukaryote could be viewed as 'vicarious EVEs' (Katzourakis and Aswad, 2014).

The genome of sputnik 2 has been found integrated in its entirety in the genome of its host, Lentille virus (Desnues et al., 2012). EVEs and proviruses are usually defined as being integrated into a cellular organism's genome (Katzourakis and Gifford, 2010), which would exclude sputnik 2. However, the origins of giant viruses are unknown and debated. Some argue that they arose through the genome expansion of smaller viruses by the acquisition of host genes (Yutin et al., 2014). Poxviruses have been shown to deploy 'genomic accordions', expanding their genome size 7%-10% and acquiring higher

fitness via recurrent viral gene amplifications (Elde et al., 2012), suggesting another possible route to permanent genome expansion. Some phylogenetic analysis supports an accordion like model of giant virus evolution, but it remains controversial (Filée, 2015). Others argue, also controversially, that giant viruses are descended from a (now apparently otherwise extinct) fourth domain of life via the genome reduction commonly seen in parasitic organisms (Claverie and Abergel, 2013). Even though the integration of sputnik 2 into Lentille virus is likely a relatively recent development, if the latter reduction theory is correct, then it raises questions on how to classify and label these integrated virophages and foreign viral genes. Nature, as always, seems to abhor our attempts at neat and tidy categories.

Mobile genetic elements, in the form of DNA transposons and transpovirons, are also present in the eukaryotic host/giant virus/virophage systems and elsewhere. Transpovirons are short (~7.5 kbp) linear plasmids that resemble other transposable elements (TEs) but are found inside the giant virus and virophage virions, as well as integrated into their genomes (Desnues et al., 2012). Additionally, regular eukaryotic miniature inverted-repeat TEs are able to colonise and subsequently proliferate within the genome of the giant virus *Pandoravirus salinus*, as opposed to being acquired en masse from its amoeba host *Acanthamoeba castellanii* (Sun et al., 2015).

Virophages appear to be related to the TEs known as polintons, large DNA transposons found in many different eukaryotes (Fischer and Suttle, 2011; Yutin et al., 2013).

Indeed, it has been argued that the mavirus virophage is more closely related to the polintons than it is to other virophages (Fischer and Suttle, 2011; Yutin et al., 2013). However, this is extremely dependent on which gene is being looked at, *e.g.* integrase, which is found in mavirus and polintons but no other known virophages. The name polinton is derived from the polymerase B (POLB) and retroviral family integrase proteins found in most of them (Kapitonov and Jurka, 2006). Previously described DNA transposons are transposed via the direct transfer of their genomic copy between sites via 'rolling circle' or 'cut and paste' transposition mechanisms, both of which are dependent on the host replication machinery for DNA synthesis (Craig, 1995; Kapitonov and Jurka, 2001). Polintons have been proposed as being able to self-synthesise their DNA using POLB after an integrase catalysed (single stranded) excision from the host chromosome, followed by genomic insertion elsewhere by integrase (Kapitonov and Jurka, 2006). As with other TEs, polintons come in autonomous (encoding the full set of proteins necessary to replicate itself) and nonautonomous (being dependent on an autonomous relative for anywhere from one to all proteins) varieties. Most autonomous polintons are ~15-22 kbp with a few cnidarian examples being closer to 40 kbp (Bao et al., 2015; Haapa-Paananen et al., 2014).

## *Troubled Taxonomies*

'Giant virus' is a morphological designation as opposed to a strictly phylogenetic one, describing viruses other than poxviruses with virions that are visible with light

microscopy (Abergel et al., 2015; Legendre et al., 2015) and as mentioned, is probably not a monophyletic grouping. Because of the diffraction limit of visible light, this has been around 300 nm, however with the creation of superlenses (lenses that use metamaterials to go beyond the diffraction limit) there is theoretically no lower limit for the size of an object that can be clearly resolved (Pendry, 2000; Wang et al., 2011), so a more concrete definition should be established. The term giant virus is used with varying degrees of synonymy for NCLDVs, and sometimes includes the closely related faustoviruses and phycodnaviruses, despite those two groups generally being less than 300 nm in size (Reteno et al., 2015).

APMV was the first giant virus identified and became the type species and genus for the viral family *Mimiviridae*, recognised by the International Committee on Taxonomy of Viruses (ICTV). In addition to more mimiviruses, other viruses such as faustoviruses (Reteno et al., 2015), Marseilleviruses (Colson et al., 2012), molliviruses (Legendre et al., 2015), pandoraviruses (Philippe et al., 2013), and pithoviruses (Legendre et al., 2014) have been discovered since then, founding families of their own. Most of these additional families are putative placements, with only *Marseilleviridae* being currently recognised by the ICTV. These recognised and putative virus families belong to a putatively monophyletic order of viruses along with the smaller (generally 100 to 400 kbp genomes) NCLDV families *Ascoviridae*, *Asfarviridae*, *Iridoviridae*, *Phycodnaviridae*, and *Poxviridae* (Colson et al., 2013). The order *Megavirales* is not an officially recognised order but the term is still widely used in the

literature to encompass all of these families. The origins of the clade are less clear, obscured by significant HGT, a large ratio of ORFans to ORFs with identifiable relations, and what appears to be a deep evolutionary history. As mentioned earlier there are some who claim *Megavirales* are descended from an otherwise extinct fourth domain of life (Claverie and Abergel, 2013; Raoult, 2013), but the hypothesis remains deeply controversial (Williams et al., 2011; Yutin et al., 2014). Different papers have presented phylogenetic trees with a widely varying levels of disagreement, depending on which genes are being used to create the trees, and which viruses (and cellular organisms) are being sampled (Aherfi et al., 2016; Benamar et al., 2016; Delaroque et al., 1999; Fischer et al., 2010; Raoult, 2013; Williams et al., 2011; Zade et al., 2015).

The ICTV has also recognised the family *Lavidaviridae*, containing two genera, *Mavirus* and *Sputnikvirus,* comprising the virophages which have been isolated: mavirus, sputnik, and zamilon (Krupovic et al., 2015). As the other virophages are only known metagenomically they have no formally recognised place. Because of the close relationship between virophages and polintons, the origins of both are if anything more obscured than those of giant viruses. As with giant viruses, trees vary by which gene is being examined (Blanc et al., 2015; Fischer and Suttle, 2011; Gong et al., 2016; Iyer et al., 2008; Yutin et al., 2015b, 2015a, 2013), which makes classification difficult. Further complicating the issue, sputnik appears to be evolutionarily linked to bacterial TEs carrying TV-Pol family genes (*e.g.* IS605 or SCCmec elements) (Iyer et al., 2008), and has a phage type bacterial insertion transposase that is entirely

unrelated to the retroviral type integrase found in mavirus and polintons (La Scola et al., 2008). This further reinforces the idea that there is extensive mosaicism, but interestingly may suggest a degree of convergent evolution as well if sputnik reacquired an integration system after losing the retroviral integrase. Putative hybrids between virophages and polintons have also been reported from metagenomic data (Yutin et al., 2015a). Since polintons are found across a broad range of eukaryotes including multicellular ones, it has been suggested that they may have arisen from an ancient mavirus-like ancestor that became able to integrate and transposition itself intragenomically (Figure 0.1a) (Fischer and Suttle, 2011; Katzourakis and Aswad, 2014). However, others have suggested that virophages are escaped polintons that recombined with an unknown virus (Figure 0.1b) (Yutin et al., 2013).

A more controversial hypothesis for the origins of polintons and virophages combines elements of both scenarios, and puts hypothetical 'polintoviruses' at the origin of not just virophages but also *Megavirales,* bidnaviruses, and adenoviruses (Figure 0.2) (Koonin et al., 2015; Krupovic and Koonin, 2014). In this hypothesis, tectiviruses (bacteriophages infecting gram-negative bacteria) came into eukaryotes as part of symbiogenesis and the capture of mitochondria (Koonin et al., 2015; Krupovic and Koonin, 2014). Tectiviruses acquired cysteine protease and integrase from already extant transposons, becoming 'polintoviruses' capable of existing exogenously and endogenously (Koonin et al., 2015). Polintoviruses then differentiated into bidnaviruses via recombination with a parvovirus, into adenoviruses through the loss of integrase, and

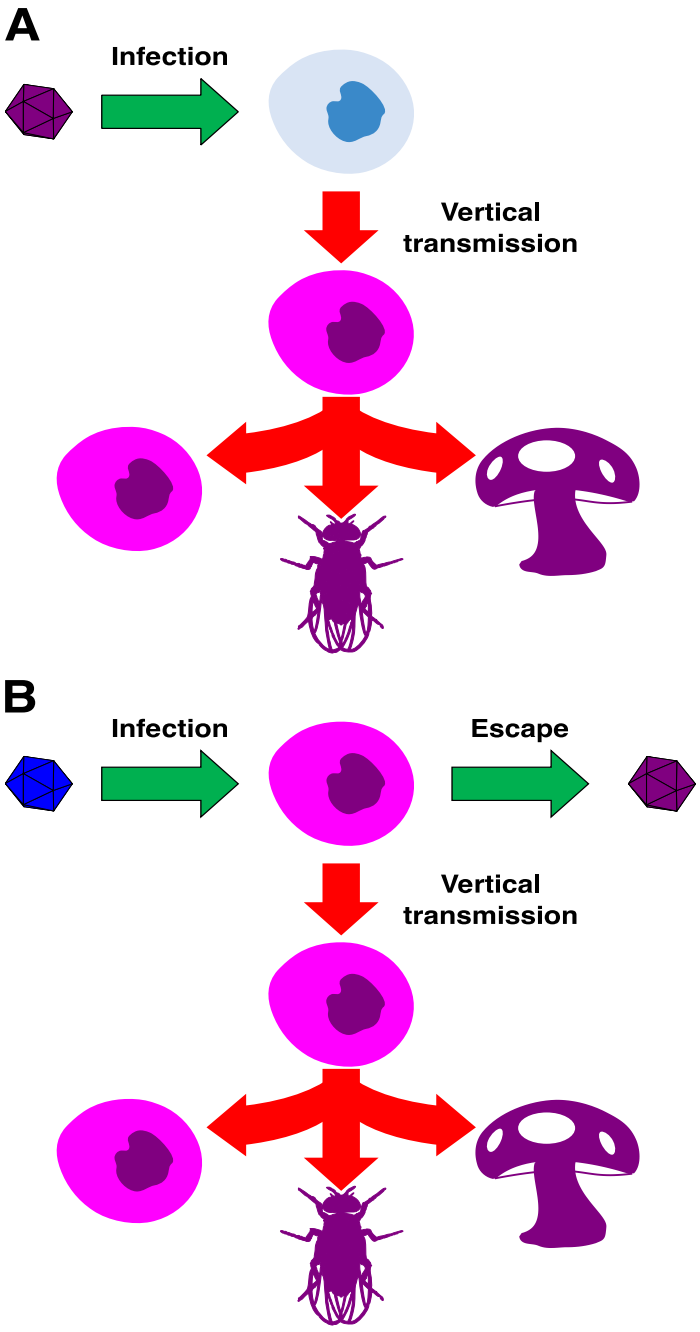



*Figure 0.1* Two proposed scenarios for the origins of polintons. **A**) A virophage (purple icosahedron) integrated into the genome of a polinton free cell (blue) and was then transmitted vertically, as a standard EVE (purple), differentiating along with its hosts. **B**) An unknown virus (blue icosahedron) infected a cell and recombined with pre-existing polintons (purple). The polintons then escaped as virophages (purple icosahedron).

into true polintons through the loss of their capsid gene (Koonin et al., 2015).

Polintoviruses also separately escaped the nucleus via the acquisition of DNA dependent

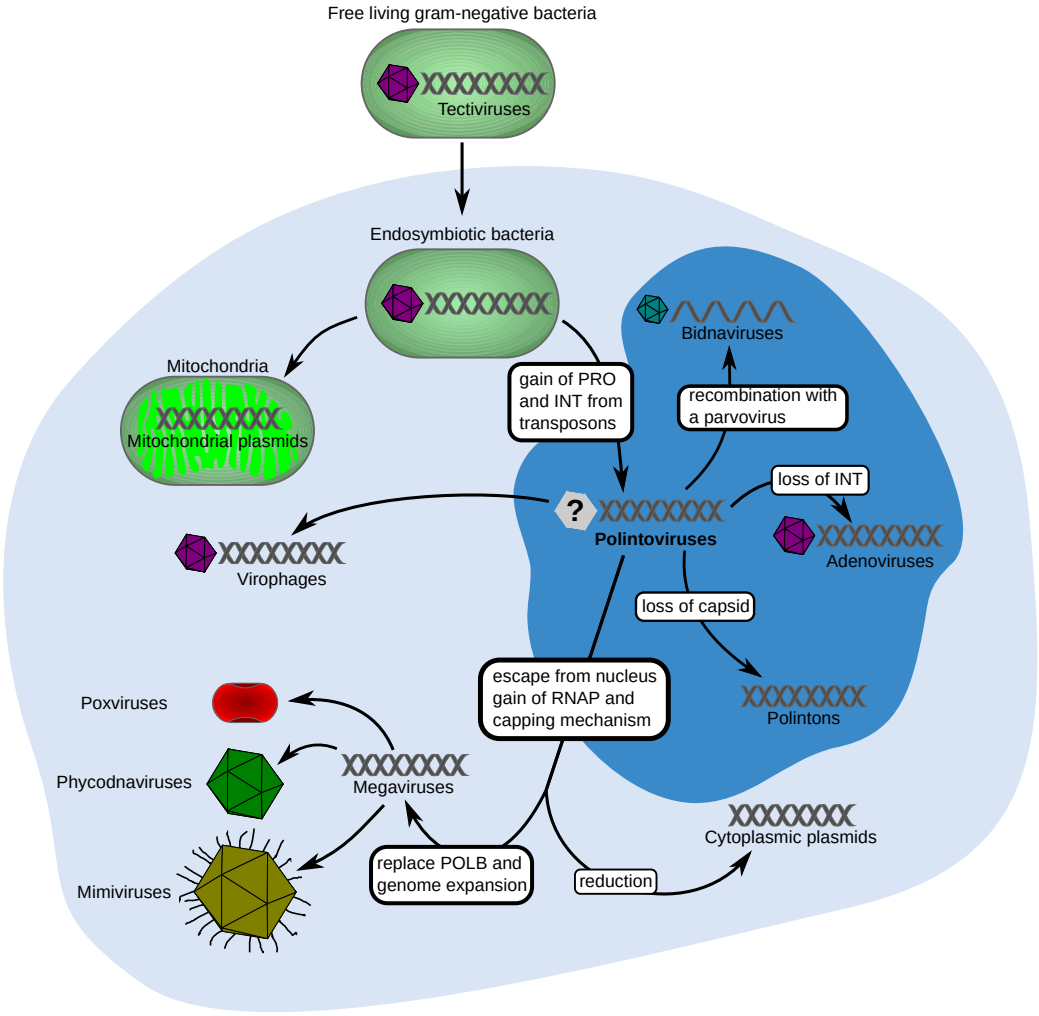



*Figure 0.2* A 'polintovirus' centred scenario for the evolution of polintons, plasmids, and various eukaryotic DNA viruses. See text for details. Icosahedrons represent icosahedral capsids. Capsid protein of bidnaviruses is unrelated to that of tectiviruses and their descendants. Greyed out icosahedron by polintoviruses indicates that (like polintoviruses themselves) the predicted capsid has never been observed. Double helices represent dsDNA. Single helix represents ssDNA. INT: retroviral family integrase; PRO: cysteine protease; POLB: DNA polymerase B; RNAP: DNA dependent RNA polymerase. Modified from Koonin et al., 2015; Krupovic and Koonin, 2014.

RNA polymerase and an mRNA capping system, forking into cytoplasmic plasmids via reduction, and into *Megavirales* via the replacement of POLB, and a 10-100 fold (or more) genome expansion (Koonin et al., 2015). Separately, polintoviruses differentiated into more permanently exogenous virophages and began parasitizing and swapping genes with *Megavirales* (Koonin et al., 2015). If the currently observed distribution of integrase

in virophages (*i.e.* only in mavirus and its close relatives) is representative of the actual distribution, then this last step would have to involve the loss of integrase from most, but not all, virophage lineages. It should be noted that there is virtually no direct support for this hypothesis. Polintoviruses have never been observed as exogenous viruses, either *in vivo* or *in vitro*. Their existence is predicated on the existence of putative capsid genes in some polintons (Koonin et al., 2015; Krupovic and Koonin, 2014). The ORFs in question do not align with virophage capsid proteins at the sequence level for nucleic acids (NA) or amino acids (AA), but are claimed to form the double jelly roll fold seen in virophage capsid proteins when protein folding is digitally simulated (Krupovic and Koonin, 2014). Even if the putative capsid proteins found in some polintons are real, they could have been acquired by polintons via recombination with exogenous viruses or be left over from the initial integration described in Figure 0.1. This latter explanation would also not require the repeated gain and loss of integrase from polintoviruses as they differentiated into adenoviruses, bidnaviruses, *Megavirales*, and most but not all virophages, an issue which the proponents of the theory do not address. There is also no explanation for why some polintoviruses escaped the nucleus and underwent enormous genome expansion while others remained more or less the same size.

These irresolvably tangled phylogenies and obscure taxonomies are unfortunately, not exclusive to the viral world. Differing phylogenies and classification systems for Eukarya have been suggested (*e.g.* Adl et al., 2012; Simpson and Roger, 2004) but regardless of the system used, polintons and polinton-like elements (PLEs) are

widespread amongst eukaryotes (Bao et al., 2015; Krupovic and Koonin, 2014; Yutin et al., 2015b). Polinton and PLEs have been found in green algae, *Cryptophyta* (flagellate algae), and the SAR super group (which includes water moulds, dinoflagellates, *Bigelowiella natans*, and *C. roenbergensis,* proposed in Burki et al., 2007), but not land plants. They are also found amongst Amoebozoa, animals (notably absent from mammals), and fungi plus other single celled organisms that vary drastically in their placement under different systems. Given how distantly related some of hosts of polintons and PLEs are (*e.g.* wasps and dinoflagellate endosymbionts of corals), there would need to have been extensive HGT between different kingdoms/super groups within eukaryotes, or else the ancestral integration would need to have happened at least a billion years ago. Although this amount of time seems like it would render polintons unrecognisable, it is claimed that phylogenetic analysis supports the primarily (though not exclusively) vertical transmission and evolution of polintons from a unicellular ancestor (Haapa-Paananen et al., 2014). However, some evidence of HGT is still seen between different eukaryotes, as well as from a braconid parasitoid wasp chromosome to its proviral endogenous bracovirus (Dupuy et al., 2011).

## *Ecological Effects*

While the relatives of the giant viruses, *Iridoviridae*, *Poxviridae*, *etc.* are often extremely serious threats to multicellular life, the giant viruses of the family *Mimiviridae* appear to only infect single celled eukaryotes, and have not been clearly associated with any

pathologies of multicellular organisms (Vanspauwen et al., 2013). Many of the single celled organisms infected by giant viruses are primary producers or otherwise underpin food webs, meaning giant viruses can have substantial effects on nutrient cycling and population dynamics (Fischer and Suttle, 2011; Yau et al., 2011). Harsh ecosystems with highly truncated food webs, such as hyper saline lakes or acidic hot springs will tend to be microbially dominated, and in such environments, viruses can be one of the dominant 'predators'. Furthermore, giant viruses have been found inside coral endosymbionts (*Symbiodinium sp.*), and associated with a coral bleaching event (Correa et al., 2016), showing how they can have major effects on macroscopic ecosystems. Because virophages parasitise giant viruses and, at least in some cases, can prevent host cell death, they can be beneficial to the helper virus's host. By increasing the eukaryote's survival, virophages can play a significant role in viral-host dynamics. This results in complex ecological interactions between virophage, eukaryote, and giant virus.

In the case of *B. natans*, a unicellular alga, a number of genes of putative virophage origin, dubbed virophage like elements (VLEs), have been incorporated into the host nuclear genome and are transcriptionally active, possibly as a defence mechanism against a giant virus (Blanc et al., 2015). However, no known giant virus has been observed parasitising *B. natans*, so the defensive nature is speculative. Nevertheless, even if there is not currently a virus that the VLEs in *B. natans* defend against, they could be indicative of a historical confrontation, serving as an indirect fossil record. The flagellate *C. roenbergensis* is host to the giant virus CroV, itself host to the mavirus

virophage. The de novo integration of mavirus into *C. roenbergensis* was reported as being experimentally observed (Fischer and Hackl, 2016). Unlike in *B. natans* these provirophages remained quiescent unless the cell was infected with CroV, which resulted in the production and eventual release of mavirus virions. Although this did not seem to protect the individual *C. roenbergensis* cell initially infected with CroV, it did protect the wider population as a whole from lysis and it was suggested that endogenous virophages can act as an antiviral defence system in natural protist populations (Fischer and Hackl, 2016). The authors do not elaborate on what the unit of selection would be, but presumably this would be via kin selection, similar to eusocial insects. *C. roenbergensis* does not have any known sexual reproduction, replicating entirely asexually via binary fission (Hackl et al., 2019). Populations are therefore clonal, in which case rescuing the individual and rescuing the host will pass on essentially the same complement of genes. Alternatively, non-integrated mavirus does help protect the host, so it is possible that there was a small level of protection provided that they did not detect, and that the population defence was an entirely secondary effect. The cases of *B. natans* and *C. roenbergensis* provide evidence of an evolutionary scenario needed for a hypothetical origin of polintons from exogenous virophages, however, they remain the only two reported cases of virophage/VLE integration in eukaryotes.

# Part 2

Chapter 2 of this thesis examines HLAs in modern humans derived from archaic hominids, Neanderthals and Denisovans. The prevalence and geographic distributions of archaic HLAs as well as possible causes for the observed patterns are explored. This chapter shows that there is a correlation between archaic HLA frequency in modern human populations and environmental predictors such as distance from the Equator, and that the different types of class I HLAs (A, B, and C) maintain different frequencies of archaic HLA variants suggesting different selective pressures.

## *Human Leukocyte Antigen*

In vertebrates, the coevolution of a host with its pathogens is probably most readily apparent in the genes of the major histocompatibility complex (MHC). Found on the short arm of chromosome six in humans, these genes are known as human leukocyte antigen (HLA) genes and comprise the most rapidly evolving portion of the genome (Bergström et al., 1998; Shiina et al., 2006). They were initially detected via similarly reacting sera as an alloantigen present on leukocytes (hence the name) and their role in self vs. nonself detection was quickly recognised, though initially in the context of organ transplantation (Dausset, 1958; Thorsby, 2009). Their principle biological role is pathogen vs. self peptide recognition and presentation and this was identified soon thereafter (Levine et al., 1963). HLAs are divided into three classes comprising multiple types and although the function of all HLA types is not known, generally Class I and

*Table 0.1* The number of known HLA alleles by type. Alleles in purple italics are considered the main Class I alleles. Alleles in blue bold are considered the main class two alleles. Data from (González-Galarza et al., 2015).

| Locus | Number of Known Alleles |
|---|---|
| *HLA-A* | *3,094* |
| *HLA-B* | *3,866* |
| *HLA-C* | *2,618* |
| HLA-DMA | 7 |
| HLA-DMB | 13 |
| HLA-DOA | 12 |
| HLA-DOB | 13 |
| **HLA-DPA1** | **39** |
| **HLA-DPB1** | **520** |
| **HLA-DQA1** | **54** |
| **HLA-DQB1** | **777** |
| HLA-DRA | 7 |
| **HLA-DRB1** | **1,719** |
| HLA-DRB2 | 1 |
| HLA-DRB3 | 59 |
| HLA-DRB4 | 15 |
| HLA-DRB5 | 21 |
| HLA-DRB6 | 3 |
| HLA-DRB7 | 2 |
| HLA-DRB8 | 1 |
| HLA-DRB9 | 1 |
| HLA-E | 17 |
| HLA-F | 22 |
| HLA-G | 50 |
| HLA-H | 12 |
| HLA-J | 9 |
| HLA-K | 6 |
| HLA-L | 5 |
| HLA-P | 5 |
| HLA-V | 3 |
| HLA-Y | 3 |

Class II are involved in antigen presenting and the Class III region contains complement genes as well as genes with no immune or inflammatory function (Shiina et al., 2004). The HLA region includes the most polymorphic loci in the human genome, with

anywhere from one to thousands (Table 0.1) of different identified alleles for the different HLA types (Campbell and Trowsdale, 1993; González-Galarza et al., 2015; Kawashima et al., 2012). Many, and at least in Class I most (>80%), of these alleles are very rare, having been identified in only a single family or individual, with only minor point mutations differentiating them from more common alleles (Robinson et al., 2017). The frequency of these single nucleotide polymorphism (SNP) differentiated alleles tracks to the human germline mutation rate and predicts 8-9 million different HLA Class I alleles globally (Robinson et al., 2017). Distinguishing functional and silent random variation from one another is both challenging and has important implications for disease and transplant outcomes (Crux and Elahi, 2017; Dendrou et al., 2018; Robinson et al., 2017).

Given the number of HLA alleles, the number of possible combinations is enormous. However, certain combinations of HLAs, what is known as a haplotype, are much more common than what would be expected by chance (Choo, 2007; Dendrou et al., 2018). This non-random association of alleles at different loci is known as linkage disequilibrium. In general, the different HLA genes are linked and the entire HLA region is inherited in a Mendelian fashion from each parent.

### *HLA Structure*

Class I HLAs (Figure 0.3a) are noncovalently bound heterodimers comprising a single transmembrane heavy chain and a $\beta_2$-microglobulin light chain (Bjorkman and

Parham, 1990; Choo, 2007). The heavy chain is encoded by the HLA region on chromosome 6, while the $\beta_2$-microglobulin is encoded by the *B2M* gene on chromosome 15. The protein domains distal to the membrane, $\alpha_1$ and $\alpha_2$, are highly variable, and this is where the peptide binding site is located. The membrane proximal domains, $\alpha_3$ and the $\beta_2$-microglobulin, are more constant and part of the immunoglobulin (Ig) superfamily which also includes T cell receptors (TCRs) (Bjorkman and Parham, 1990). Class II HLAs (Figure 0.3b) are heterodimers composed of two transmembrane heavy chains encoded by the HLA region of chromosome 6.

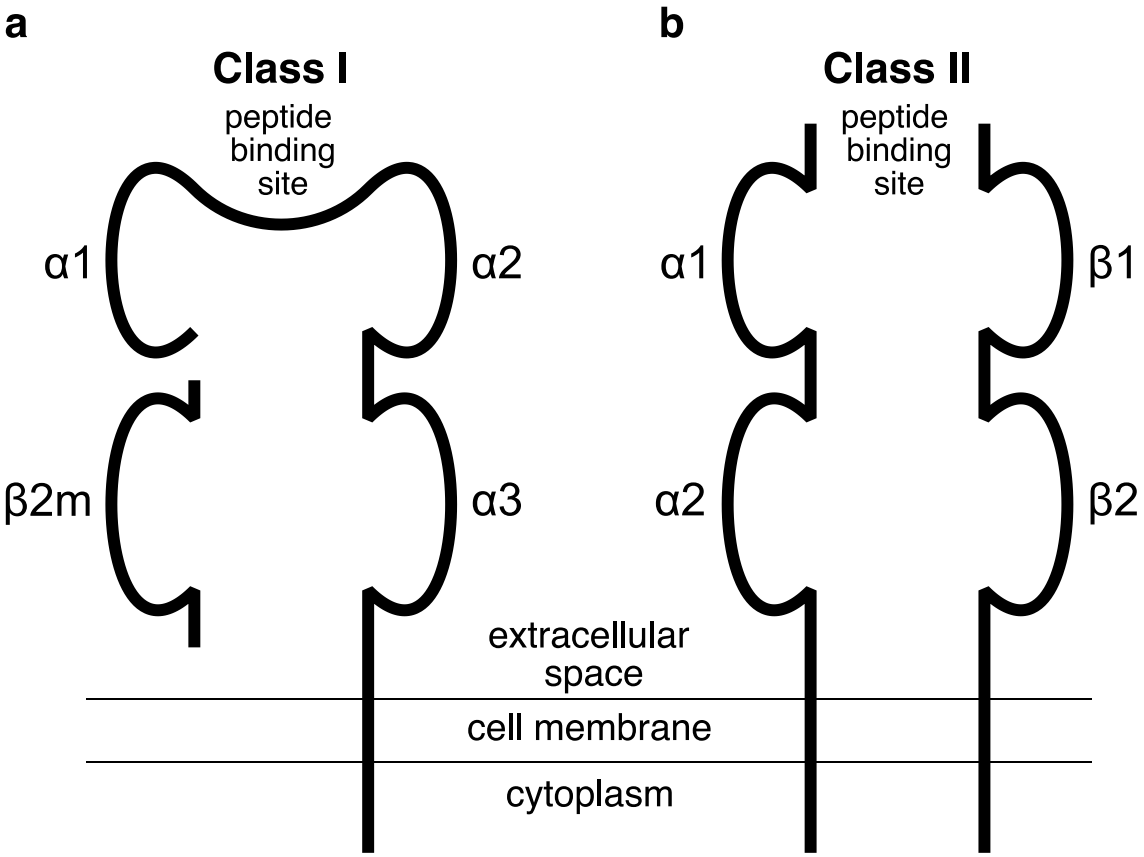



*Figure 0.3* Schematic representation of Class I (**a**) and Class II (**b**) HLAs. Class I HLAs are noncovalently bound heterodimers and are composed of a single heavy chain derived from the HLA gene region and an extracellular light chain $\beta_2$-microglobulin derived from elsewhere. The $\alpha_1$ and $\alpha_2$ domains are highly variable while the $\alpha_3$ and $\beta_2$-microglobulin are more constant. Class II HLAs are a heterodimer consisting of two heavy chains. As with Class I HLAs, the membrane distal $\alpha_1$ and $\beta_1$ domains are highly variable while the membrane proximal $\alpha_2$ and $\beta_2$ domains are more constant.

The $\alpha_1$ and $\alpha_2$ domains of the α-chain are respectively very similar to the $\alpha_1$ and $\beta_2$-microglobulin of Class I HLAs (Brown et al., 1993). On the β-chain, the $\beta_1$ domain is very similar to the Class I $\alpha_2$ domain while the $\beta_2$ domain matches the Class I $\alpha_3$ domain less closely (Brown et al., 1993). As with Class I the membrane distal portions are more variable while the Ig-like proximal portions are more constant (Wieczorek et al., 2017).

The peptide binding sites of both Class I and Class II HLAs are broadly similar. For both, the site is made up of two domains, derived either from a single chain in Class I or the two chains found in Class II. The peptide binding site is a groove at the membrane distal end of the protein. The walls of the groove are formed by two raised antiparallel α-helices with a β-sheet forming the floor (Bjorkman et al., 1987a, 1987b; The UniProt Consortium, 2019; Wieczorek et al., 2017). The groove is closed at both ends in Class I HLAs which generally restricts peptide size to around 8-10 residues (Matsumura et al., 1992; Wieczorek et al., 2017). In Class II HLAs the groove is open at both ends and able to accommodate longer peptides, 13-25 residues though a modal average of 15, with some of the peptide being able to project out of the groove entirely (Chicz et al., 1992; The UniProt Consortium, 2019). The groove necessarily must be able to accommodate a broad array of different epitopes. For example, using a 20 amino acid alphabet over a 10 residue long peptide has over $10^{13}$ (short scale: 10 trillion) different possible combinations (Sewell, 2012). However not all of these possible epitopes bind with equal affinity or even at all to any given HLA variant, and this binding affinity is controlled by the fine grain structure of the groove (Falk et al., 1991).

The fine grain structure of the groove is highly polymorphic and binds different peptides with varying degrees of affinity based on shallow and deep pockets in the groove being occupied by peptide side chains, and by hydrogen bonds between the HLA and the peptide backbone (Brown et al., 1993; Falk et al., 1991). In practice, this means that a given HLA will be better at binding (and presenting to T cells, see next section) the epitopes of some pathogens and worse at others. Upon exposure to a novel virus, this affinity is likely going to be fairly random. HLA B*27:05 for example has been linked to CD8+ T cell responses to both Gag and polymerase conserved epitopes and is strongly associated with a slow progression of HIV (Adland et al., 2017). Due to the relatively recent appearance of HIV, it is unlikely that HLA-B*27:05 evolved specifically to combat HIV, it just happens to be very efficient at it. Genome wide studies have linked different HLA alleles to the risk, severity, and/or treatment outcome of many major infectious diseases as well autoimmune diseases (Chapman and Hill, 2012; Dendrou et al., 2018)

Obviously, an individual that is better able to present epitopes of an endemic pathogen and thus better able to fight that pathogen gains a survival advantage. In this way, endemic pathogens can place a selective pressure on a population, leading to an increased prevalence of that HLA within the population over time (Hughes and Yeager, 1998; Prugnolle et al., 2005). However, a population also benefits from maintaining some diversity in its HLAs due to the potential exposure to novel pathogens (Hughes

and Yeager, 1998). There seems to be a mechanism to promote this in humans, where HLA dissimilarity has been shown to correlate with mate partnership (Kromer et al., 2016). Although well established in other mammals, birds, and fish, the extent of this influence on human behaviour and the mechanism, generally presumed to be olfactory, remains fairly controversial (Kromer et al., 2016; Natsch, 2014). Regardless, pathogens generally have much faster generation times and evolution rates when compared to their vertebrate hosts. The high degree of polymorphism found in the HLA system is widely believe to be a defensive measure evolved by vertebrates in order to compensate for what might otherwise be a crippling handicap when competing with their more rapidly evolving pathogens (Markov and Pybus, 2015; Sommer, 2005).

### *HLA Function*

Although involved in other processes such as natural killer cell licensing, the prime function of HLAs is antigen presentation (Figure 0.4). Class I HLAs comprising, HLA-A, HLA-B, and HLA-C alleles, are expressed on the surface of nearly all nucleated cells (Crux and Elahi, 2017; Daar et al., 1984). In contrast, the products of the Class II HLA genes, *DP*, *DQ*, and *DR* are principally found on professional antigen presenting cells such as B cells, dendritic cells, and macrophages (Choo, 2007; Cresswell, 1994). Class I HLAs predominantly bind peptides derived from cytosolic proteins (though with some nuclear proteins as well) produced by the cell whether these are self (Figure 0.4a) or pathogen (Figure 0.4b) in origin (Jardetzky et al., 1991). Class II HLAs predominantly

present antigens extracted from the extracellular milieu (Figure 0.4c). There are some reports of Class II HLAs being able to bind peptides directly from the extracellular milieu without an internalisation process but Class II HLAs are primarily loaded with

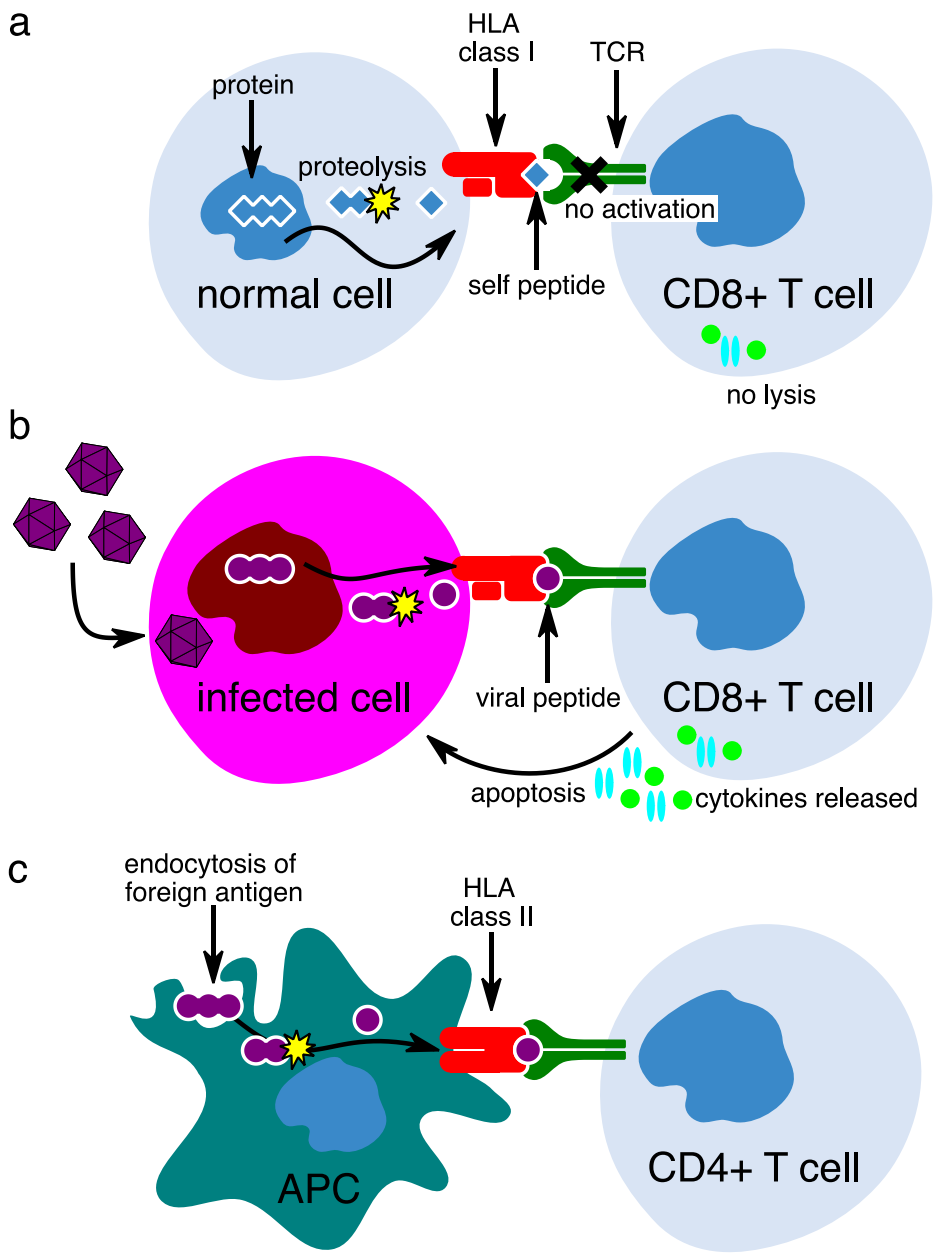



*Figure 0.4* Cartoon schematic of the use of HLA class I (**a** and **b**) and class II (**c**). (**a**) normal cellular proteins are broken down into smaller peptides and presented on the cell surface but do not activate CD8+ T cells. (**b**) peptides derived from viral proteins are presented to CD8+ T cells which causes the T cell to induce lysis in the infected cell. (**c**) APCs endocytise foreign antigens and present them to CD4+ T cells. HLA: human leukocyte antigen. TCR: T cell receptor. APC: Antigen presenting cell.

peptides derived from endogenised antigens (Cresswell, 1994; Wieczorek et al., 2017).

The combined Class I HLA/peptide complex serves as a target for TCRs on CD8+ T cells (Figure 0.4a and 0.4b) which bind to exposed residues on the peptide as well as accessible portions of the α-helices of the HLA (Pamer and Cresswell, 1998). When naïve CD8+ T cells are activated by binding to HLA/pathogen peptide complexes, and the presence/absence of other positive/negative costimulatory signals (e.g. CD28 or IL-2), they become either effector cytotoxic T lymphocytes (CTLs) or memory CD8+ T cells. Activated CTLs can then destroy infected cells by inducing apoptosis (Figure 0.4b). The Class II HLA/peptide complex is the target for TCRs on CD4+ T cells, also known as 'T helper cells' (Figure 0.4c). As with CD8+ T cells, CD4+ T cells require additional costimulatory signals in order to activate. Depending on the signals involved when the naïve CD4+ T cells are activated they can differentiate into numerous different subtypes (e.g. Figure 0.5) which play different roles in the adaptive immune response (Bonelli et al., 2014; Eyerich et al., 2009; Owen et al., 2013). It was originally thought that these were terminal differentiations but there is actually a large degree of plasticity between different subtypes (Bonelli et al., 2014; Geginat et al., 2014).

CD4+ T helper cells are critical for both the antibody mediated (humoural) and cell based adaptive immune response, as they activate and recruit other immune cells such as basophils, B cells, CD8+ CTLs, eosinophils, macrophages, mast cells, and neutrophils (Geginat et al., 2014; Zhu and Paul, 2010). Though there are many

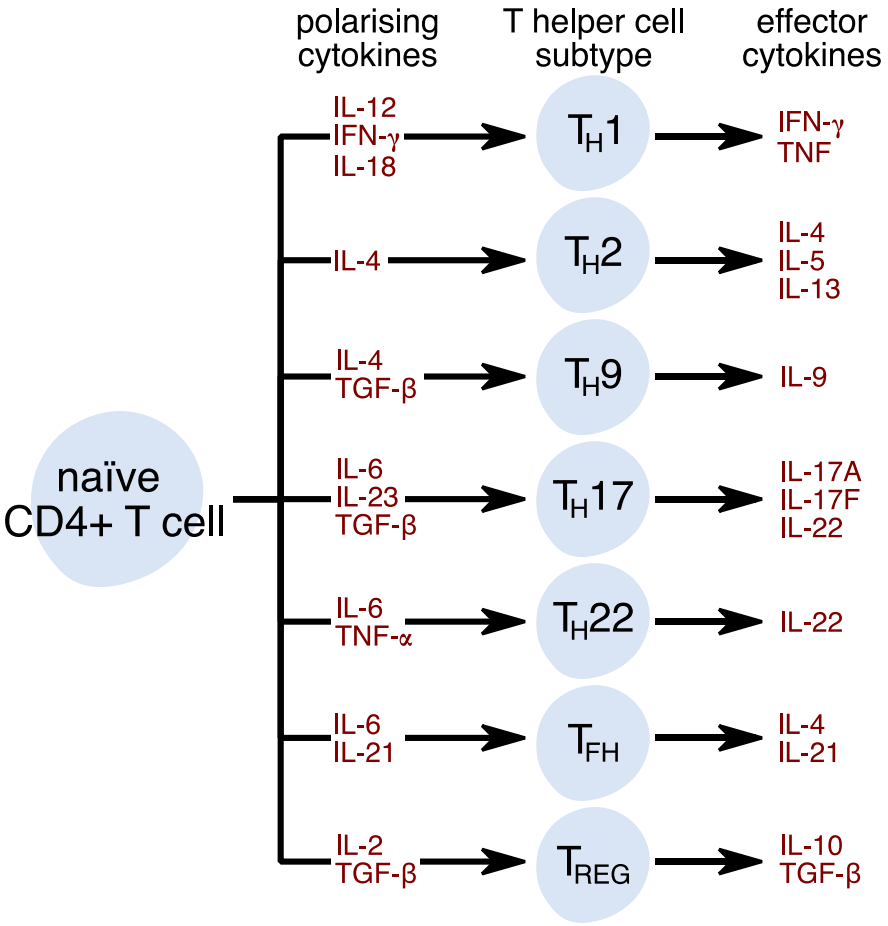



*Figure 0.5* A small sample of T helper cell subtypes that naïve $CD4^+$ T cells can differentiate into based on exposure to various polarising cytokines. Different subtypes express different effector cytokines upon differentiation. IL Interleukin; IFN Interferon; TGF Transforming growth factor; TNF Tumour Necrosis Factor; $T_H$ T Helper; $T_{FH}$ T Follicular Helper; $T_{REG}$ Regulatory T Cell. Data from (Bonelli et al., 2014; Eyerich et al., 2009; Owen et al., 2013)

identified CD4+ subtypes, and the list continues to grow, five 'principle' lineages are recognised: Follicular helper T cells ($T_{FH}$) which help and regulate B cells, regulatory T cells ($T_{REG}$) which regulate self-tolerance of the immune system, and $T_H1$, $T_H2$, and $T_H17$ which target specific classes of pathogen (Crotty, 2011; Geginat et al., 2014; Korn et al., 2007; Sakaguchi, 2005). In addition, there are also a variety of memory T cells, such as central memory and effector memory T cells which can provide lasting protection after a primary infection by responding quickly to secondary infections, in some cases by

changing into an effector lineage such $T_H1$ (Geginat et al., 2014; Owen et al., 2013).

Humans are exposed to a constant barrage of pathogens many of which have evolved sophisticated methods of evading the innate immune response. The adaptive immune response mediated by HLAs and CD4+ cells works to counter this. The nature of CD4+ T cells as the linchpin of the adaptive immune response becomes immediately apparent in their absence such as with HIV infection. Infected individuals become susceptible to opportunistic infections and cancers due CD4+ depletion and some of these HIV related diseases can even be stratified by CD4+ count (Jung and Paauw, 1998).

# Part 3

Chapter 3 of this thesis is in some ways the inverse of chapter 2, looking at the coevolution of HIV with humans. The dominant paradigm (discussed in detail in Chapter 3) for more than 20 years has been that HIV subtype diversity is merely an artefact of random founder effects rather than being a result of adaptive processes (Hu et al., 1999; Rambaut et al., 2001). The work here shows that while founder effects likely still play a role, adaptive processes are at work.

## *Origin of HIV*

Each human immunodeficiency virus 1 (HIV-1) lineage or 'group' is the result of a single independent cross species transmission of simian immunodeficiency virus (SIV) and are

labelled group M, group N, group O, and group P. Phylogenetic evidence indicates that groups M and N crossed over from chimpanzees (*Pan troglodytes troglodytes*) while groups O and P likely stem from western lowland gorillas (*Gorilla gorilla gorilla*), , though they may have still been transferred to humans via chimpanzee (Keele et al., 2006; Plantier et al., 2009; Van Heuverswyn et al., 2007). Other cross species transmissions from sooty mangabeys (*Cercocebus atys*), have led to various HIV-2 lineages (Marx et al., 1991).

The different crossover events had widely varying degrees of success. HIV-1N accounts for only a small minority of HIV cases in Cameroon (Vallari et al., 2010). Group O likewise has only low prevalence primarily in Gabon (Liégeois et al., 2013) and HIV-1P has only been identified in two patients to date (Plantier et al., 2009; Vallari et al., 2011). All the HIV-2 strains combined only account for 1 to 2 million infections and are primarily constrained to West Africa (Campbell-Yesufu and Gandhi, 2011; Esbjörnsson et al., 2012). HIV-1M is the strain responsible for the global pandemic currently infecting ~38 million people and estimated to have infected ~75 million people since its start (UNAIDS, 2019).

These crossover events are thought to likely have occurred from bushmeat hunting, and indeed modern bushmeat hunters continue to be exposed to and infected by simian retroviruses (Kalish et al., 2005; Wolfe et al., 2004). Similar crossover events thus have likely been occurring for thousands if not tens of thousands of years, before the right alchemy of biological traits and human social factors allowed HIV-1M to go global.

## *Evolution of Pandemic HIV-1 group M*

HIV-1M emerged around 1900 in either the Belgian Congo, now the Democratic Republic of the Congo (DRC) or Cameroon (Kalish et al., 2004; Keele et al., 2006; Korber et al., 2000; Salemi et al., 2001; Van Heuverswyn et al., 2007; Vidal et al., 2000; Worobey et al., 2008). Originally it spread only slowly from its rural point of origin until human social and cultural factors, many brought about by colonialism in the region, accelerated its spread. These factors included the construction of roads and railroads, urbanisation, increased mobility of migrant labour, non-sterile medical practices, and changes in sexual activity (Faria et al., 2014; Morse, 1995; Sharp and Hahn, 2008). In an effort to treat endemic and imported diseases, colonial powers attempted both environmental and human interventions in varying proportions (Worboys, 1994). In their efforts to stamp out sleeping sickness, the British tended to focus more on environmental intervention, whereas the French and Belgians focused more on trying to cure infected individuals in order to prevent them from transmitting the disease onwards (Headrick, 2014). The Belgians were particularly brutal in this regard, establishing forced internment camps, known as *lazarets*, for sick Africans, where those suspected of being sick were given injections of the relatively toxic arsenic compound Atoxyl (Headrick, 2014). These injections were not voluntary or optional, and the widespread reuse of non-sterile needles, compounded by immune systems already weakened by horrific conditions in the camps, has been proposed as exacerbating if not kicking off the nascent HIV-1 pandemic (Pepin, 2011). The pandemic only started in

earnest when the virus reached Léopoldville (now Kinshasa) in the Belgian Congo in the 1920s, the time of the most recent common ancestor for Group M (Faria et al., 2014). From there, it began spreading along the colonial transport networks both within the country and beyond (Faria et al., 2014; Sharp and Hahn, 2008).

Both Group M and Group O were epidemiologically similar until ~1960, exhibiting a relatively slow exponential growth essentially keeping pace with the rate of population growth in Kinshasa (Faria et al., 2014). In ~1960, while Group O's growth remained relatively the same, Group M's growth rates more than doubled, and it is around this time that we see an explosive diversification of Group M subtypes (Faria et al., 2014; Worobey et al., 2008). The exact cause of this shift is unknown as is whether the increased growth rate led to diversification or vice versa. However, several possible potentially overlapping hypotheses have been put forward. These include increased transmission rates leading to a much larger number of infections, substantial geographic expansion, and a shift away from a small fraction of infections generating the majority of new cases (e.g. what could be expected with commercial sex workers receiving non-sterile injections) to more homogenous and less frequent transmissions (Duke-Sylvester et al., 2013; Faria et al., 2014; Magiorkinis et al., 2013). Why this only occurred in Group M and not O is also not known, but the increased susceptibility of Group O to tetherin has been put forward as a possible cause (Sauter et al., 2009). Regardless of the reason, Group M now accounts for 95% of infections worldwide (Sharp and Hahn, 2008).

### *Group M Subtypes*

In 1992, about a decade after the discovery of HIV and its identification as the etiological agent of AIDS (Barré-Sinoussi et al., 1983; Gallo and Montagnier, 2003), it became apparent there were at least five distinct lineages, or subtypes, of HIV-1M (Myers et al., 1992). Originally based on *env* sequences, these were designated subtypes A, B, C, D, and E, however, when efforts were made to validate these designations with *gag* sequences, subtype E was not recovered (Myers, 1994). What was originally designated subtype E was shown to be a recombinant virus between subtype A and an ancestral subtype E virus (Gao et al., 1996), an early harbinger of the havoc recombination can wreak on phylogenetic efforts. A 'pure' version of the ancestral subtype E virus has not yet been located and the strains formerly designated as E were reclassified as the first identified circulating recombinant form CRF01_AE (Lau and Wong, 2013).

Currently, subtypes A-K, and over a hundred different circulating and unique recombinant forms are recognised (subtype I was also reclassified as a CRF), with subtypes A and F being further divided into sub-subtypes (Buonaguro et al., 2007; "Los Alamos National Laboratory HIV Sequence Database," 2019; Robertson et al., 2000). CRFs stem from co-infection and recombination by different subtypes of the virus within an individual. Subtypes are roughly phylogenetically equidistant from each other and

genetic variation between them is usually 25% to 35% (Hemelaar et al., 2006; Korber et al., 2001; Taylor et al., 2008). Genetic variation within a subtype is usually around 15% to 20% (Korber et al., 2001), while genetic variation within sub-subtypes is around 6% to 8% (Triques et al., 2000). There are also geographically distinct lineages which, while not phylogenetically equidistant within a subtype, are still phylogenetically distinguishable (Taylor et al., 2008). Examples of geographically distinct lineages include Ethiopian C, Former Soviet Union A, Thai B, and West vs. East African D. The ability to distinguish subtypes, sub-subtypes, and geographically distinct lineages is not just an interesting phylogenetic quirk. Numerous studies have repeatedly shown that HIV subtype affects replicative fitness e.g. (Ariën et al., 2005; Njai et al., 2006), disease progression e.g. (Baeten et al., 2007; Kaleebu et al., 2002), viral load e.g. (Fischetti et al., 2004; Sarr et al., 2005), and transmission rates (Hudgens et al., 2002; Kiwanuka et al., 2009; Sarr et al., 2005).

In addition to this phylogenetic determination, subtypes can also be distinguished by subtype specific positions (SSPs) in the virus' AA sequence ("Los Alamos National Laboratory HIV Sequence Database," 2019; Tenzer et al., 2014). Subtypes and CRFs show a strong geographic distributional 'preference', being more or less prevalent in different regions, for example: Subtype A is common in Central, East, and West Africa, as well as Eastern Europe and Central Asia. Subtype B is the most geographically dispersed strain (Junqueira and Almeida, 2016) and predominates in Australia, Europe, New Zealand, and the United States. Subtype C is prevalent in Ethiopia and Southern Africa as well as India (Table 0.2).

*Table 0.2* Non-recombinant subtypes of HIV-1M, their areas of prevalence, and their prevalence globally from 2010 to 2015. Areas of prevalence compiled from ("Los Alamos National Laboratory HIV Sequence Database," 2019). Percent of global infections from (Hemelaar et al., 2019).

| Subtype | Percent Global Infections | Areas of Prevalence |
|---|---|---|
| A | 10.3% | Central, East, and West Africa; Eastern Europe, Central Asia |
| B | 12.1% | Americas, Australia, Japan, South Korea, New Zealand, Western Europe |
| C | 46.6% | Southern and Eastern Africa; South Asia |
| D | 2.7% | Sudan and Uganda |
| F | 0.6% | Angola, South America |
| G | 4.6% | West Africa |
| H | 0.1% | Central Africa |
| J | 0.1% | Central Africa |
| K | 0.1% | Central Africa |

Due to their presumed clinical relevance and importance to future vaccine efforts, it is unsurprising that the questions of where subtypes come from and why were asked and debated as soon as they were identified (Myers, 1994; WHO network for HIV isolation and characterization, 1994). The idea of HIV coevolution and human host specificity was summarily dismissed early on as well (Myers, 1994). However, this dismissal seems to have been somewhat premature and possibly due to limited data as it also discounted the idea that there was geographic distribution to the subtypes, something that is self-evident today. It was proposed, first in an epidemiological framework (Hu et al., 1999) and then independently in a phylogenetic one (Peeters and Sharp, 2000; Rambaut et al., 2004, 2001), that the global subtype diversity we see is due entirely to random founder effects. Under this hypothesis, there is a pool of circulating HIV strains in the

DRC, and the random exportation of one of these strains to a new region is what causes a new localised epidemic of a specific subtype. Geographic distribution of subtypes then is random and does not indicate any kind of host specificity or coevolution. This idea that global subtype diversity is merely an artefact of chance founder events has since become the dominant paradigm in the field and has changed little in the intervening years e.g. (Archer and Robertson, 2007; Faria et al., 2019; Ferdinandy et al., 2015; Grant et al., 2020; Roques et al., 2002; Tebit and Arts, 2011).

## HIV Biology

HIV-1 is a retrovirus (family *Retroviridae*) belonging to the *Lentivirus* genus. The genome of the most commonly used reference strain, HXB2 (accession number K03455), is 9719 base pairs (bp) long; however, due to the unstable nature of HIV-1's genome there are numerous insertions and deletions in other strains relative HXB2, particularly in the envelope protein ("Los Alamos National Laboratory HIV Sequence Database," 2019). Like other retroviruses, the positive sense single stranded RNA genome encodes three larger polyproteins: Gag, Pol, and Env, as well as a number of accessory proteins and numerous regulatory elements (Figure 0.6a).

The virion (Figure 0.6b) possesses a host derived lipid envelope that retains some host proteins. The collection of these proteins appear to be conserved, indicating a non-random exit pathway and that the uptake of host proteins may not be a passive process

(Burnie and Guzzo, 2019; Linde et al., 2013). The envelope is also studded with *env* derived trimeric glycoprotein complexes comprising gp41, gp120, and extensive (~50% by mass of gp120) glycans (Leonard et al., 1990). HIV-1 is dependent on these Env glycoproteins for cellular entry and changes to it can affect cellular tropism (Arrildt et al., 2012). The *gag* gene encodes structural proteins that make up the matrix, capsid, and the nucleocapsid, while *pol* encodes the replication machinery of the virus: reverse transcriptase and integrase. The viral genome is diploid with each virion containing two strands of RNA.

The mutation rate of HIV-1's RNA genome is similar to that of other non-retro RNA viruses, around $3.5 \times 10^{-5}$ mutations per nucleotide per replication (Sanjuan et al., 2010; Schlub et al., 2014). Differing levels are reported by different groups depending on whether experiments are *in vitro* or *in vivo*, what cell lines are being worked with, and the normal variability of biological experiments. The various estimates work out to about .1 to 1 mutation per replication cycle, with the bulk of this mutation being thought to occur during reverse transcription due its lack of exonucleolytic proofreading (Abram et al., 2010; Cromer et al., 2016; Schlub et al., 2014). Other possible sources of mutation likely include the poor fidelity of RNA polymerase and

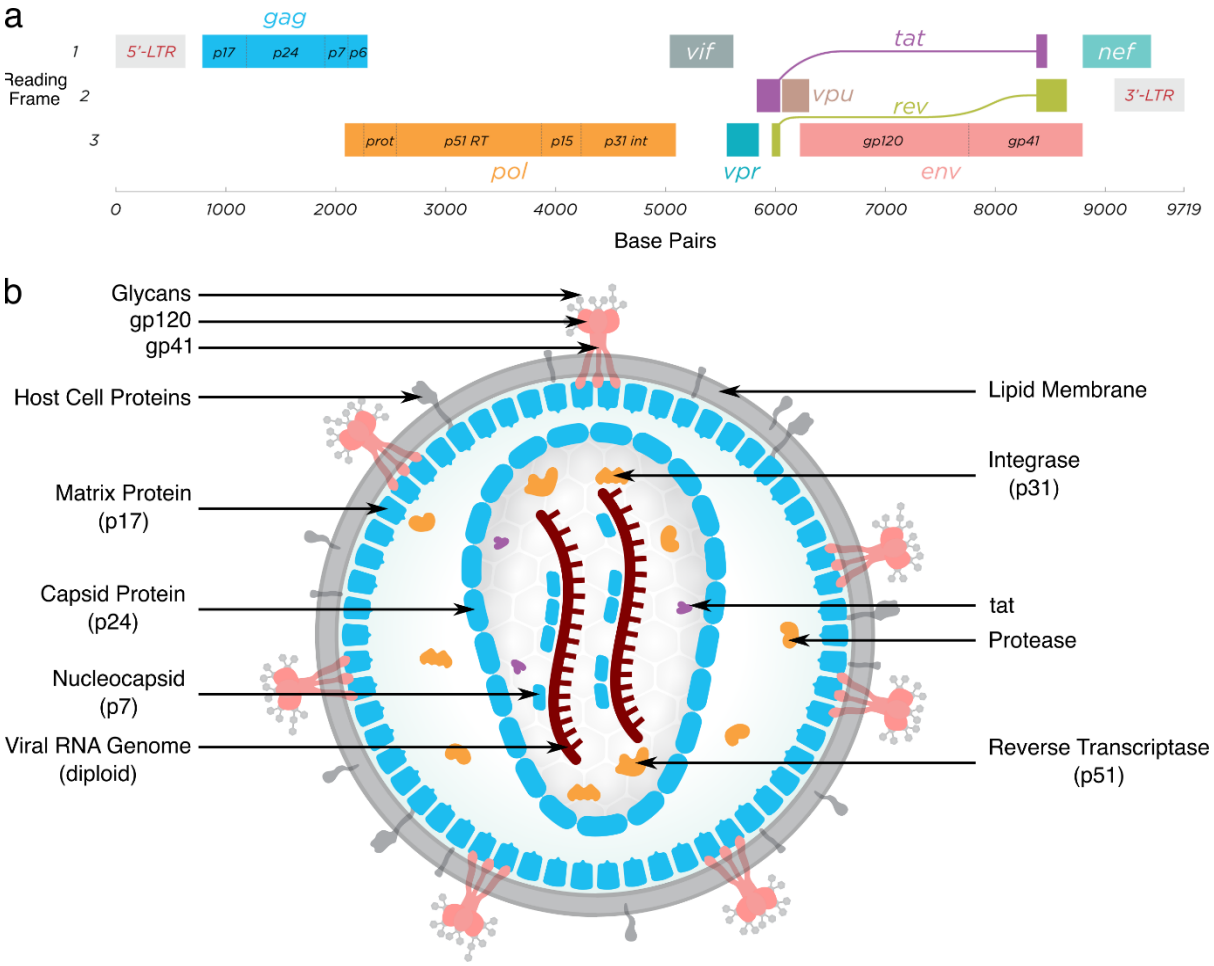



*Figure 0.6* Schematic diagrams of HIV genome (top, **a**) and virion (bottom, **b**). Adapted with permission from (Thomas Splettstoesser, 2014).

host factors such as APOBEC3F increasing mutation rates. Additionally, 15-20% of mutations are directly linked to recombination events, suggesting that the misincorporation of nucleotides increases as the RNA is swapped in and out during strand transfers (Schlub et al., 2014).

Because of this, any given virion's two RNA strands are unlikely to be 100% identical. During reverse transcription the growing DNA strand undergoes an obligatory strand transfer between RNA strands, in a process called minus strong-stop DNA transfer (Basu et al., 2008). A second strand transfer also occurs later during the synthesis of the

complementary (positive sense) strand of DNA. In addition to these obligatory strand transfers, the growing DNA strand can also switch between RNA strands at any point during reverse transcription, and certain sites seem to be preferential break points, promoting strand transfer (Simon-Loriere et al., 2010; Smyth et al., 2014). These preferential sites appear to be mediated by secondary and tertiary structures in the RNA genome itself as well as chaperone activity by the nucleocapsid protein (René et al., 2018; Simon-Loriere et al., 2010). During a superinfection, wherein a cell is infected by multiple different HIV-1 subtypes, progeny virions can be packaged with RNA strands from two different subtypes. Subsequent strand swapping during reverse transcription can lead to recombinant forms of the virus, some of which can establish themselves as CRFs, apparently outcompeting 'pure' subtypes in specific regions (Njai et al., 2006). This can also occur with two viruses from the same subtype, potentially transferring functional mutations such as drug resistance, but this is generally harder to detect.

Like any replicating system, HIV-1 can be visualised as existing in a 'fitness landscape', where height measures the reproductive success of a given genotype and horizontal distance between genotypes is a measure of how closely related they are (McCandlish, 2011; Wright, 1932). HIV-1's rapid rate of mutation, and even more so its recombination, allow for the virus to quickly adapt to changes in the fitness landscape such as a host's Cytotoxic-T-Lymphocyte (CTL) response, moving to a new host with different HLA genes, or the application of antiviral drugs (Burke, 1997). Recombination allows for the escape of 'Muller's ratchet', wherein purely asexual

reproduction leads to the accumulation of deleterious mutations and the 'ratcheting down' of population fitness, and for rapid leaps between local fitness maxima that would otherwise be difficult to move away from (Burke, 1997; Felsenstein and Yokoyama, 1976; Muller, 1964).

Despite this overall propensity for mutation and recombination in HIV-1, different viral proteins appear to have vastly different tolerances for these changes (Crawford et al., 2011; Dahirel et al., 2011; Rolland et al., 2010). Env, the protein most exposed to the host immune response, can tolerate substantial mutations, including relatively large insertions and deletions ("Los Alamos National Laboratory HIV Sequence Database," 2019). Indeed, gp120, the outermost portion of Env (Figure 0.6b), contains five variable loops which can vary in length and amino acid composition affecting everything from cellular tropism to sensitivity to neutralising antibodies (Yuan et al., 2013). In contrast proteins like Gag are much more structurally and functionally constrained, for example the CTL escape mutations required by HIV-C to evade HLA-B*57/B*58:01-restricted responses incur significant costs to viral replicative capacity (Boutwell et al., 2009; Schneidewind et al., 2008). While escape mutations and co-occurring compensatory mutations exist in Gag, broad Gag-specific $CD8^+$ T-cell responses (i.e. ones targeting multiple Gag epitopes) are strongly linked to reduced viral loads (Crawford et al., 2011; Honeyborne et al., 2007; Kiepiela et al., 2007; Rolland et al., 2010; Tarosso et al., 2017; Zuñiga et al., 2006). This has important implications for vaccine design. A proposed HIV-1 vaccine strategy is 'cornering' the virus by targeting Gag epitopes that can

impose fitness costs that cannot be compensated for (Crawford et al., 2011), either because the compensatory mutations structurally destabilise the virus or because they make it vulnerable in other ways that can be simultaneously targeted and exploited.

## Addressing the Pandemic

Patients with access to antiretroviral therapy (ART) or highly active antiretroviral therapy (HAART) have their lifespans dramatically increased, with many surviving long enough to die of non-HIV related causes such cardiovascular disease (Antiretroviral Therapy Cohort Collaboration, 2008; Burchell et al., 2019). Still, even for populations with access, adherence rates are often quite poor, ranging from 27% to 80% depending on the population, falling short of the required 95% adherence (Iacob et al., 2017).

Although great improvements have been made to increase the availability of life saving treatments, in 2018 only ~60% of the estimated ~38 million people living with HIV were accessing any form of antiretrovirals (UNAIDS, 2019). This is a dramatic improvement from earlier years and has finally become a small majority, but further gains will likely be increasingly difficult as many regions lack the healthcare and transport infrastructure to allow effective delivery and dissemination of ART.

Because of these issues, despite the tremendous successes of ART and preventative modalities, a permanent cure remains sought after. An HIV vaccine is likely the best

route to a robust and lasting end to the pandemic (Fauci et al., 2014a). The US Secretary of Health and Human Services famously predicted in 1984 that a vaccine would be ready for testing in approximately two years (Esparza, 2013). This was same year that HIV-1 was discovered to be the etiological agent of Acquired Immune Deficiency Syndrome (AIDS), and only a year after the virus itself had been isolated (Barré-Sinoussi et al., 1983; Gallo and Montagnier, 2003). While after over thirty years without much progress on a vaccine, this statement may seem absurd, it should be noted that she was actually essentially accurate in her prediction (Fauci et al., 2014a). The problem was that those trials, and the many subsequent ones, do not seem to have brought us any closer to an HIV vaccine. Even with decades of research, we appear to remain very much in the early stages of vaccine development (Fauci et al., 2014b).

Significant effort has been spent targeting Env, and there is some evidence for the efficacy of broadly neutralising antibodies (bNAbs), at least *in vitro*, but it is not clear how effective these are at reducing viral load and suppressing viremia in chronically infected patients (Halper-Stromberg and Nussenzweig, 2016; Kumar et al., 2018). The hope is that by acting on Env directly, bNAbs prevent viral entry into cells. However, application of bNAbs rapidly selects for mutations either in Env or compensatory ones elsewhere that confer resistance to the antibodies with only a few amino acid changes in the virus (Halper-Stromberg and Nussenzweig, 2016; Klein et al., 2012). *In vivo*, the number of Env specific $CD8^+$ T-cell responses directly correlates with viral load (Kiepiela et al., 2007), i.e. the more the body is targeting Env the less controlled viremia is. Even

a mixed response to both Env and Gag is worse than Gag alone as Env responses serve to dampen the efficacy of Gag responses (Kiepiela et al., 2007). The variable loops of Env play an important role in the integrity of both the protein and viral envelope and so ultimately there are some structural constraints on Env (Pritchard et al., 2015; Veillette et al., 2015; Yuan et al., 2013). However, there appears to be more structural constraints on Gag than Env (Woo et al., 2010), and these structural constraints are reflected in the different clinical effects of CTL responses towards the two proteins (Kiepiela et al., 2007).

# Part 4

## *Phylogenetics*

> As buds give rise by growth to fresh buds, and these, if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications.
>
> -*On the Origin of Species* (Darwin, 1859)

Phylogenetic methods are used at various points throughout this thesis and so deserve their own overview. Phylogenetics is the study of relationships between evolving entities by modelling their shared ancestries. These 'entities' can range from entire species and more inclusive taxa, down to individuals, and even single genes within an individual, or even languages (Gray et al., 2009). Phylogenies allow us to organise our understanding of and test hypotheses about the evolution of diversity (Williams and Heaps, 2014). Now, phylogenetics is used in one form or another throughout the

biological sciences (Yang and Rannala, 2012). Indeed, as has been pointed out, "nothing makes sense in biology except in the light of evolution" (Dobzhansky, 1964).

Originally, phylogenies were based on morphological features, and used to describe broad taxonomic relationships (Fitch and Margoliash, 1967; Kapli et al., 2020; Sanghvi, 1953). Morphological features can be utilised when looking at extant macroscopic creatures, like the beaks of Darwin's finches, and even fossilized remains. However, the vast bulk of life on Earth is microbial in nature where morphological distinctions are unfortunately less useful. Prokaryotes could be separated from eukaryotes on the basis of the absence/presence of a nucleus, but, for example, the question of what exactly constituted a 'bacterium' and how were they related to blue-green algae (cyanobacteria) was an unresolved question prior to the advent of molecular sequencing (Heather and Chain, 2016; Stanier and Niel, 1962). Some biochemical distinctions could be drawn between microorganisms, but in general only the quantities of constituent elements of a gene sequences could be determined, not the order (Holley et al., 1961; Williams and Heaps, 2014).

The development of methods to determine sequence order in 1965 (Holley et al., 1965; Sanger et al., 1965) started initially just with small RNA fragments, but moved relatively quickly to gene coding regions (Min Jou et al., 1972) and eventually whole (albeit very small) viral genomes (Fiers et al., 1976). With the door open on molecular sequencing, researchers had access to a new source of phylogenetic information, not

just for microorganism but also for multicellular life and individual genes. This rapidly led to shake ups in our understanding of biology, for example breaking prokaryotes into the two distinct lineages of Bacteria and Archaea (Woese and Fox, 1977). One of the greatest benefits of molecular phylogenetics is scientific rigour. The unambiguous nature of the possible character states in sequence data (4 nucleic acid, ~20 amino acids) vs. arbitrarily encoded morphological character states allows for substantially more rigorous statistical analysis (Stevens, 1991; Williams and Heaps, 2014). However, some subjectivity is still present in molecular phylogenetics. While automated alignment systems exist and generally do well for similar sequences, sequence alignments sometimes require manual corrections which are by their nature somewhat subjective.

The availability of molecular sequences, and advances in the statistical and computational methodologies surrounding molecular phylogenetics has greatly broadened its utility. Molecular phylogenetics can be used to sort and classify the numerous different species from metagenomic samples, such as with the Phymm software (Brady and Salzberg, 2009). Phylogenetics can also be used for genome interpretation: identifying genes and non-coding regulatory elements in both microbial (Kellis et al., 2003) and multicellular genomes, including humans (Lindblad-Toh et al., 2011). In medicine it can be used to reconstruct the progression and mechanisms of cancer development and treatment (Schwartz, 2019) or the epidemiological dynamics and evolutionary history of pathogens (Kalish et al., 2004; Kist et al., 2020). An extension of phylogenetics, phylogeography, pinpointed Kinshasa as the epicentre of

the HIV-1M pandemic (Faria et al., 2014). Another extension, phylodynamics, looks at interactions between epidemiological and pathogen evolutionary processes (Grenfell et al., 2004), and can be used to estimate the basic reproduction number $R_0$ (Rife et al., 2017) and the dispersion parameter $k$ (Li et al., 2017) of disease outbreaks.

A phylogenetic tree is a form of acyclic graph (Figure 0.7), meaning that between any two points there is exactly one path connecting them. Internal nodes bifurcate with each internal node having one parent node and two offspring, though in unrooted trees the directionality of this is, by definition, unknown. Terminal nodes, more commonly known as 'leaves' or 'tips' are end points on the graph and are only connected to one other node. The terminal nodes consist of the input data, such as sequences in the alignment, while

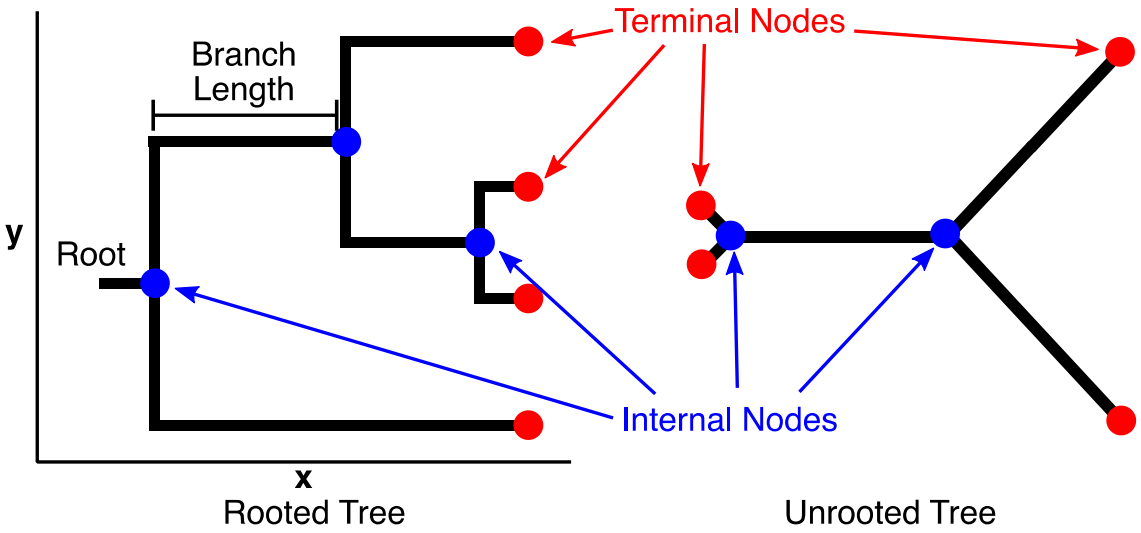



*Figure 0.7* Anatomy of phylogenetic trees. Internal nodes (blue points) are sometimes called "ancestral nodes", "most recent common ancestors", or "hypothetical taxonomic units". Terminal nodes (red points) are commonly referred to as "leaves", "tips", or "operational taxonomic units". Branches are sometimes referred to as edges, and differentiation is made between internal branches connecting internal nodes and external or peripheral branches which connect to tips. For a rooted tree (left), the y axis is arbitrary and meaningless. The x axis can be in substitutions per site or if the tree is dated, the x axis can denote time (e.g. days, years, etc.). Similarly, for an unrooted tree (right) the angle between branches is arbitrary and meaningless, only the branch length matters.

internal nodes are inferred constructs. Each internal node represents a split into new lineages, and each branch is the persistence of a lineage through time. What exactly 'a new lineage' constitutes is dependent on the kind of data being analysed. For a gene family it could represent a duplication event. For a genus of related finches, it represents a speciation event. Similarly, what exactly the metric of time is can also vary. In an undated tree, this 'evolutionary time' is measured in substitutions per site. In a dated tree it can represent units of actual time, such as hundreds of millions of years when looking at early animal evolution (Dohrmann and Wörheide, 2017), or much shorter timescales such as days when examining within-host HIV diversity (Raghwani et al., 2018). Trees are inferred from the input data rather than being directly observed, the final tree selected is the tree that best fits the sequence alignment as determined by the phylogenetic reconstruction methods used to construct it.

### *Distance Based Reconstruction Methods*

Reconstruction methods can be broadly divided into either distance based or character based methods (Saitou, 2018; Yang and Rannala, 2012). For distance based methods, the distance between every pair of sequences is calculated by comparing the aligned sequences, and the resulting distance matrix (Figure 0.8a) is used iteratively to construct a tree (Figure 0.8b) (Kapli et al., 2020). These types of methods include unweighted (or weighted) pair group method with arithmetic mean (Sokal and Michener, 1958), Minimum Evolution methods (Edwards and Cavalli-Sforza, 1965),

a

| | A | B | C | D |
|---|---|---|---|---|
| A | 0 | 2 | 6 | 4 |
| B | 2 | 0 | 6 | 4 |
| C | 6 | 6 | 0 | 6 |
| D | 4 | 4 | 6 | 0 |

b

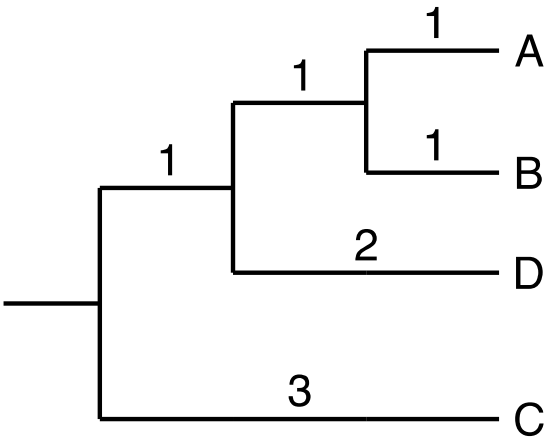



*Figure 0.8* Neighbour joining example of (**a**) a distance matrix and (**b**) its corresponding tree. Numbers in (**a**) are the sum of the branch lengths shown in (**b**). Created with artificial data from (Saitou and Nei, 1987)

Least Squares (Cavalli-Sforza and Edwards, 1967), Minimum Deviation methods (Fitch and Margoliash, 1967), Transformed Distance methods (Farris et al., 1970; Klotz and Blanken, 1981), Neighbour Joining (Saitou and Nei, 1987), and the recently proposed External Edge Elimination (Saitou, 2018). Neighbour joining and its myriad descendent variations such as BIONJ (Gascuel, 1997) are currently the most popular distance methods (Kapli et al., 2020). Neighbour joining is very computationally efficient in selecting the optimal tree within tree space (Saitou and Nei, 1987). This is important, because tree space, that is all possible trees, rapidly becomes enormous as the number of sequences under analysis goes up due to combinatorial explosion. The number of possible rooted trees for three different taxa is three. The number of possible rooted trees for 30 taxa is ~$4.95*10^{38}$, more than the number of stars in the observable universe. At 60 taxa it becomes a trillion-fold more than the number of atoms in the observable universe. While distance based methods tend to be computationally efficient, they often perform poorly when examining sequences with high levels of divergence such as might

be seen in distantly related species (Kapli et al., 2020; Yang and Rannala, 2012). There are several reasons for this. Distance methods are often sensitive to gaps in sequence alignments despite some efforts to address this issue (Bruno et al., 2000). Additionally, most distance methods do not account for the high variance found in large distance estimates, nor do they handle well the large sampling errors often found with highly divergent sequences (Kapli et al., 2020; Yang and Rannala, 2012).

## *Character Based Reconstruction Methods*

Character based methods include maximum parsimony, maximum likelihood, and Bayesian inference. These look at character states within a site (a column of an alignment, Figure 0.9): either one of four possible nucleotides in an NA based alignment, or one of 20 possible residues in an AA alignment. Both types can also include gap/unknown character states as well, and some methods can also incorporate additional data types such as a binary yes/no presence of biochemical markers (Saitou,

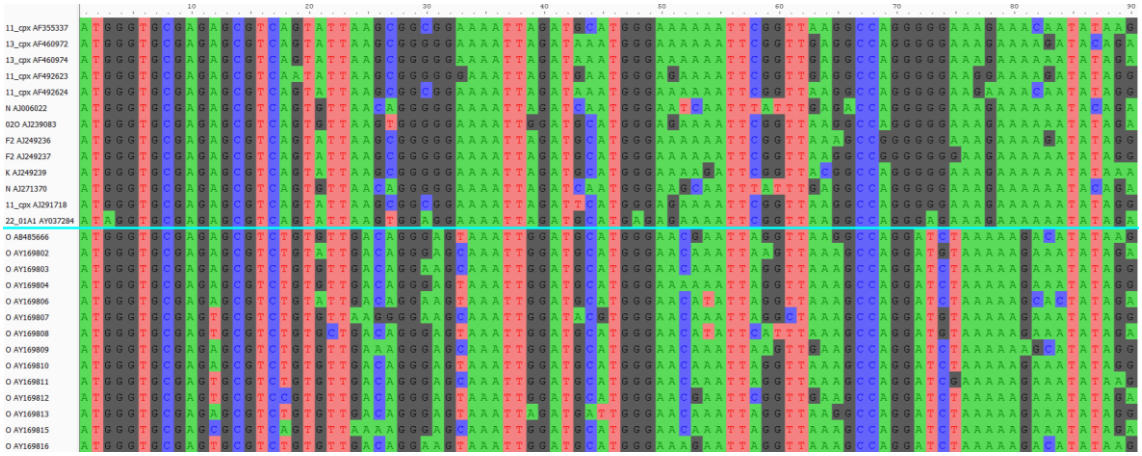


*Figure 0.9* Example alignment of the first 90 sites of HIV *Gag*. Even before tree construction, there are two obvious clades (above and below cyan line). This can also be seen by looking at the sequence subtypes. Those above the cyan line are HIV-1M subtypes while those below are HIV-1O subtypes. Image created in Aliview (Larsson, 2014) from data used in Chapter 2.

2018). From these character states a 'tree score' is calculated for each tree. What exactly this tree score looks like varies between the different character-based methods, but tree scores are used to quantitively compare trees to each other in order to select the best tree. With infinite time and computing power, all possible trees in tree space would be compared and the best possible tree would be selected. As previously discussed under distance based methods, because of combinatorial explosion this becomes infeasible for all but the smallest datasets. Heuristic algorithms are used to find a 'good enough' tree, making character-based methods feasible for larger data sets, even though the likelihood that the true optimal tree has been found is low. These methods will often use a computationally faster algorithm such as neighbour joining to generate a starting tree and then attempt to improve the tree score via local rearrangements of the starting tree (Yang and Rannala, 2012).

### *Maximum Parsimony*

Maximum parsimony methods were originally applied to phylogenetics in the context of morphological data (Camin and Sokal, 1965), but were proposed as being applicable to molecular data almost immediately (Dayhoff et al., 1965). There are a number of variant maximum parsimony methods but all of them are based around the 'principle of maximum parsimony' (Saitou, 2018). Maximum parsimony methods seek to minimise the number of character state changes for a given tree topology, with the tree score in this method being some function of the sum of these changes. Sites that do not change, i.e.

sites that are identical for all sequences in an alignment, and sites where only a single sequence is different ('singletons') are labelled as 'uninformative' and do not factor into tree topology (Saitou, 2018).

The simplicity of maximum parsimony methods facilitates the development of efficient computer algorithms. One of the most appealing aspects of maximum parsimony is it is logically clear and easy to explain and understand (Saitou, 2018; Yang and Rannala, 2012). However, maximum parsimony methods have a number of problems. Maximum parsimony methods are not model based and so lack explicit assumptions incorporating knowledge of sequence evolution. This leads to misleading results if there are significant differences in the rate of evolution amongst different lineages (Felsenstein, 1978). Additionally, when sequence divergence is high, maximum parsimony methods tend to underestimate the total number of changes and has considerable uncertainty with branch length estimation (Saitou, 1989). In light of this, some trees constructed with maximum parsimony methods only show tree topology and forgo showing branch length entirely (Saitou, 2018). Even this is not without risks though as maximum parsimony methods are vulnerable to systematic errors that distort tree topology such as long branch attraction. This is a type of systematic error where if the true topology has two long branches separated by a short internal branch, maximum parsimony methods or model based methods with too simple models will incorrectly group the long branches together. Maximum parsimony methods are particularly susceptible to long branch attraction as they fail to correct for multiple

substitutions at the same site (Felsenstein, 1978). This can still occur in more elaborate maximum parsimony methods that include some in built assumptions or in artificial examples where nucleotide substitution is held constant across taxa (Takezaki and Nei, 1994; Zharkikh and Li, 1993).

### *Maximum Likelihood*

Although commonly attributed R.A. Fisher in 1912, 1921, or 1922, the methods of maximum likelihood were in use prior to these dates, e.g. (Gauss, 1816), albeit under other names such as 'the most probable values of the unknowns' (Hald, 1999). Fisher did however, create the modern version of maximum likelihood between 1912 and 1922 starting with the 'absolute criterion' (Fisher, 1912) and settling in 1922 on the still in use terminology of 'maximum likelihood estimates' (MLEs) (Fisher, 1922; Hald, 1999) with later developments by others e.g. (Wilks, 1938). The method was originally developed for statistics as a method for estimating unknown parameters in a model and was only later extended to working with sequence data (Felsenstein, 1981). The improvements made since this extension have principally been in computational speed up and efficiently searching bifurcating trees, rather than to the essential framework which has remained largely unchanged (Saitou, 2018).

In this framework, for a model parametrised by the unknown parameter $\theta$, the likelihood $L(\theta)$ is the probability of the observed data, that is the alignment, as a function of $\theta$. As sites in the alignment are independent of each other under most

models, the likelihood is the product of the probability of observing the data across all sites (Felsenstein, 1981; Kapli et al., 2020). In other words, the likelihood function is the probability of getting the data seen in the input alignment given as a function of the (unknown) parameters. These parameters θ might include anything from branch lengths of the tree to the various parameters of the substitution model used. Technically, every tree topology is actually a model rather than a parameter, and substitution parameters and branch lengths for any given tree are the parameters of that model. Under this view inferring trees is equivalent to comparing multiple models with identical numbers of parameters (Yang and Rannala, 2012). Regardless, the underlying assumption here is that parameter values that make it highly likely to observe the given data are closer to the true parameter values than ones that make the given data highly improbable. The parameter values that maximise this likelihood are referred to as the MLEs mentioned earlier, and the tree score is the log-likelihood for any given tree (Saitou, 2018; Yang and Rannala, 2012). As with maximum parsimony, given a small dataset or infinite time and computing power for a large dataset the true maximum can be located by exhaustively examining every possible tree. Absent these things, most researchers must settle for a 'good enough' trees, obtained via iterative optimisation algorithms, such as implemented in RAxML (Stamatakis, 2014). These optimisations involve two steps, both optimising the branch lengths for any given tree as well as searching the vast forest of tree space for the maximum likelihood tree.

Maximum likelihood methods have a number of strengths. With large samples, MLEs are

unbiased, consistent, and efficient (Yang and Rannala, 2012). Its model assumptions for sequence evolution are also explicitly stated and can therefore be directly assessed and improved upon. There are numerous different evolutionary models that have been developed for both nucleotide and AA analyses. Methods also exist for examining the fit of these evolutionary models to the data, e.g. the likelihood ratio test (Goldman, 1993). Historically, maximum likelihood was viewed as computationally expensive, particularly the tree searching portion of the optimisation process, which was considered its primary drawback. Although it is still more computationally expensive than distance based methods such as neighbour joining, hardware and software advances have made this increasingly less of a concern. For example, version 3 of the widely used RaxML could handle 1000 taxa on a single processor desktop PC in under 24 hours, a ten-fold increase from earlier programs (Stamatakis et al., 2005). More recent versions making use of parallelisation are even faster (Stamatakis, 2014). An unresolved issue with maximum likelihood is one of interpretability. It is not possible to define a confidence interval for a phylogenetic tree (Yang, 2006; Yang and Rannala, 2012). The bootstrap method is widely used as a replacement in an effort to put confidence limits on phylogenies (Felsenstein, 1985), however actually interpreting what this means has proven difficult despite numerous attempts over several decades e.g. (Berry and Gascuel, 1996; Efron et al., 1996; Felsenstein and Kishino, 1993; Susko, 2010), as well as sometimes controversial, e.g. (Hillis and Bull, 1993). Unsurprisingly given the number of efforts to explain how to interpret the bootstrap, there are multiple largely mutually exclusive explanations (Yang, 2006).

### *Bayesian Inference*

The Bayesian method is named after and originates with Thomas Bayes, an 18th century English statistician. The method was applied to phylogenetics in the late 1990s (Rannala and Yang, 1996) and has become increasingly popular since then with the implementation of various computational packages such as MrBayes (Huelsenbeck and Ronquist, 2001) and BEAST (Suchard et al., 2018). Bayesian inference is a general methodology for statistical inference, not just phylogenetics, and is closely related to maximum likelihood methods. Like maximum likelihood, Bayesian inference relies on both the likelihood function and on explicitly stated models of sequence evolution. However, where maximum likelihood treats model parameters as unknown fixed constants, Bayesian inference treats the parameters as random variables with a statistical distribution (Yang and Rannala, 2012). It uses these statistical distributions to quantify uncertainties within the parameters (Kapli et al., 2020)

The distribution of these tree and model parameters before the data are examined is known as the prior distribution and it represents prior knowledge such as an objective assessments of prior evidence or the researcher's subjective opinion (Yang, 2006). Combining the prior distribution with the data gives the posterior distribution which is the conditional distribution given the data (Felsenstein, 2004; Yang, 2006). The posterior is an update to the prior and is made by rescaling (to form a proper distribution) the product of the prior and the likelihood (Kapli et al., 2020). Inferences

about the tree and model parameters are based on the posterior (Yang, 2006). The basic equation of Bayesian inference in the context of phylogenetics is:

Equation 0.1

$$Prob(Tree,\ Model|Data) = \frac{Prob(Data|Tree,\ Model) * Prob(Tree,\ Model)}{Prob(Data)}$$

More generally, Prob($A|B$) is the conditional probability of observing $A$ given $B$. Prob(Tree, Model | Data) then is the posterior probability of a tree and model given the input sequence alignment. Prob(Data | Tree, Model) is the likelihood and Prob(Tree, Model) is the prior probability of the tree and the evolutionary model. The term 'Model' here encompasses a variety of parameters such as the specifics of the evolutionary model being used (e.g. transition/transversion ratios) as well as branch lengths of the tree. Prob(Data) is a normalising constant equal to the sum of all possible Prob(Data | Tree, Model) × Prob(Tree, Model). Practically speaking, the posteriors of trees cannot be directly calculated because Prob(Data) involves high-dimensional integrals over all possible model parameter values and summation over all possible trees (Yang and Rannala, 2012). However, because Prob(Data) is constant, it cancels out when comparing tree posteriors to one another.

The computation for Bayesian phylogenetics is done using an MCMC algorithm (Figure 0.10) such as the Metropolis-Hastings algorithm (Hastings, 1970; Metropolis et al., 1953; Peskun, 1973). The first MCMC algorithm was developed by Metropolis for use in physics (Metropolis et al., 1953) but was later generalised for other applications

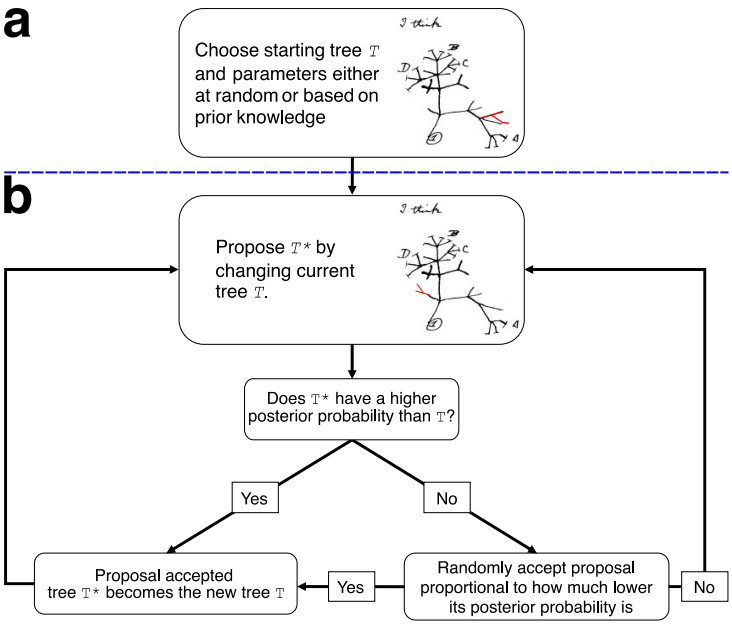



*Figure 0.10* Flowchart diagram of the Markov Chain Monte Carlo algorithm. (**a** above dashed line) A starting tree and parameters are chosen at random or based on prior knowledge. (**b** below dashed line) The main MCMC loop updates the current tree. An identical main loop is also carried out for model parameters.

by Hastings (1970) and Peskun (1973). Despite these generalisations, it was not until much later that MCMC algorithms were shown to be useful in solving nontrivial practical problems (Gelfand and Smith, 1990; Yang, 2006). MCMC algorithms generate a sample using random numbers taken from a target distribution, such as the posterior distribution in the context of Bayesian phylogenetics (Yang and Rannala, 2012). Each sample is an individual tree with its own parameters. The MCMC is used to generate a sample distribution of tree topologies and parameters, moving from one tree or parameter value to another in tree/parameter space, and visiting each one in proportion to their posterior probabilities (Kapli et al., 2020). This frequency then becomes an estimate of the posterior probability for that tree and also serves as the tree score for Bayesian inference methods (Kapli et al., 2020; Yang and Rannala, 2012).

Like maximum likelihood, Bayesian inference uses the likelihood function and so shares many of its strengths such as consistency, efficiency, and numerous well developed evolutionary models to choose from (Yang and Rannala, 2012). The explicit selection of the model allows for the impact of the model on the analysis to be directly tested. Another benefit of Bayesian inference is one of interpretability. Unlike the unresolved problems of confidence interval substitutes discussed earlier for maximum likelihood, the posterior probability of a tree is easy to interpret. Given the data and the model, the posterior is simply the probability of the tree being correct (Yang, 2006). However, "given the data and the model" is not an insignificant caveat. Posteriors are highly sensitive to model selection and misspecification, and small changes in priors can significantly affect posterior distributions (Huelsenbeck and Rannala, 2004; Yang and Rannala, 2005). Posterior probabilities have been shown to be overly credible with some trees showing posteriors of ~100% for nearly all nodes in both simulated and real datasets (Lewis et al., 2005; Suzuki et al., 2002; Yang and Rannala, 2012, 2005).

In light of this, the specification of a priori knowledge can be a double-edged sword. Such information is often not available, and high dimensional priors are notoriously difficult to parameterise, and so most analyses are conducted using default priors in computer programs (Yang and Rannala, 2012). In order to mitigate this, Bayesian robustness analyses can be performed to assess the impact of the prior. Also similar to maximum likelihood, the principal drawback of Bayesian inference is computational intensity. Despite various algorithmic and computational improvements that have been

made, e.g. (Larget and Simon, 1999), MCMC algorithms in particular remain very computationally demanding.

### *Error & Noise*

All phylogenetic methods will have some associated error with them due to models of all kinds being imperfect representations of nature (Box, 1979). These can be errors in the branch lengths of the inferred tree or its topology. These are sometimes viewed as being largely distinct and of differing importance depending on whether one is trying to find divergence dates or taxonomies (Ho and Jermiin, 2004). However, topological inference often relies on branch length (Saitou, 2018), so drawing a sharp line between branch length errors and topological ones is not fully justified. It is true though that for phylogenetic reconstruction, error can be broadly divided into two groups: random sampling errors and systematic errors (Yang and Rannala, 2012). Random sampling errors are uncertainties in parameter estimates due to the finite nature of the data. As the data increase (e.g. the sequence length goes to infinity if the sample size is the number of sites), random sampling errors will decrease towards zero. Systematic errors arise from incorrect assumptions about the model. In contrast to random sampling errors, systematic errors will not only remain as sample size increases, but they actually become worse. This property of solidifying the incorrect conclusion with the addition of more data is sometimes referred as inconsistency or being positively misleading (Felsenstein, 1978). Next generation sequencing technologies are generating enormous

amounts of molecular sequence data leading to huge datasets. This helps to reduce random sampling errors, but simultaneously increases the importance of reducing systematic errors (Yang and Rannala, 2012). The robustness of a model is a measure of its insensitivity to departures from ideal assumption (Box, 1979). Or put another way, a model is robust if it still performs well, giving correct answers when its assumptions are violated. Some assumptions will matter more than others and robustness can be tested via simulation generating data via one model and testing them via another (Yang, 2006). As the amount of data increase, model robustness becomes more important in order to avoid systematic errors.

Regardless of whether the methodology is distance based or Bayesian, estimating the phylogeny requires extracting the true evolutionary signal from input data such as a multiple sequence alignment. Unfortunately, the evolutionary signal will not be the only signal present in any real data. Other signal types also exist: compositional heterogeneity or bias (Gu and Li, 1998; Lockhart et al., 1992), covariotides or covarions (Fitch, 1986a; Miyamoto and Fitch, 1995), multiple substitutions and saturation (Mossel and Steel, 2005; Philippe et al., 2011), invariant sites (Fitch, 1986b; Steel et al., 2000), rate heterogeneity among lineages (Felsenstein, 1978; Steel et al., 2000) and rate heterogeneity among sites (Yang, 1996, 1993). Recombination can also lead to multiple non-concurring evolutionary signals (Edwards, 2009). Ultimately, these are all forms of structured noise. If the noise to signal ratio becomes too great, it can overcome support for the correct tree, leading researchers to incorrect conclusions. These trees can be statistically well

supported due to the magnitude of noise, despite being ultimately inaccurate (Jeffroy et al., 2006). Because these are generally a form of systematic error, additional data will only reinforce these incorrect conclusions (Swofford et al., 1996). This is particularly problematic for deep divergences which may not have much phylogenetic signal left (Mossel and Steel, 2005). The resulting tree in such a situation will often have long terminal branches with numerous multiple substitutions (Philippe et al., 2011), an example of the long-branch attraction error mentioned earlier under maximum parsimony. It should also be noted that for this noise to lead to an incorrect conclusion it needs to not only be larger in magnitude than the evolutionary signal but it also needs to be in an incorrect direction. It is also possible for the noise to increase the apparent support for the correct evolutionary signal, and so while it does not necessarily lead to the wrong conclusion, the presence of these competing signals can cast doubt on inference process (Swofford et al., 1996). Increasingly more sophisticated models try to account for non-phylogenetic signal but at the expense of more parameters and computational time (Philippe et al., 2011).

## *Shortcomings*

> All models are wrong, but some are useful.
>
> -*Robustness in the strategy of scientific model building* (Box, 1979)

Although ever more elaborate models of evolution are constantly being drawn up, all phylogenies eventually fail to fully encompass and account for the complexity of the real world. The acyclic and bifurcating graphical nature of the default phylogeny is

amenable to computation and 'tree-thinking' is intuitively easy to grasp, but this structure is fundamentally ill equipped to deal with messy biological concepts such recombination and introgression. In these processes, branches or parts of branches would be reconnecting downstream and violate the acyclicity of a phylogenetic tree. Because of these processes, gene trees and species trees will not always match up, and so efforts have been made to define 'core genes'. Resulting trees should be interpreted as a 'prevailing trend' rather than the full phylogenetic history (Philippe and Douady, 2003; Wolf et al., 2002). Phylogenetic networks (Wang et al., 2001) and heat maps (Bapteste et al., 2005) have been suggested as possible alternatives to trees, but have yet to supplant phylogenetic trees on a large scale. Even if the trees are imperfect, they still sometimes yield useful fruits.

## *This Thesis*

In Chapter One, I examine the relationships between a class of transposable elements known as polintons, the satellite viruses known as virophages, giant viruses, and their eukaryotic hosts. I uncover a number of new polinton like elements that potentially call into question the monophyly of the two classic polinton clades, and find evidence answering the longstanding question of how homing endonuclease genes (HEGs) jump between species and disparate phyla. In Chapter Two, I show how the introgression of archaic hominid (Neanderthal, Denisovan, and 'Other') HLA genes into anatomically modern humans is granularly adapted to locations. This indicates there was likely a

substantial survival advantage to picking up preadapted immune genes from the archaic hominids when modern humans moved into Eurasia. In Chapter Three of this thesis, I show the link between human genetic diversity and HIV genetic diversity. This strongly calls into question the dominant paradigm in HIV research that HIV diversity is principally an artefact of founder effects (Ferdinandy et al., 2015; Korber et al., 2000; Rambaut et al., 2001).

# Chapter 1
# Virophages, Polintons, Giant Viruses, & their Cellular Hosts

## Overview

My research here is on the phylogenetic relationships between the transposable elements (TEs) known as polintons/mavericks (hereafter polintons), polinton-like elements, virophages, giant viruses, and their host organisms. Viruses do not leave behind mineralised fossils, but they can become heritably embedded in their host organisms' genomes, forming endogenous viral elements (EVEs). Utilising a variety of bioinformatics tools, these viral fossils can then be found in the rapidly expanding available genomic data. This report will provide a brief background on the relevant biology and literature before presenting my results, hypotheses, and future avenues of research. This work was done under the supervision of Aris Katzourakis.

# DISCOVERY OF POLINTON-LIKE ELEMENTS IN WHOLE GENOME SHOTGUN DATA

## Introduction

Transposable elements (TEs) are found in both prokaryotes and eukaryotes where they play major roles in the evolution of their host genomes. Like viruses, they can act not just as parasitic selfish genetic elements (elements that further their own evolutionary interests at a cost to the individual/genome bearing them), but can also serve as sources of novel genes, regulatory elements, and genomic rearrangement (Bourque, 2009; Feschotte and Pritham, 2007; Forterre, 2006; Frost et al., 2005; Kazazian, 2004; Koonin, 2016; Koonin et al., 2006). TEs can be broadly divided into retrotransposons and DNA transposons. DNA transposons are further divided into 'rolling circle' and 'cut and paste' transposons which are dependent on the host's replication machinery for DNA synthesis (Craig, 1995; Kapitonov and Jurka, 2001), and self-synthesizing DNA transposons known as mavericks/polintons (hereafter polintons) (Yutin et al., 2014). Like other TEs, polintons are further subdivided into autonomous (encoding the full set of proteins necessary to replicate itself) and nonautonomous (being dependent on an autonomous relative for anywhere from one to all proteins) varieties. The name

polinton is derived from the polymerase B (POLB) and retroviral-family integrase found in autonomous members of the group. Polintons range from 9 to 40 kilobase pairs (kbp), although most autonomous members are 15-22 kbp (Bao et al., 2015; Haapa-Paananen et al., 2014). They are flanked by terminal inverted repeats ranging from a few hundred bp to over seven thousand bp, and are thought to encode up to 20 proteins (Bao et al., 2015; Fischer and Suttle, 2011).

Interestingly, polintons are related to the viral family *Lavidaviridae* (and its metagenomically identified putative members), sharing varying amounts of genes with different members of that family (Desnues et al., 2012; Fischer and Suttle, 2011; Yutin et al., 2013b). More commonly known as virophages, *Lavidaviridae* are satellite viruses of the giant viruses in the putative order *Megavirales*. Although polintons and polinton-like elements (PLEs) are widely distributed throughout *Eukarya*, giant viruses (and by association, virophages) are only known to infect single celled eukaryotes: amoebae, heterokonts, and haptophytes. The widespread distribution of polintons has been taken as evidence for an ancient origin of the elements (Kapitonov and Jurka, 2006), while their relation to the virophages has spawned several hypotheses as to both groups' origins. One theory suggests that virophages are escaped polintons that acquired a capsid protein through recombination with an unknown virus (Yutin et al., 2013b). Others have suggested that they may have arisen from an ancient mavirus-like virophage ancestor that became able to integrate and transposition itself intragenomically (Fischer and Suttle, 2011; Katzourakis and Aswad, 2014). In the latter

case polintons would be endogenous viral elements (EVEs), vertically inherited 'viral fossils' that are formed by the heritable horizontal gene transfer (HGT) from a virus to a host's germline cells (Katzourakis and Gifford, 2010). A third theory suggests that both groups originated from tectiviruses in bacteria which became 'polintoviruses' (polintons with a putative capsid protein) via the acquisition of integrase from extant eukaryotic TEs (Koonin et al., 2015; Krupovic and Koonin, 2014). These putative polintoviruses then diversified into polintons, virophages, bidnaviruses, adenoviruses, cytoplasmic plasmids, and the entire order *Megavirales* through the gain and loss of integrase, capsid proteins, and other genes, as well as genome expansion (100+ fold in the case of some *Megavirales* members) and reduction in various lineages (Koonin et al., 2015; Krupovic and Koonin, 2014). Exogenous polintoviruses have never been observed, and the putative capsid proteins they are predicated on do not align at the amino acid (AA) or nucleic acid (NA) level, instead being classified based on predicted protein folding (Koonin et al., 2015).

Numerous transcriptionally active genes of putative virophage origin (apparently related to mavirus), dubbed virophage like elements (VLEs) have been found incorporated into the genome of the unicellular alga *Bigelowiella natans* (Blanc et al., 2015). No known giant virus has been observed in *B. natans*, but the VLEs are argued to be evidence of a historical encounter (Blanc et al., 2015). Additionally, the flagellate *Cafeteria roenbergensis* is host to the giant virus *Cafeteria roenbergensis virus* (CroV), itself host to mavirus virophage. The de novo integration of mavirus into the unicellular

flagellate *C. roenbergensis* has been experimentally observed (Fischer and Hackl, 2016). Unlike in *B. natans* these provirophages remained quiescent unless the cell was infected with CroV, which resulted in the production and eventual release of mavirus virions. Both cases indicate an evolutionary scenario in which a mavirus-like ancestor could have become integrated into a host genome and evolutionarily maintained long enough to become fixed in the host population as EVEs. Nevertheless, the debate as to the origins of polinton, PLEs, and virophages remains unresolved.

Here we searched for the presence of VLEs and polintons in metagenomic and whole genome shotgun (WGS) data. We report the identification of a novel group of PLEs in the genome of the coral endosymbiotic dinoflagellate *Symbiodinium minutum* (SMPLEs). Our analysis also indicates that other transposons previously labelled as polintons found in oomycetes (water moulds) possibly cluster with other previously unidentified water mould PLEs outside of the two canonical polinton clades.

## Results

**Virophage hits in WGS.** The full complement of predicted proteins from 17 isolated and metagenomically identified virophages was used to probe the NCBI eukaryotic WGS database using tblastn. In addition to virophage ORFs, ORFs from metagenomically identified putative virophage/polinton hybrids (Yutin et al., 2015a), and transpovirons (Desnues et al., 2012) (hereafter referred to as combined virophage),

were used also used as probes. Excluding *B. natans* which has previously been shown to contain VLEs (Blanc et al., 2015), the best and largest number of hits were to the dinoflagellate *S. minutum*. Hits for each protein query in *S. minutum* were collated and sorted yielding 808 unique contigs, 118 of which had multiple hits per contig with an E value < 1e -10 (Supplementary data 1). The largest 30 of these contigs were selected for further examination.

**SMPLEs in *S. minutum*.** Tblastn hits from both combined virophage ORFs and the full complement of annotated polinton ORFs from RepBase (Bao et al., 2015) against the retrieved contigs clustered together. In the examined contigs, seven full length SMPLEs were found, being defined as having a terminal inverted repeat (TIR) at both ends (purple blocks, Figure 1.1). In the other contigs examined the SMPLEs were truncated artificially by the end of the contig. The regions where hits clustered together in both full length and artificially truncated contigs had a lower GC content relative to the rest of the contig (Supplementary Figure S1.1). Interestingly, mapping *S. minutum* transcriptome shotgun assembly (TSA) data to contig BASF01000492 indicated that the SMPLE makes up approximately 95% of the space between two *S. minutum* exons (cyan arrows, Supplementary Figure S1.1). TSA data did not map to regions around the SMPLEs in the other examined contigs.

ORFs over 300 bp long and located between the TIRs were extracted from the SMPLEs and reciprocally blasted using tblastn against the other SMPLEs. Tblastn hits were then

mapped to the SMPLEs in order to determine which genes were homologous between them (coloured arrows, Figure 1.1). Several of the putative reconstructed ORFs that showed homology but were truncated to fewer than 300 bp by frameshifting mutations and other internal stop codons were also discovered this way. These 'reconstructed ORFs' were also reciprocally blasted between all SMPLEs but this did not reveal any new truncated ORFs. We performed a conserved domain database (CDD) search on each of the homologous ORFs (Marchler-Bauer et al., 2015). Eight of them returned results (labelled arrows, Figure 1.1): Aldehyde:Ferredoxin Oxidoreductase super family (SF) (1.63e-03), COG6 SF (2.67e-03), DNA Topoisomerase I SF (4.24e-03), DNA-Cytosine Methylase SF (1.44e-11), GIY-YIG Homing Endonuclease Gene (HEG) SF (2.65e-11), P-loop containing Nucleoside Triphosphate Hydrolase SF (3.02e-04), Pox A32 SF (FtsK-HerA ATPase packaging protein) (4.9e-04), RVE (retroviral) Integrase SF (3.41e-07), UvrD-like helicase C-terminal domain SF (2.00e-03). The largest ORF in the SMPLEs returned both the UvrD-like helicase C-terminal domain and P-loop containing Nucleoside Triphosphate Hydrolase SF hits. Other homologous ORFs returned no results in CDD.

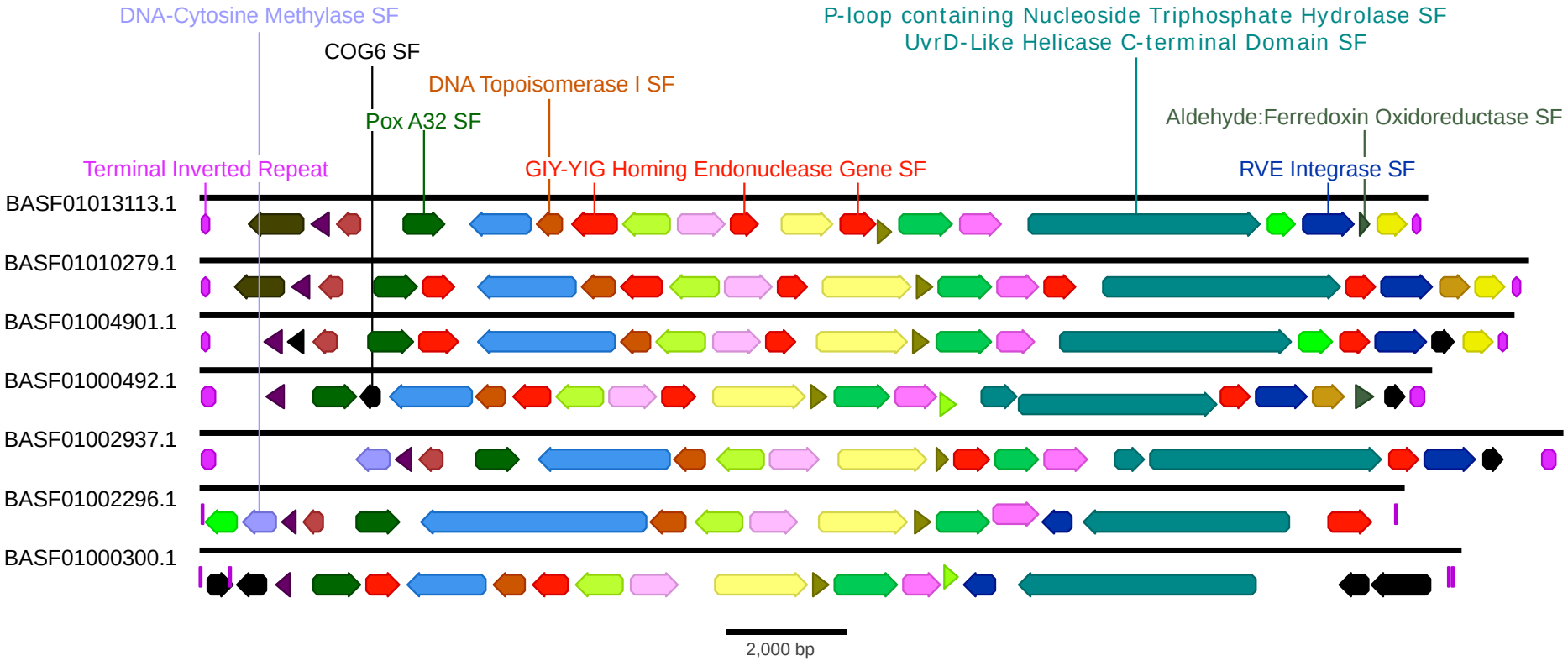



*Figure 1.1* Organization of seven SMPLEs. The position and strand orientation for each ORF (≥300 bp) and homologous reconstructed ORF is indicated by an arrow. Homologous ORFs across SMPLEs are indicated with the same color. ORFs with no homologs in the other SMPLEs are shown in black (including COG6 SF). Terminal inverted repeats are purple boxes. Labels on arrows indicate hits from searching the conserved domain database (CDD). Unlabelled arrows did not get any hits in CDD. E values for CDD hits: Aldehyde:Ferredoxin Oxidoreductase SF 1.63e-03, COG6 SF 2.67e-03, DNA Topoisomerase I SF 4.24e-03, DNA-Cytosine Methylase SF (1.44e-11), GIY-YIG Homing Endonuclease Gene SF 2.65e-11, P-loop containing Nucleoside Triphosphate Hydrolase SF 3.02e-04, Pox A32 SF (FtsK-HerA ATPase packaging protein) 4.9e-04, RVE Integrase SF 3.41e-07, UvrD-like helicase C-terminal domain SF 2.00e-03. The large grey green arrow on the right of the figure with multiple hits in CDD is the possible POLB referred to in the text. The ORFs labelled Pox A32 SF are the FtsK-HerA ATPases referred to in the text. SF: superfamily; COG: conserved oligomeric golgi complex; RVE: retroviral integrase.

**Comparing SMPLEs, virophages, and polintons.** We performed a reciprocal tblastn search in order identify candidates for evolutionary analysis from all three groups: polintons, SMPLEs, and virophages. All the individual ORFs and reconstructed ORFs shown in Figure 1. were used as probes against a local database of the genomes of combined virophages and the 100 polintons listed in Repbase (Bao et al., 2015). The best hit in both polintons and combined virophages for each individual ORF were recorded (Table 1.1). For each homologous set of ORFs, the number of instances of best hit to one group or the other was tabulated ('# better hits', Table 1.1). The single best hit in any of the individual ORFs for each homologous set, to both polintons and combined virophages were also compared ('Best hit' and 'E value', Table 1.1). Seven of

*Table 1.1* Best hits from each group of SMPLE ORFs, tblastned against a combined database of polintons, virophages, transpovirions, and the putative rumen polinton/virophage hybrids.

| ORF | # better hits in P. | # better hits in V. | Best hit in | Best Polinton Hit | E value | Best Virophage hit | E value |
|---|---|---|---|---|---|---|---|
| ORF-1 | 3 | 4 | Virophages | Polinton-1 DEu | 3.70E+00 | YSLV3 | 2.70E-02 |
| ORF-2 | 7 | 0 | Polintons | Polinton-1 CB | **4.67E-11** | PgVV | 3.70E-02 |
| ORF-3 | 1 | 0 | Polintons | Polinton-3 HM | 3.20E-01 | Qinghai Lake virophage | 1.20E+00 |
| ORF-4 | 3 | 4 | Polintons | Polinton-1 PSo | **9.79E-08** | YSLV3 | 6.80E-01 |
| ORF-5 | 3 | 4 | Virophages | Polinton-4 PBa | 1.70E+00 | Dishui lake virophage 1 | 4.40E-02 |
| ORF-6/18 | 0 | 21 | Virophages | Polinton-1 SSa | 2.20E+00 | YSLV6 | **1.24E-34** |
| ORF-7 | 3 | 4 | Virophages | Polinton-2 NV | 5.10E-02 | YSLV3 | 1.00E-02 |
| ORF-8 | 0 | 7 | Virophages | Polinton-1 NV | 5.30E-01 | Qinghai Lake virophage | 4.04E-04 |
| ORF-9 | 4 | 3 | Polintons | Polinton-4 SP | 1.00E-03 | Rumen 2|gb|AUXO017923253.1| | 1.40E+00 |
| ORF-10 | 3 | 4 | Polintons | Polinton-1 DGr | 1.20E-01 | YSLV6 | 2.80E-01 |
| ORF-11 | 4 | 3 | Polintons | Polinton-1 NV | 1.00E-01 | Ace Lake Mavirus | 3.40E+00 |
| ORF-12 | 0 | 3 | Virophages | Polinton-4 NV | 5.10E-01 | Dishui lake virophage 1 | 3.10E-02 |
| ORF-13 | 3 | 4 | Virophages | Polinton-3 TC | 2.80E+00 | Qinghai Lake virophage | 6.40E-02 |
| ORF-14 | 0 | 2 | Virophages | Polinton-1 EI | 3.60E-01 | YSLV1 | **2.73E-16** |
| ORF-15/16 | 7 | 2 | Polintons | Polinton-1 TV | 1.37E-04 | Moumouvirus Monve transpoviron | 2.40E-01 |
| ORF-17 | 3 | 1 | Virophages | Polinton-2 TV | 8.30E-01 | Megavirus courdo7 transpoviron | 8.70E-02 |
| ORF-19 | 7 | 0 | Polintons | Polinton-4 NVi | **2.34E-26** | Rumen 2|gb|AUXO017923253.1| | 1.40E+00 |
| ORF-20 | 0 | 2 | Virophages | Polinton-2 ACe | 1.50E+00 | Mavirus strain Spez | 1.00E-01 |
| ORF-21 | 2 | 0 | Polintons | Polinton-2 DEu | 6.10E-01 | YSLV4 | 4.40E+00 |
| ORF-22 | 0 | 1 | Virophages | Polinton-3 PBa | 7.20E-01 | Ace Lake Mavirus | 7.40E-02 |
| ORF-23 | 0 | 1 | Virophages | Polinton-1 PI | 2.00E+00 | YSLV7 | 4.50E-02 |
| ORF-24 | 0 | 1 | Virophages | Polinton-1 TV | 8.80E-01 | Ace Lake Mavirus | 2.40E-01 |
| ORF-25/44 | 1 | 1 | Polintons | Polinton-1 DGr | 5.70E-02 | YSLV5 | 1.60E+00 |
| ORF-26/28 | 0 | 2 | Virophages | Polinton-2 PBa | 4.80E+00 | YSLV7 | **3.34E-24** |
| ORF-27 | 1 | 0 | Polintons | Polinton-2 CB | 1.20E+00 | Qinghai Lake virophage | 2.70E+00 |
| ORF-31 | 1 | 0 | Polintons | Polinton-1 DAn | 3.00E-01 | YSLV3 | 1.80E+00 |
| ORF-32 | 0 | 1 | Virophages | Polinton-1 DR | 3.60E+00 | YSLV7 | 1.10E+00 |
| ORF-34 | 0 | 1 | Virophages | Polinton-2 TV | 2.90E+00 | Zamilon virophage | 4.10E-01 |
| ORF-35 | 0 | 2 | Virophages | Polinton-4 SP | 4.30E+00 | Sputnik virophage 3 | 3.20E-02 |
| ORF-37 | 1 | 0 | Polintons | Polinton-3 PBa | 1.90E+00 | YSLV2 | 3.40E+00 |
| ORF-33/43 | 2 | 3 | Virophages | Polinton-3 PBa | 5.30E-01 | YSLV3 | 1.20E-02 |
| ORF-36/41 | 0 | 3 | Virophages | Polinton-3 TC | 1.50E+00 | Mavirus strain Spez | 1.80E-01 |
| ORF-38/42 | 2 | 0 | Polintons | Polinton-3 NV | **7.58E-23** | YSLV4 | 3.90E-01 |
| ORF-46 | 0 | 1 | Virophages | Polinton-2 DK | 3.90E+00 | Rumen 2|gb|AUXO017923253.1| | 5.80E-01 |

**P.** polintons; **V.** virophages, transpovirions, and the putative rumen polinton/virophage hybrids; **# of best hits in P./V.** number of individual ORFs in all seven SMPLEs that hit in polintons or virophages. E values less than 1.0E-7 are highlighted in red with bold text. **PgVV:** Phaeocystis globosa virus virophage; **YSLV:** yellowstone lake virophage.

the homologous ORFs matched either combined virophages or polintons with low E values (E value < 1.0e-7).

**Phylogenetic reconstruction.** We chose ORF-2 (Pox A32 SF, dark green arrows, Figure 1.) from the SMPLEs as a first candidate for phylogenetic tree reconstruction, as it was one of the seven ORFs which showed high sequence similarity in one group (polintons E value: 4.67E-11), and of these it had the highest similarity hit in the other group (virophages E value: 3.7E-2). Polinton and virophage hits, as well as reciprocal blast against the NCBI nr/nt database showed similarity to the FtsK–HerA superfamily ATPase DNA packaging proteins (ATPase), in agreement with CDD results. The ATPase homologues from SMPLEs, polintons, and combined virophages were also used to query the eukaryotic WGS database which retrieved a previously described green algae PLE in a *Monoraphidium neglectum* contig (Yutin et al., 2015b). The homologues also retrieved a number of water mould contigs and a 922,537 bp contig of the green algae *Gonium pectorale* isolate NIES-2863. SMPLE ATPase also showed sequence similarity to a reconstructed ORF with numerous frameshifts inside the ~30 kbp intron of another dinoflagellate, *Crypthecodinium cohnii* whose unedited mRNA was located in NCBI's nr/nt database. As expected, tblastn of the ATPases against NCBI nr/nt limited to viruses gave numerous hits within the putative order *Megavirales*, and an *Acanthamoeba polyphaga moumouvirus* (moumouvirus) ATPase was selected to serve as an outgroup. Clustal (Larkin et al., 2007) and MUSCLE (Edgar, 2004) were unable to align the full NA or AA sequences of the diverse groups, so

reciprocal blasts were used as starting points for manual alignment. Gblocks implemented in Seaview (Castresana, 2000; Gouy et al., 2009) was used on codon based NA alignments and manually expanded to 393 NA positions for the creation of phylogenetic trees.

Autonomous polintons require Polymerase B (POLB) in order to self-synthesise their own DNA, and integrase in order to insert themselves into a host genome. Therefore, both of these enzymes would be informative for determining the placement of SMPLEs relative to polintons in phylogenetic trees. However, tblastn of identified POLBs against the SMPLEs, and tblastn of all SMPLE ORFs against the NCBI nr/nt database did not indicate a clear POLB candidate in SMPLEs. SMPLE ORF-15/16 (grey green arrows with multiple CDD hits, Figure 1.1) are approximately the expected size, but returned tblastn hits with limited query coverage and relatively high E values from polinton and virophage POLBs (*e.g.* a hit with 8.8E-5 against only 9% of ORF-15/16) and were unalignable at the AA level.

Reciprocal tblastn of ORF-19 from the SMPLEs against the NCBI nr/nt database, showed sequence similarity to integrase, in agreement with CDD results. As with the ATPase, integrase homologues from SMPLEs, polintons, and combined virophages were also used to query the eukaryotic WGS database, although putative integrases have only been identified in two of the virophages, mavirus and the closely related ace lake mavirus, as well as in some of the VLEs in *B. natans*. This query retrieved green algae and water mould contigs with both ATPase and integrase in close proximity (within 15 kbp) to one another which were considered putative PLEs. Tblastn of SMPLE ORF-19 against the *C. cohnii* intron described

earlier retrieved another reconstructed ORF. As expected, tblastn of the integrases against NCBI's nr/nt database, limited to viruses, returned numerous retroviral sequences, and the core domain of murine leukemia virus polyermase was chosen for use as an outgroup. As with ATPase, Clustal and MUSCLE were unable to align the NA or AA sequences of the diverse groups so reciprocal blasts were used as starting points for manual alignment. Gblocks implemented in Seaview was used on codon based NA alignments and manually expanded to 516 NA positions for the reconstruction of phylogenetic trees.

ATPase and integrase alignments were concatenated and used to reconstruct Bayesian inference trees (Figure 1.2). All green algae PLEs (green subtrees, Figure 1.2), water mould PLEs (ochre subtrees, Figure 1.2), and SMPLEs (blue subtrees, Figure 1.2) have both ATPase and integrase. However, not all polintons, virophages, and outgroup viruses have both ATPase and integrase, and two trees were created to account for this. The first tree used moumouvirus as an outgroup, and all 71 tips possessed ATPase with 15 not possessing integrase (Figure 1.2*Figure*A). ATPase is found in all 15 of the examined virophages (sputnik 2 and sputnik 3 are over 99.9% identical to sputnik 1 and not included in the trees) and the VLEs (red subtrees, Figure 1.2). Integrase is only found in two of the virophages and the VLEs. Polintons 1 (upper collapsed purple subtree, Figure 1.2A) has 23 polintons, which all have both ATPase and integrase. Polintons 2 (lower collapsed purple subtree, Figure 1.2A) has 14 polintons with ATPase, all but one of which have integrase. Moumouvirus has ATPase but not integrase. In Figure 1.2A, virophages were basal to all polinton and PLE groups. The two water mould elements previously described as

polintons ('Polinton 1 PI' and 'Polinton 1 PSo') clustered consistently with the other putative water mould PLEs. Polintons otherwise clustered together as has been previously reported. The placement of the green algae PLEs at the base of the polinton/PLE group had low support (28), but the clustering of the two green algae PLEs was well supported (80). Unsurprisingly, the *C. cohnii* reconstructed ORFs clustered with the SMPLEs (blue subtrees, Figure 1.2) but formed an outgroup in that subtree.

The second tree used murine leukemia virus as an outgroup, and all 71 tips possessed integrase with 14 not having ATPase (Figure 1.2B). Polintons 1 (upper collapsed purple subtree, Figure 1.2B) has 33 polintons with integrase, 23 of which also have ATPase. Polintons 2 (lower collapsed purple subtree, Figure 1.2B) has 17 polintons, 13 of which also have ATPase. Integrase (red subtree, Figure 1.2B) is only found in two virophages and the VLEs. Murine leukemia virus has integrase but not ATPase. In Figure 1.2B, the green algae PLEs were basal to the virophages, water mould PLEs, SMPLEs, and polintons. The two water mould elements previously described as polintons again clustered consistently with the putative water mould PLEs. Other types also consistently clustered together, and there was better support for nodes within the integrase tree as compared to the ATPase tree. Besides the placement of green algae PLEs, the major discrepancy between the integrase tree and the ATPase tree was the placement of water moulds and virophages together as a distinct clade separate from SMPLEs and polintons, although the support for this placement was relatively low (57). Unfortunately, the amount of substitution (see scale bars) likely renders both these trees unreliable (see Discussion).

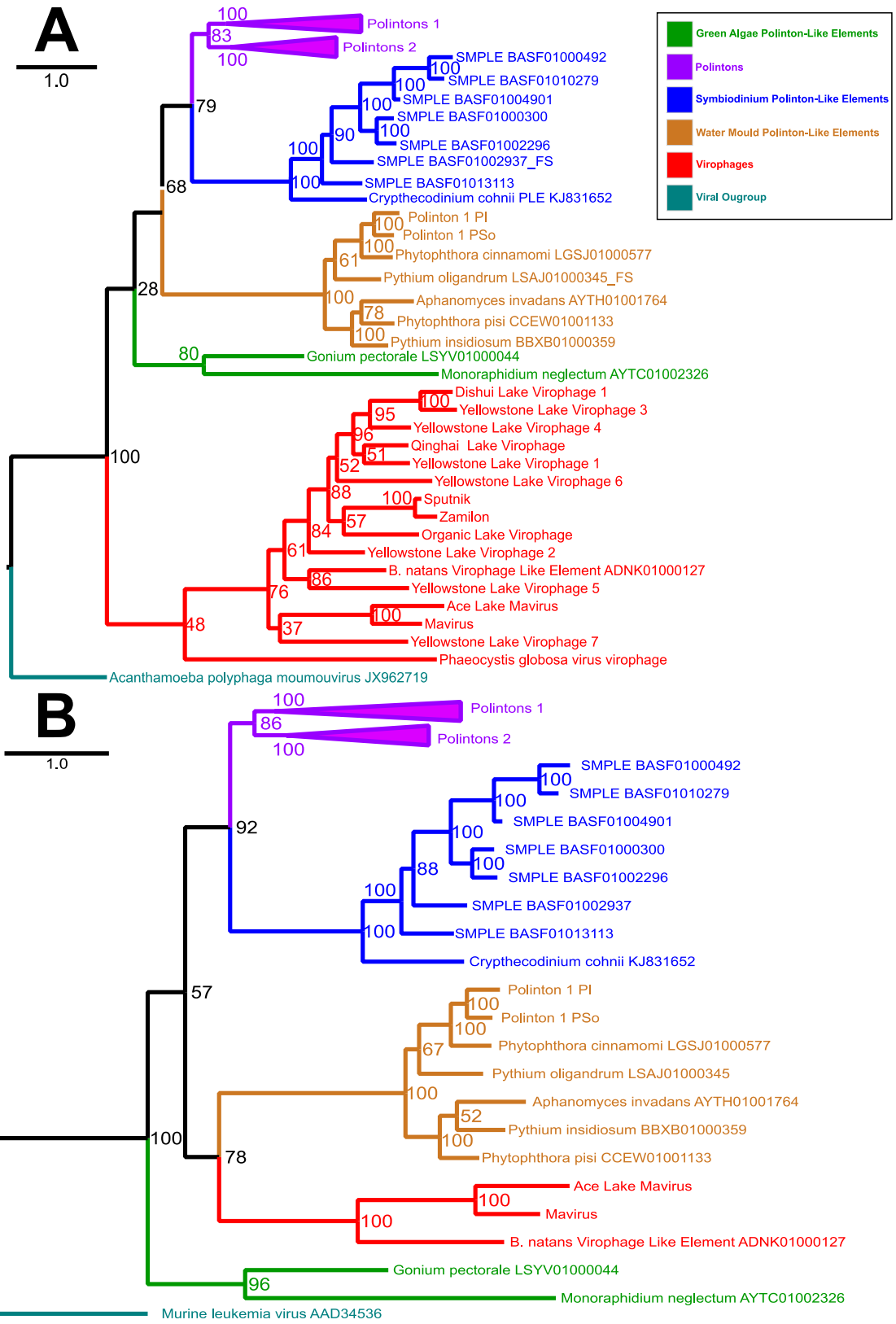



*Figure 1.2* Bayesian inference trees created by concatenating integrase and ATPase and their possible homologues. **A)** All tips have integrase but not necessarily ATPase. **B)** All tips have ATPase but not necessarily integrase. Polintons are shown in purple, green algae PLEs in green, water mould PLEs in ochre, virophages in red, SMPLEs in blue and viral outgroups in teal. Branches that split between groups are in black. The sequence of divergence varies between trees and is not well supported, but different types of PLEs, virophages and polintons cluster together consistently indicating that these are distinct classes of TEs. Polintons are collapsed for clarity. In **A**) there are 23 polintons in Polintons 1 and 14 polintons in Polintons 2. In **B**) there are 33 polintons in Polintons 1 and 17 polintons in Polintons 2. Bayesian inference trees were reconstructed using the SRD06 model (Shapiro et al., 2005). Bayseian posterior probability percentages are shown beside nodes. Scale bar represents the number of substitutions per nucleic acid site.

**ORF-618 phylogenetic reconciliation analysis.** Tblastn of SMPLE ORF-618 showed essentially no sequence similarity to the polintons in the RepBase database, but got hits with low E values to combined virophages (best hit to virophages 1.24e-34 vs. best hit to polintons 2.2e0, Table 1.1). It occurs multiple times in all but one of the examined SMPLEs (red arrows, 1.1). The exact function of ORF-618 is unknown, but a CDD search gave hits (2.65e-11) to the GIY-YIG nuclease domain superfamily.

Given the apparent collinearity of the ORF-618 genes seen in Figure 1.3 one might expect co-speciation between ORF-618 and the SMPLEs. However, phylogenetic reconciliation analysis with Jane (Conow et al., 2010) did not support this (Figure 1.4). A Bayesian inference 'host tree' based on SMPLE ATPase was created, using the putatively homologous reconstructed ORF from the *C. cohnii* intron as an outgroup (black tree Figure 1.4). Tblastn of ORF-618 against the *C. cohnii* mRNA allowed us to extract two reconstructed ORFs with multiple frameshifting mutations and these were used alongside the SMPLE ORF-618s to reconstruct a Bayesian inference tree of the ORF-618 'parasite' (coloured subtrees, Figure 1.4), separate from the SMPLE host tree. None of the collinear blocks appear to have replicated exclusively through co-speciation (symmetrical forks in trees, Figure 1.4), but appear to have replicated substantially via duplication and species switching events (asymmetrical side branches). Interestingly, the pink and green subtrees arise out of the sections of the gold subtree linked to the *C. cohnii* reconstructed ORFs.

**Distribution of ORF-618 in PLEs, polintons, and virophages.** Tblastn of ORF-618 against the eukaryotic WGS database retrieved various water mould contigs as well as two previously described (Yutin et al., 2015b) PLE containing contigs from the cryptomonad alga *Guillardia theta*. Interestingly, mapping ORF-618 against Yellowstone Lake virophage (YSLV) 6 and Organic Lake virophage indicated that the putative genes YSLV6_ORF16 (YP_009177831.1) and OLV1 (ADX05762.1) are both truncated by a frameshift mutation towards the 3' end of the ORF. Tblastn of ORF-618 against viruses (excluding virophages) retrieved *Phaeocystis globosa* virus strain 16T (PgV-16T), a member of *Megavirales* that infects haptophytes (algae related to but distinct from cryptomonads and dinoflagellates). We used these retrieved ORFs alongside virophage ORFs to reconstruct a Bayesian inference tree (Figure 1.5). The two *G. theta* derived homologues clustered together as did those from the retrieved water mould contigs. Congruent with the phylogenetic reconciliation analysis, the *C. cohnii* homologues were split by SMPLE ORF-618s. Surprisingly, the virophage homologues were distributed throughout the tree. Two of the ORFs from YSLV6 clustered together but are separated from the third YSLV6 ORF by SMPLE ORF-618s and a *Phaeocystis globosa* virus virophage (PgVV) homologue. Although not as saturated as the trees created by concatenating integrase and ATPase (scale bars, Figure 1.2), the level of substitution found in the ORF-618 tree (Figure 1.5) also renders it suspect (see Discussion).

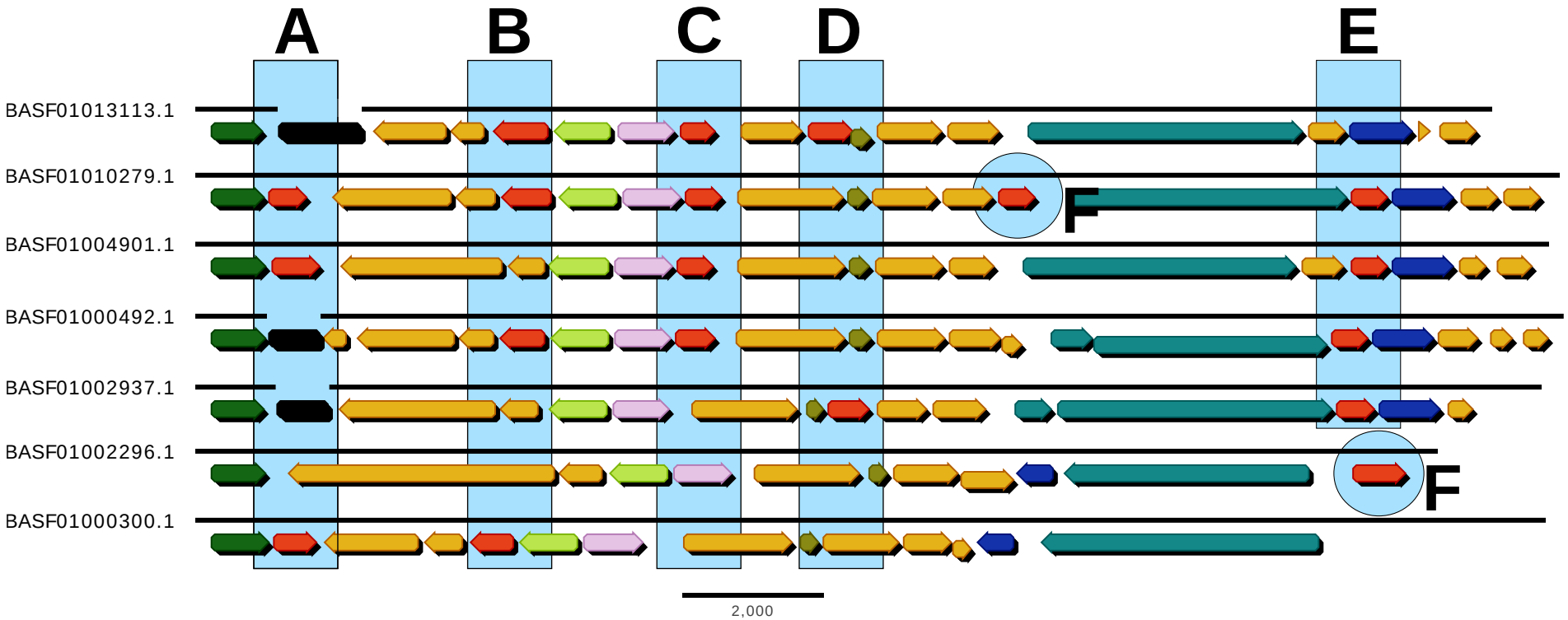



*Figure 1.3* ORF-618 grouped by apparent collinearity. Groups are presumed based on position relative to other nearby ORFs. Red arrows are multiple copies of ORF-618, a putative GIY-YIG Homing Endonuclease Gene. Dark green arrows are putative FtsK–HerA superfamily ATPase DNA packaging proteins. Blue arrows are putative retroviral family integrase. Orange arrows are automatically determined ORFs (≥300 bp). Other arrows coloured to show homology near ORF-618 copies. Black blocks are artificial spaces added to visually align groups.

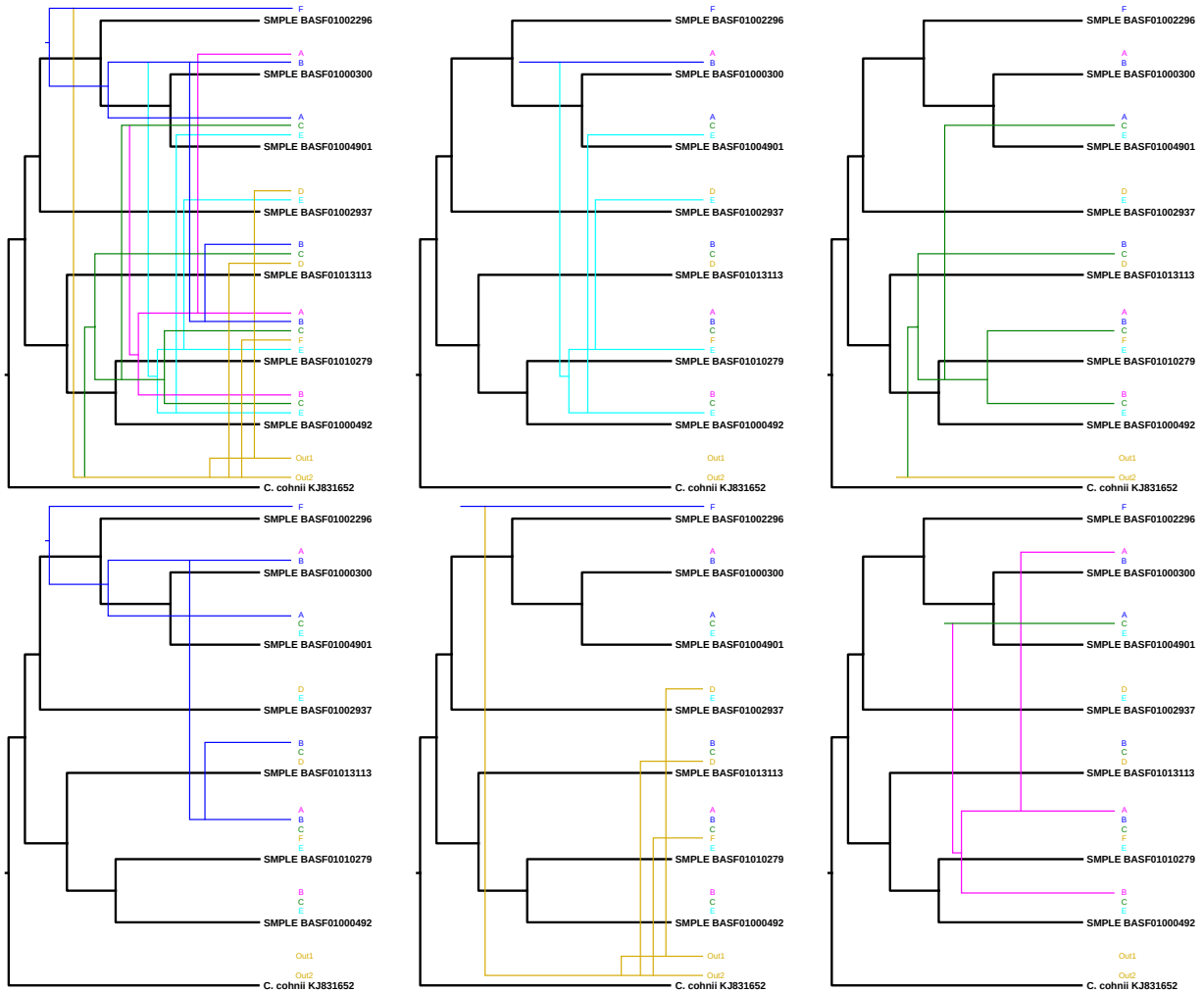



*Figure 1.4* Phylogenetic reconciliation analysis of Bayesian inference trees for host SMPLE ATPase (black) and ORF-618 'parasites' (colours). Host tree is rooted on a reconstructed ORF from a *Crypthecodinium cohnii* dinoflagellate intron. Parasite tree is all SMPLE ORF-618 homologues labelled by grouping in Figure 1.3, plus two reconstructed ORF-618 homologues from the *C. cohnii* intron (Out1 and Out2). Subtrees are shown for clarity. Symmetrical forks indicate co-speciation with the host. Asymmetrical side branches indicate duplication and host switching. Clade C (green subtree) and E (cyan subtree) appear to have been duplicated collinearly. Other subtrees branch across collinear groupings indicating that the HEGs are colonizing new locations rather than only being copied into cognate alleles. The gold subtree is connected to an outside organism and also gives rise to green and pink subtrees, indicating that gene flow is occurring with outside systems. These could be other SMPLEs within *S. minutum*, PLEs from other species, or viruses. Bayesian inference trees reconstructed using the SRD06 model (Shapiro et al., 2005). Reconciliation tree constructed using Jane (Conow et al., 2010).

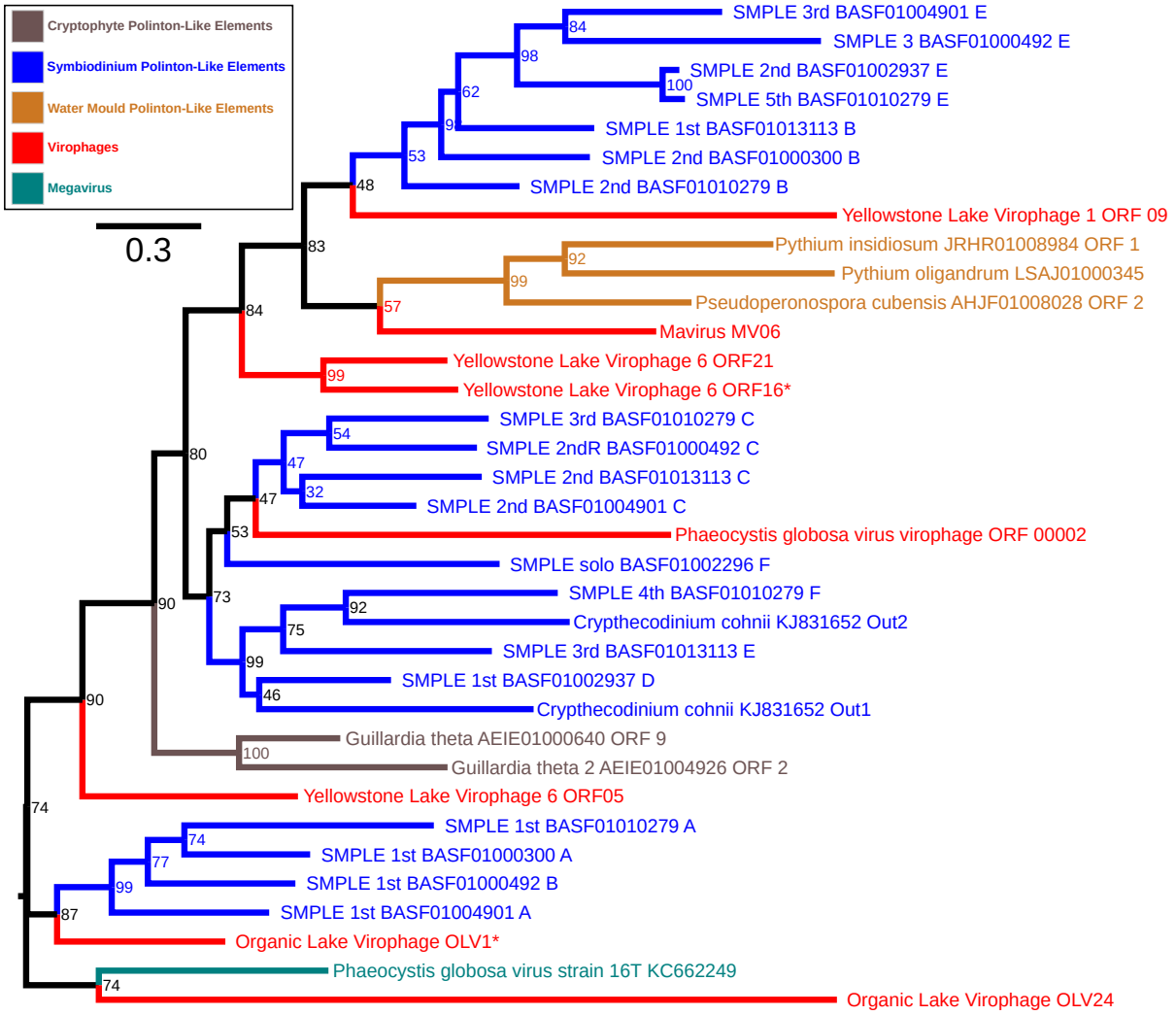



*Figure 1.5* Midpoint rooted phylogenetic tree of SMPLE ORF-618 and possible homologues. The Bayesian inference tree was created from an alignment of conserved regions in polinton like elements (PLEs), contigs retrieved from the WGS database, virophages, and SMPLEs. Cryptophyte examples are shown in grey, water mould examples in ochre, virophages in red, SMPLEs including *C. cohmi* in blue, and *Megavirales* in teal. Branches that split between groups are in black. Bayseian posterior probability percentages are shown beside branches. Asterisks indicate ORFs taken from the NCBI database that possess a previously unidentified frameshifting mutation at their 3' end. Bayesian inference trees reconstructed using the SRD06 model (Shapiro et al., 2005). Scale bar represents the number of substitutions per nucleic acid site.

## Discussion

**Novel PLEs.** Virophages, polintons, and PLEs have conventionally been grouped together into a broad class of selfish genetic elements, the origins of which are highly debated (Fischer and Suttle, 2011; Kapitonov and Jurka, 2006; Katzourakis and Aswad, 2014; Krupovic and Koonin, 2014; Yutin et al., 2013b, 2015b). The majority of known members of this group are endogenous elements within Eukarya (Feschotte and Pritham, 2007; Krupovic and Koonin, 2014; Yutin et al., 2015b). Here we characterised a novel group of elements in the genomes of dinoflagellate endosymbionts of corals and in doing so expanded the known range of this class of elements to alveolates. Interestingly, most of the ORFs in the SMPLEs did not have high sequence similarity to either virophages or polintons, nor did they get high sequence similarity tblastn hits in NCBI WGS or nr/nt databases, underscoring the amount of diversity that exists within this probably overly broad grouping. We have also retrieved previously unknown putative PLE elements from already described classes such as green algae PLEs.

Although there are a number of internal stop codons, internal inversions, rearrangements, non-homologous ORFs, and the varying copy number (between one and five) and locations of ORF-618, the gene order is generally conserved amongst the SMPLEs examined (Figure 1.1), indicating a common origin. Since the SMPLE from contig BASF01000492 was the only one examined that mapped to the inside of mRNAs from TSA, it is likely coincidental that it appears to make up most of an intron. It is

not surprising that some SMPLEs would end up in *S. minutum* introns, as the dinoflagellate's genes are highly intron rich: 95% of its genes have introns with an average of 19.6 exons per gene and some having as many as 200 (Shoguchi et al., 2013). Despite this, the average length of *S. minutum* introns is only 499 bp, substantially smaller than the 20,000+ bp SMPLEs. The fact that a TSA mRNA mapped to a region around the SMPLE indicates that it remains transcriptionally active despite the insertion. This idea is supported by the SMPLE hits from *C. cohnii* being from a larger intron in a sequenced mRNA from NCBI's nr/nt database. However, the draft assembly of *S. minutum*'s genome is only 616 megabase pairs (Mbp) of the estimated 1500 Mbp full genome (less than 42%) (Shoguchi et al., 2013). Therefore the presence of a duplicated functional copy of the gene in the unsequenced portion cannot be conclusively ruled out.

**Saturation and Non-phylogenetic Signal.** Sequence alignments contain multiple signals within them, one of which is the sought after phylogenetic signal used to infer phylogenetic trees (Ho and Jermiin, 2004). However, other signals can arise from phenomena such as variable evolutionary rates across lineages, sites, or positions, and heterogeneous nucleotide/amino acid compositions (Ho and Jermiin, 2004; Philippe et al., 2005). Individual sequence sites can have undergone multiple substitutions leading to inaccurate estimates of genetic distance between two sequences, and if enough of this occurs it leads to saturation, where sequences are effectively random with respect to each other (Mossel and Steel, 2005; Philippe et al., 2011). These are in essence structured noise, which adversely affect the recovery of the phylogenetic signal and can

compete with it in tree reconstruction. For deep divergences, there may be very little phylogenetic signal left (Mossel and Steel, 2005), and resulting trees will often have long terminal branches with numerous multiple substitutions (Philippe et al., 2011). In extreme cases, non-phylogenetic signal may overwhelm phylogenetic signal and yield statistically well supported but ultimately inaccurate trees (Jeffroy et al., 2006). Long branch attraction, when lineages with much longer branches than others in the tree are clustered regardless of any actual relationship, is a well known example of this kind non-phylogenetic signal artefact.

Even before the inclusion of viral elements, the taxa examined in this study are extremely distant from one another, with some of the organisms involved only sharing their taxonomic domain (Eukarya) with one another and so are separated by at least 1.6 billion years (Knoll et al., 2006) of evolutionary time. The multiple alignments created from these distantly related organisms and viruses had to be created manually from a tBlast base, as fully automated methods such as Clustal (Larkin et al., 2007) or MUSCLE (Edgar, 2004) were unable to align the sequences at either the NA or AA level, indicating very divergent sequences and unclear homology. The effects of this are readily observable in Figure 1.2. Even though some of the nodes are well supported, indicating consistent clustering, the length of some of the branches indicates ~6 substitutions per site. Unfortunately, this indicates that the sequences are probably saturated with regards to each other and so the largescale trees are unreliable. Figure

1.5 is better but similarly has problems with saturation, with some of the longer branches indicating ~2 substitutions per site.

The originally discovered canonical polintons, PLEs, and virophages have all been lumped into a broad, ill-defined grouping with fuzzy borders and the nature of the different subgroups' relations to one another is difficult to discern. Previously described green algae and cryptophyte elements (Yutin et al., 2015b) were already been labelled 'polinton-like' as the substantial structural differences (Figure 1.6) and overall lack of homology between their ORFs has made it difficult to justify grouping them with the originally discovered canonical polintons, which are themselves divided into two distinct groups. Elements identified in this study belonging to these already described PLE clades as well as the novel SMPLEs found in alveolates have also been labelled under the catch-all term PLE. Some members of the canonical polintons, such as those elements found in water moulds, were identified as PLEs in this study due to having structural differences from the other canonical polintons e.g. size discrepancies, presence/absence of TIRs, polymerase, and GIY-HEG.

While there are clearly connections between them, there are also significant structural differences between these various related clades (Figure 1.6). The virophage clade (except PgVV) unambiguously have capsids whereas the other clades do not. Only two known virophages have TIRs whereas all canonical polintons, SMPLEs, and cryptophyte PLEs have TIRs. Only some water mould PLEs have TIRs and neither the

previously described green algae PLE nor the one identified in this study possess them. The size of the TIRs also varies substantially between and within clades, from 51 bp in virophages, to a few hundred bp in SMPLEs, to hundreds or thousands of bp in canonical polintons. SMPLE and cryptophyte PLEs also lack the unambiguous POLB found in the other described clades. Similarly, integrase is found in the SMPLEs, water mould and green algae PLEs, and autonomous canonical polintons, but is only found in two virophages and has not been clearly identified in the cryptophyte PLEs. While the SMPLEs are generally syntenic with one another, canonical polintons found in the same host species are not always syntenic with one another.

Drawing a unified tree and trying to determine a discrete origin for polintons, the various disparate kinds of PLEs, and virophages may be impossible due to insufficient phylogenetic signal surviving the deep gulf of time separating modern available tips from this hypothetical root. It may also be impossible due a fundamental flaw in the assumption that there is a single unified tree in the first place. Although the importance of HGT in eukaryotes is debated, it appears to occur quite frequently (but not exclusively) in microorganisms (Bapteste et al., 2005; Huang, 2013; Philippe and Douady, 2003; Wolf et al., 2002). Even when "core" genes are defined, any resulting tree should be interpreted as a "prevailing trend" rather than the full phylogenetic history (Philippe and Douady, 2003; Wolf et al., 2002).

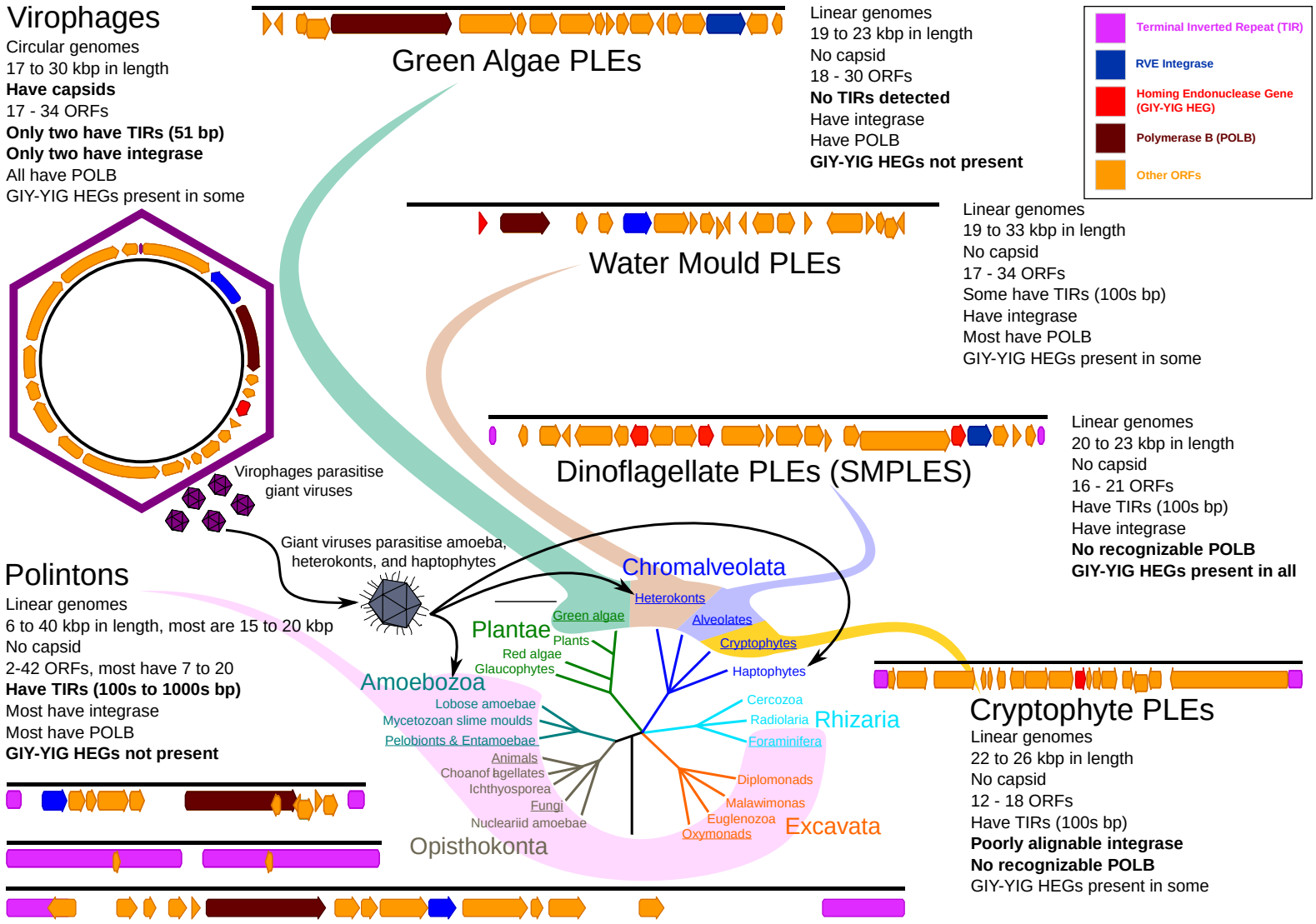



Figure 1.6 Distribution of polintons and PLEs in the eukaryotic tree of life. Polintons (light pink shaded region) have been identified in Amoebozoa, Opisthokonta, and Excavata. Hits to key polinton proteins have been identified in Rhizaria. Polintons have also been described in water moulds (heterokonts) but these are likely water mould PLEs rather than canonical polintons. PLEs have been identified in algae (green shaded region), water mould heterokonts (tan shaded region), dinoflagellate alveolates (blue shaded region) and cryptomonad algae (yellow shaded area). Underlined text indicates branches where polintons/PLEs have been described or detected. Virophages parasitise giant viruses which themselves parasitise amoebae, bicosoecid heterokonts, and haptophytes. There is significant morphological diversity between and within the various groups in terms of presence or absence of capsids and TIRs, and also the size of the TIRs when present. In categories where TEs are known from multiple species (polintons, algae, and water mould PLEs) gene order and presence is often poorly conserved between species. The broad but heterogenous distribution and variations within polintons and PLEs suggest that they are not a monophyletic feature of their hosts. Eukaryotic tree and putative root adapted from (Simpson and Roger, 2004).

HGT is not limited to transfers between cellular organisms. Genes can go from virus to host in the form of proviruses and EVEs, and exogenous viruses can recombine with and acquire genes from these endogenous elements (Chakrabarti et al., 1994; De Paepe et al., 2014; Golovkina et al., 1994; Pandey et al., 1991; Rasmussen, 1997; Sheets et al., 1992). Nor is HGT restricted to the transfer of individual genes. Virally mediated HGT of entire TEs between species has been shown to occur with insect transposons in baculoviruses (Gilbert et al., 2014), and similar process could have occurred multiple times between the PLEs, giant viruses, and virophages. Modern *Symbiodinium* are endosymbionts of cnidarians (known hosts of polintons), living at potentially extremely high densities within specialised compartments inside host cnidarian cells (Stimson et al., 2002; Wakefield and Kempf, 2001). Many dinoflagellates are endosymbionts, and these conditions would give ample opportunity for cross species transmission on evolutionary time scales.

Based on ORF-618's distribution within PLEs, as well as within giant viruses and virophages infecting disparate group (Figure 1.7), it appears to be moving horizontally in and out of both PLEs and viruses. ORF-618 and its homologues in other hosts are putatively members of the GIY-YIG nuclease domain superfamily, part of a functional grouping of genes known as homing endonuclease genes (HEGs). HEGs are mobile genetic elements that encode a single protein which acts on nucleic acids, and which exhibit traits similar to full length TEs (Belfort et al., 1995; Dunin-Horkawicz et al., 2006; Gimble, 2000). HEGs can reproduce intragenomically through homing, a process

wherein the HEG is copied from a donor to a HEG-less cognate allele (Dunin-Horkawicz et al., 2006). However, homing can also be induced by any event that causes a double stranded break (Eddy and Gold, 1992). The ability to home to any double stranded break, as opposed to just cognate alleles, may help clarify why HEGs seem to be able to transmit horizontally between distantly related phyla and also offers an explanation of how ORF-618 colonised new loci within the SMPLEs (Figure 1.3). Our phylogenetic reconciliation analysis showed that only the collinear blocks C and E (green and cyan subtrees, Figure 1.4) appear to have been duplicated strictly collinearly into cognate alleles. Despite this, both arise from another subtree (gold and blue respectively), indicating that they originally colonised the site from a different non-cognate allele, and in the case of clade C this appears to have arisen outside the SMPLEs examined.

The ability of HEGs to move to new loci likely also explains the presence of putative homologues in viral genomes and other PLEs (Figure 1.5, Figure 1.6). The colonisation of viral genomes also offers a possible solution to the standing question over the mechanism/vector of HEG horizontal transmission. If HEGs function similarly to larger TEs, then like insect TEs in baculoviruses (Gilbert et al., 2014), they too could be using the genomes of exogenous viruses to transport themselves to new eukaryotic genomes. An unknown cosmopolitan megavirus has been shown to be associated with *Symbiodinium* (Correa et al., 2016), and megaviruses can carry ORF-618 homologues which gives a possible mechanism for how the HEGs entered

SMPLEs. There is also the possibility of an unknown virophage associated with the giant virus, although no evidence for this has been reported at this time.

The putative HEG seen in PgV-16T appears to be more closely related to Organic Lake virophage protein OLV24 than it is to the putative HEG found in PgVV (Figure 1.5), the plasmid like virophage that parasitises PgV-16T. This apparent discrepancy could be an artefact of non-phylogenetic signal and saturation. If it is not simply a systematic error, then it may be explained by the fact that 75% of the best matches for the predicted proteins in PgV-16T are to the two metagenomically identified viruses named Organic Lake phycodnaviruses (OLPV) 1 and OLPV2 (Santini et al., 2013). OLPV1 and OLPV2 have been putatively reclassified as belonging to *Mimiviridae* (Yutin et al., 2013a), rather than *phycodnaviridae*, and it is assumed (but not demonstrated) that they are the hosts for the metagenomically detected Organic Lake virophage. The presence and movement of the HEGs laterally between viruses and distantly related eukaryotic hosts further demonstrates that HGT of these TEs between species is possible. This is important for explaining some of the discrepancies seen in the phylogeny of polintons and PLEs, such as the non-basal relation of the single known fungi polinton to those found in animals, or the possible transfer of a proto-SMPLE/polinton to an early metazoan.

The discovery of SMPLEs brings the range of PLEs and polintons one step closer to fully encompassing all the major tips of the eukaryotic tree of life (Figure 1.6). The

identification of well conserved ORFs (other than the laterally transferring HEGs) in cryptophyte PLEs would allow useful comparisons between them and the other lineages described here (virophage, canonical polinton, green algae, water mould, and dinoflagellates). Overall, our work helps illustrate the complexity of the evolutionary relationship between virophages, giant viruses, TEs, and eukaryotic hosts. This work also shows that the nomenclature is imprecise, as the catch all term PLE groups together elements that show as much homology (or lack thereof) between each other as they do with the canonical polintons. All of these disparate elements, including the viral members, could be labelled as "polinton like" with some degree of accuracy. It would be more accurate to label them all as polinton related elements (PREs), for backwards compatibility with currently extant literature, and then create subcategories based on hosts or general type e.g. metazoan PREs or viral PREs. Trying to draw a single unified tree for the entire PRE system is likely futile as no single part (gene or otherwise) appears to be universal to all members of the group (Figure 1.6), making defining a "core" impossible. Trying to draw gene networks based on gene sets (see Future Research) would likely be more evolutionarily supported.

## Methods

**Sequence analysis.** 17 full or partial genomes of previously identified virophages were downloaded from the NCBI database: Ace Lake mavirus (partial genome) (KC556923.1), Dishui Lake virophage 1(KT894027.1), mavirus(HQ712116.1), Organic Lake

virophage(HQ704801.1), *Phaeocystis globosa* virus virophage (NC_021333.1), Qinghai Lake virophage (KJ854379.1), sputnik 1 (EU606015.1), sputnik 2 (JN603369.1), sputnik 3 (JN603370.1), YSLV1 (KC556924.1), YSLV2(KC556925.1), YSLV3 (KC556926.1), YSLV4 (KC556922.1), YSLV5 (KM502589.1), YSLV6 (KM502590.1), YSLV7 (KM502591.1), and zamilon (NC_022990.1). The contigs containing putative polinton/virophage hybrids (AUXO017923253.1, AUXO016180925.1, AUXO014440522.1, and AUXO011225015.1) as well as transpoviron sequences (JQ063129.1, JQ063128.1, JQ063127.1, JQ063126.1) were also downloaded from the NCBI database. The 375 ORFs with accession numbers were taken from these sequences and translated into proteins. 75 manually extracted ORFs from the putative virophage/polinton hybrids were also extracted and translated in proteins. These proteins were used to query the NCBI WGS database excluding *B. natans* using tblastn. The resulting contigs retrieved by this were then collated based on size and number of tblastn hits from different ORFs. Polinton ORFs from Repbase were extracted and tblastned against the collected contigs.

*S. minutum* contigs that had a terminal inverted repeat (TIR) at both ends of the region in which hits from the virophage and polinton ORFs clustered (the GC depleted region) were defined as full length SMPLEs and used for further analysis. ORFs from the seven contigs were reciprocally blasted against all seven contigs to identify any reading frames with internal frameshifting mutations and other stop codons (reconstructed ORFs), which yielded a total of 151 individual ORFs and reconstructed ORFs, with approximately 35 different putative genes.

These 151 ORFs were used as probes in a tblasn search against a local database combining polintons, virophages, transpovirions, and the putative metagenomics gut rumen

polinton/virophage hybrids. The best blast hits of each group were labelled as either having a best hit against polintons or against the combined virophages.

When extracting sequences from polintons and virophages for the reconstruction of phylogenetic trees, in order to ensure that all ATPase homologues were gathered, and to correct misannotated features in RepBase, the ORFs in the polintons that had hits to the putative SMPLE ATPase as well as any ORFs that had been annotated as integrase were extracted and reciprocally tblastned against the polintons from RepBase. The resultant ORFs were then aligned with each other, and any ORFs that were unalignable at the amino acid (AA) level (such as those that were misannotated) were discarded. An identical process was carried out using integrase. Identical processes were carried out with both integrase and ATPase for the previously described *B. natans* VLEs and combined virophages. An identical process was carried out with ORF-618 in PLEs and virophages. Both ATPase and integrase homologues from both groups were also used to query the NCBI eukaryotic WGS database. Retrieved contigs that had both integrase and ATPase in close proximity were considered to be putative PLEs. Close proximity was defined as less than 40 kbp, the maximum documented size of a polinton, virophage, or PLE. Within this more lenient definition, all examined putative PLEs had ATPase and integrase within 15 kbp of one another. ORF-618 homologues were tblastned against the NCBI eukaryotic WGS database and hits were added to the alignment of ORF-618.

**Phylogenetics.** Alignments were initiated using reciprocal tblastn results and then completed manually. GBlocks implemented in Seaview on its most lenient settings was used on a codon based nucleic acid alignment and then expanded and trimmed manually (Alignments available in supplementary data 2). The ATPase and integrase alignments were concatenated and used to

reconstruct Bayseian inference trees, while the ORF-618 alignment was used un-concatenated. An alignment of ATPase in the SMPLEs and *C. cohnii* only was used to reconstruct the host tree for phylogenetic reconciliation analysis. An additional alignment of ORF-618 was created for only SMPLEs and *C. cohni* for use as the parasite tree in phylogenetic reconciliation analysis. All trees were made in the parallel version of MrBayes version 3.2 (Altekar et al., 2004; Ronquist and Huelsenbeck, 2003), using the SRD06 model (Shapiro et al., 2005). SRD06 is a codon position based model, a model that partitions a nucleotide-based model into categories based on codon position. It defines a substitution model for codon positions one and two and an independent model for codon position three, with a different transition-transversion, α parameter, and relative rate between the partitions (Shapiro et al., 2005). Phylogenetic reconciliation analysis was conducted using Jane version 4 (Conow et al., 2010) with the cost settings: cospeciation 0, duplication 1, duplication and host switch 2, loss 4, failure to diverge 3. The exact ratios of the cost settings did not have an effect on the results so long as the order of values was maintained. Setting 'duplication' and 'duplication and host switch' to equal values introduced a single duplication event without a host switch but was otherwise similar to the results shown.

**Images.** Sequences were visualised and annotated in Geneious version 9.1.7 (Kearse et al., 2012) and exported as scalable vector graphic files (SVGs) for use in Inkscape version .91 (Inkscape Project, 2017). Trees output from MrBayes were visualised using Figtree version 1.4.2 (Rambaut, 2012) and exported as SVGs for use in Inkscape. Tree output from Jane was exported as rasterised Portable Network Graphics files and then vectorised and colourised in Inkscape.

## Future Research

Multiple avenues of research present themselves as follow ups to this project. Although comparisons of polinton sequence identity between species has been examined since their discovery, to my knowledge no widespread attempt to date polintons or PLEs based on divergence in their TIRs has been carried out. Given that many of the host species have documented evolutionary rates and divergence times, fortunately including the genus *Symbiodinium* (Pochon et al., 2006), this would be a fairly straightforward project. This would entail taking the polintons and PLEs described in the literature, as well as those identified over the course of the research described above and identifying their TIRs where present. Unfortunately almost none of the polintons in RepBase have annotated TIRs, many are entirely unannotated of any features, and some polintons are annotated incorrectly. This will therefore require going through over a hundred polintons and extracting the TIRs directly (skipping polintons whose host species do not have documented evolutionary rates). Some polintons listed in RepBase are also created from consensus sequences, so the originals will need to be chased down so as to not lose signal. Almost all PLEs described in the literature are extracted and described directly from unannotated contigs, as are the PLEs discovered over the course of this project, and these will require similar treatment (although some were annotated over the course of this work). Fortunately, none of these are consensus sequences. Once the TIRs are extracted it is simply a matter of comparing evolutionary rates of host species to the TIR's divergence.

Beyond the SMPLEs, applying similar search and retrieval strategies to higher plants has yielded some promising leads (Figure 1.7). However, many of these elements are entire contigs, rather than embedded in a larger contig as was the case with the SMPLEs. They may have been discarded as repetitive elements, in which case some assembly work may be able to tentatively place them, although there is always the danger that they are simply results of contamination. Admittedly, some of the PLEs described in the literature also suffer from this issue, but a more thorough analysis is warranted given that polintons and PLEs have been conspicuously absent from land plants up until now. Reciprocal blasts of putative plant PLE proteins against nr/nt, WGS, and TSA databases should show any immediate signs of contamination (although I would also presume some form of similar filtering goes on when these genomes are being sequenced). This will not conclusively disprove contamination, but should indicate whether the project is worth following up on. Downloading the raw reads and reassembly may be able to expand the contigs past their original borders and hopefully place the putative polinton/PLE within the broader context of host genes. This would be a strong indicator that the elements are in fact genuine.

Numerous potential PLEs, particularly water mould PLEs, were identified over the course of this study. Characterisation and analysis of these elements could help flesh out the polinton/PLE/virophage 'tree of life'. Green algae PLEs are only known from a few examples. Neither of these have the ORF-618 GIY-YIG HEG, but is this simply a case of a small sample size? Not all water mould PLEs possess it either, but it is

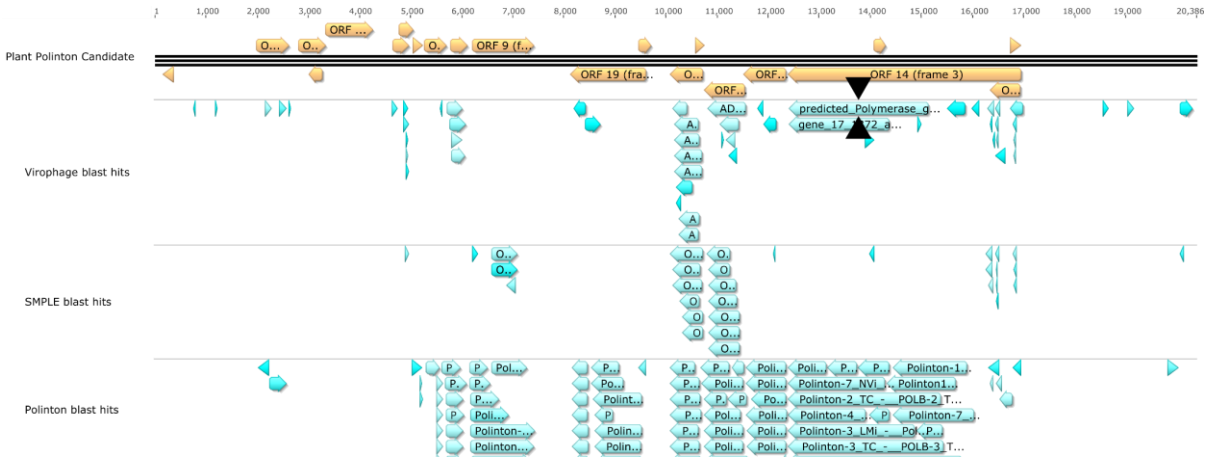


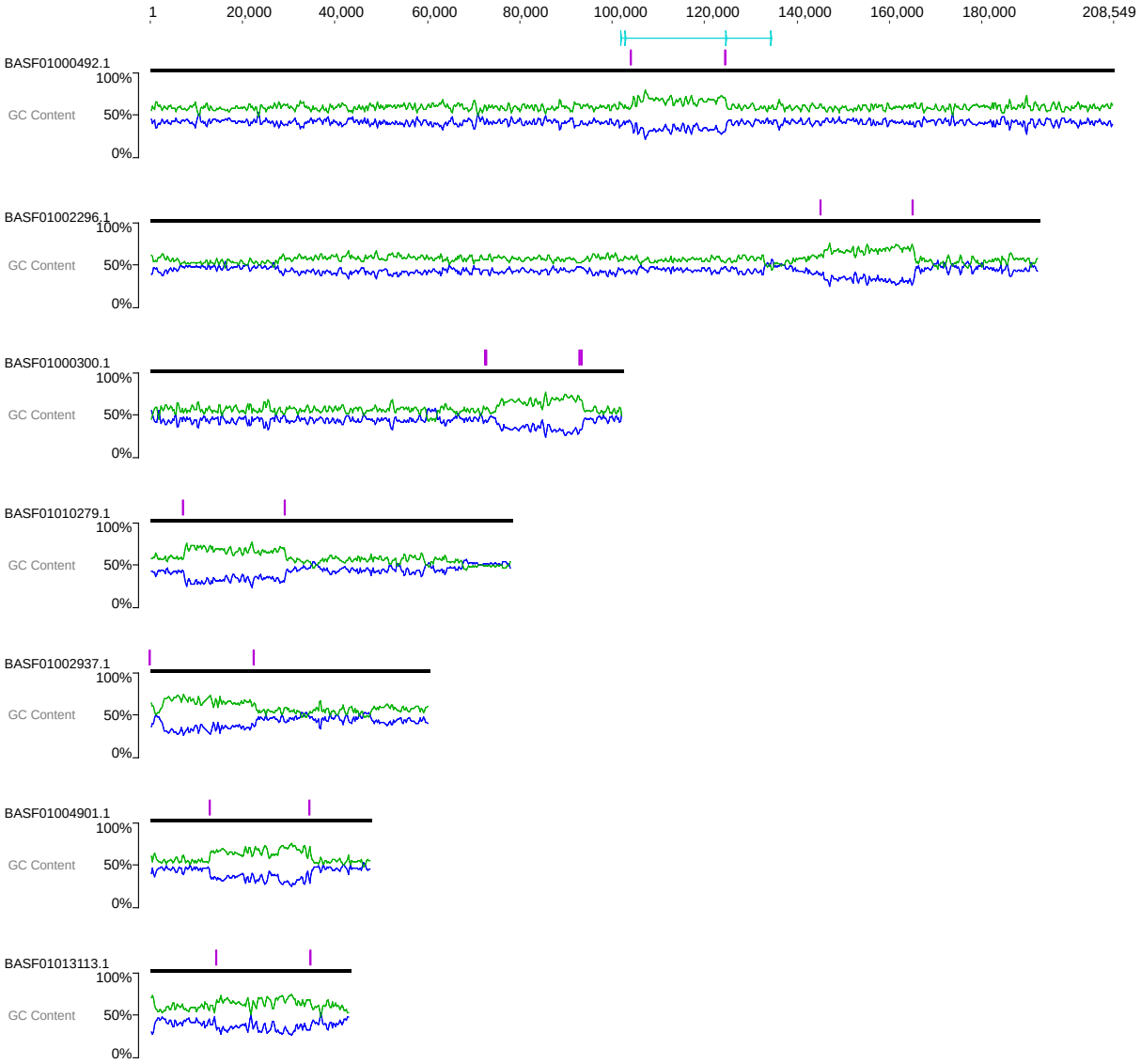

*Figure 1.7* A polinton or PLE candidate in a higher plant. Cyan arrows are blast hits, sorted into where the query ORF was drawn from. Arrows are shaded according to E value, with lighter being a lower E value. The recognizable polymerase indicates that this is probably not closely related to SMPLEs. For reference, the blast hit designated by the black arrows has an E value of 1.11E-48. Orange arrows are automatically detected ORFs over 200 bp in length.

definitely present in some. Identification and characterisation of more green algae PLEs should shed light on this. Similarly cryptophyte PLEs are only known from three poorly characterised contigs from a single species and remain enigmatic. Polinton/PLE proteins have been detected in Rhizaria via tblastn, but no Rhizaria polintons or PLEs have been extracted and characterised.

Most polinton genes are poorly characterised. A large scale mapping of polinton genes to see if common or mutually exclusive gene sets exist could reveal diversity within the canonical polintons. This project would involve the extraction of the hundreds of polinton ORFs from the polinton sequences already gathered, plus any new polinton sequences deposited in RepBase by the time this project commences. In order to correct for misannotations in RepBase, as well as ORFs truncated by frameshifting mutations and other internal stop codons, the ORFs would then go through multiple rounds of

reciprocal tblastn, ORF reconstruction, and extraction. An identical process would also be carried out for the various PLEs discovered and catalogued up until this point, so genes and gene sets within the PLEs can be compared to each other and to polintons. This last step should also shed light on the relationships between the various PLE groups and the canonical polintons. If polintons and PLEs are transmitted primarily vertically (or at least have preferential host switching within major host clades), then they are apparently quite stable as they remain recognisable between disparate groups of eukaryotes, despite a presumed origin over a billion years ago. The cause of this stability is unknown. Successfully finding and characterising stably vertically transmitted polintons and PLEs in land plants and other parts of the eukaryotic tree would also help in resolving some of the major controversies surrounding eukaryotic phylogeny.

# Supplementary Information




*Supplementary Figure S1.1* Full view of contigs from which the seven SMPLEs in Figure 1.1 are extracted. Blue line is GC content. Green line is AT content. Purple boxes are terminal inverted repeats (TIRs). Cyan arrows are Symbiodinium mRNA from transcriptome shotgun assembly data mapped onto BASF01000492, with cyan lines connecting mRNA segments. The other SMPLEs are not located in areas that map to TSA data. Sliding window size for determining GC content is 500 bp.

# Works Cited


Altekar, G., Dwarkadas, S., Huelsenbeck, J.P., Ronquist, F., 2004. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 20, 407–415. https://doi.org/10.1093/bioinformatics/btg427

Bao, W., Kojima, K.K., Kohany, O., 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6. https://doi.org/10.1186/s13100-015-0041-9

Bapteste, E., Susko, E., Leigh, J., MacLeod, D., Charlebois, R.L., Doolittle, W.F., 2005. Do orthologous gene phylogenies really support tree-thinking? BMC Evol Biol 5, 33. https://doi.org/10.1186/1471-2148-5-33

Belfort, M., Reaban, M.E., Coetzee, T., Dalgaard, J.Z., 1995. Prokaryotic introns and inteins: a panoply of form and function. Journal of Bacteriology 177, 3897–3903. https://doi.org/10.1128/jb.177.14.3897-3903.1995

Blanc, G., Gallot-Lavallée, L., Maumus, F., 2015. Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses. Proceedings of the National Academy of Sciences 112, E5318–E5326. https://doi.org/10.1073/pnas.1506469112

Bourque, G., 2009. Transposable elements in gene regulation and in the evolution of vertebrate genomes. Current Opinion in Genetics & Development 19, 607–612. https://doi.org/10.1016/j.gde.2009.10.013

Castresana, J., 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Molecular Biology and Evolution 17, 540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334

Chakrabarti, R., Hofman, F.M., Pandey, R., Mathes, L.E., Roy-Burman, P., 1994. Recombination between feline exogenous and endogenous retroviral sequences generates tropism for cerebral endothelial cells. Am. J. Pathol. 144, 348–358.

Conow, C., Fielder, D., Ovadia, Y., Libeskind-Hadas, R., 2010. Jane: a new tool for the cophylogeny reconstruction problem. Algorithms for Molecular Biology 5, 16. https://doi.org/10.1186/1748-7188-5-16

Correa, A.M.S., Ainsworth, T.D., Rosales, S.M., Thurber, A.R., Butler, C.R., Vega Thurber, R.L., 2016. Viral Outbreak in Corals Associated with an In Situ Bleaching Event: Atypical Herpes-Like Viruses and a New Megavirus Infecting Symbiodinium. Frontiers in Microbiology 7. https://doi.org/10.3389/fmicb.2016.00127

Craig, N.L., 1995. Unity in Transposition Reactions. Science 270, 253–253. https://doi.org/10.1126/science.270.5234.253

De Paepe, M., Hutinet, G., Son, O., Amarir-Bouhram, J., Schbath, S., Petit, M.-A., 2014. Temperate Phages Acquire DNA from Defective Prophages by Relaxed Homologous Recombination: The Role of Rad52-Like Recombinases. PLoS Genetics 10, e1004181. https://doi.org/10.1371/journal.pgen.1004181

Desnues, C., La Scola, B., Yutin, N., Fournous, G., Robert, C., Azza, S., Jardot, P., Monteil, S., Campocasso, A., Koonin, E.V., Raoult, D., 2012. Provirophages and transpovirons as the diverse mobilome of giant viruses. Proceedings of the National Academy of Sciences 109, 18078–18083. https://doi.org/10.1073/pnas.1208835109

Dunin-Horkawicz, S., Feder, M., Bujnicki, J.M., 2006. Phylogenomic analysis of the GIY-YIG nuclease superfamily. BMC Genomics 7, 98. https://doi.org/10.1186/1471-2164-7-98

Eddy, S.R., Gold, L., 1992. Artificial mobile DNA element constructed from the EcoRI endonuclease gene. Proceedings of the National Academy of Sciences 89, 1544–1547. https://doi.org/10.1073/pnas.89.5.1544
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32, 1792–1797. https://doi.org/10.1093/nar/gkh340
Feschotte, C., Pritham, E.J., 2007. DNA Transposons and the Evolution of Eukaryotic Genomes. Annual Review of Genetics 41, 331–368. https://doi.org/10.1146/annurev.genet.40.110405.090448
Fischer, M.G., Hackl, T., 2016. Host genome integration and giant virus-induced reactivation of the virophage mavirus. Nature 540, 288–291. https://doi.org/10.1038/nature20593
Fischer, M.G., Suttle, C.A., 2011. A Virophage at the Origin of Large DNA Transposons. Science 332, 231–234. https://doi.org/10.1126/science.1199412
Forterre, P., 2006. The origin of viruses and their possible roles in major evolutionary transitions. Virus Research 117, 5–16. https://doi.org/10.1016/j.virusres.2006.01.010
Frost, L.S., Leplae, R., Summers, A.O., Toussaint, A., 2005. Mobile genetic elements: the agents of open source evolution. Nature Reviews Microbiology 3, 722–732. https://doi.org/10.1038/nrmicro1235
Gilbert, C., Chateigner, A., Ernenwein, L., Barbe, V., Bézier, A., Herniou, E.A., Cordaux, R., 2014. Population genomics supports baculoviruses as vectors of horizontal transfer of insect transposons. Nature Communications 5. https://doi.org/10.1038/ncomms4348
Gimble, F., 2000. Invasion of a multitude of genetic niches by mobile endonuclease genes. FEMS Microbiology Letters 185, 99–107. https://doi.org/10.1016/s0378-1097(00)00069-0
Golovkina, T.V., Jaffe, A.B., Ross, S.R., 1994. Coexpression of exogenous and endogenous mouse mammary tumor virus RNA in vivo results in viral recombination and broadens the virus host range. J. Virol. 68, 5019–5026.
Gouy, M., Guindon, S., Gascuel, O., 2009. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Molecular Biology and Evolution 27, 221–224. https://doi.org/10.1093/molbev/msp259
Haapa-Paananen, S., Wahlberg, N., Savilahti, H., 2014. Phylogenetic analysis of Maverick/Polinton giant transposons across organisms. Molecular Phylogenetics and Evolution 78, 271–274. https://doi.org/10.1016/j.ympev.2014.05.024
Ho, S.Y.W., Jermiin, L.S., 2004. Tracing the Decay of the Historical Signal in Biological Sequence Data. Systematic Biology 53, 623–637. https://doi.org/10.1080/10635150490503035
Huang, J., 2013. Horizontal gene transfer in eukaryotes: the weak-link model. Bioessays 35, 868–875. https://doi.org/10.1002/bies.201300007
Inkscape Project, 2017. Inkscape.
Jeffroy, O., Brinkmann, H., Delsuc, F., Philippe, H., 2006. Phylogenomics: the beginning of incongruence? Trends Genet 22, 225–231. https://doi.org/10.1016/j.tig.2006.02.003
Kapitonov, V.V., Jurka, J., 2006. Self-synthesizing DNA transposons in eukaryotes. Proceedings of the National Academy of Sciences 103, 4540–4545. https://doi.org/10.1073/pnas.0600833103

Kapitonov, V.V., Jurka, J., 2001. Rolling-circle transposons in eukaryotes. Proceedings of the National Academy of Sciences 98, 8714–8719. https://doi.org/10.1073/pnas.151269298
Katzourakis, A., Aswad, A., 2014. The origins of giant viruses, virophages and their relatives in host genomes. BMC Biology 12. https://doi.org/10.1186/s12915-014-0051-y
Katzourakis, A., Gifford, R.J., 2010. Endogenous Viral Elements in Animal Genomes. PLoS Genetics 6, e1001191. https://doi.org/10.1371/journal.pgen.1001191
Kazazian, H.H., 2004. Mobile Elements: Drivers of Genome Evolution. Science 303, 1626–1632. https://doi.org/10.1126/science.1089670
Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649. https://doi.org/10.1093/bioinformatics/bts199
Knoll, A.H., Javaux, E.J., Hewitt, D., Cohen, P., 2006. Eukaryotic organisms in Proterozoic oceans. Philosophical Transactions of the Royal Society B: Biological Sciences 361, 1023–1038. https://doi.org/10.1098/rstb.2006.1843
Koonin, E.V., 2016. Viruses and mobile elements as drivers of evolutionary transitions. Philosophical Transactions of the Royal Society B: Biological Sciences 371, 20150442. https://doi.org/10.1098/rstb.2015.0442
Koonin, E.V., Krupovic, M., Yutin, N., 2015. Evolution of double-stranded DNA viruses of eukaryotes: from bacteriophages to transposons to giant viruses. Annals of the New York Academy of Sciences 1341, 10–24. https://doi.org/10.1111/nyas.12728
Koonin, E.V., Senkevich, T.G., Dolja, V.V., 2006. The ancient Virus World and evolution of cells. Biol. Direct 1, 29. https://doi.org/10.1186/1745-6150-1-29
Krupovic, M., Koonin, E.V., 2014. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. Nature Reviews Microbiology 13, 105–115. https://doi.org/10.1038/nrmicro3389
Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947–2948. https://doi.org/10.1093/bioinformatics/btm404
Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., Lanczycki, C.J., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Wang, Z., Yamashita, R.A., Zhang, D., Zheng, C., Bryant, S.H., 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res. 43, D222-226. https://doi.org/10.1093/nar/gku1221
Mossel, E., Steel, M., 2005. How much can evolved characters tell us about the tree that generated them?, in: Mathematics of Evolution and Phylogeny. Oxford University Press, Oxford.
Pandey, R., Ghosh, A.K., Kumar, D.V., Bachman, B.A., Shibata, D., Roy-Burman, P., 1991. Recombination between feline leukemia virus subgroup B or C and endogenous env elements alters the in vitro biological activities of the viruses. J. Virol. 65, 6495–6508.
Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T.J., Manuel, M., Wörheide, G., Baurain, D., 2011. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLoS Biology 9, e1000602. https://doi.org/10.1371/journal.pbio.1000602

Philippe, H., Delsuc, F., Brinkmann, H., Lartillot, N., 2005. Phylogenomics. Annual Review of Ecology, Evolution, and Systematics 36, 541–562. https://doi.org/10.1146/annurev.ecolsys.35.112202.130205
Philippe, H., Douady, C.J., 2003. Horizontal gene transfer and phylogenetics. Curr Opin Microbiol 6, 498–505. https://doi.org/10.1016/j.mib.2003.09.008
Pochon, X., Montoya-Burgos, J.I., Stadelmann, B., Pawlowski, J., 2006. Molecular phylogeny, evolutionary rates, and divergence timing of the symbiotic dinoflagellate genus Symbiodinium. Molecular Phylogenetics and Evolution 38, 20–30. https://doi.org/10.1016/j.ympev.2005.04.028
Rambaut, A., 2012. FigTree v1. 4.
Rasmussen, H.B., 1997. Interactions between exogenous and endogenous retroviruses. Journal of Biomedical Science 4, 1–8. https://doi.org/10.1007/bf02255587
Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574. https://doi.org/10.1093/bioinformatics/btg180
Santini, S., Jeudy, S., Bartoli, J., Poirot, O., Lescot, M., Abergel, C., Barbe, V., Wommack, K.E., Noordeloos, A.A.M., Brussaard, C.P.D., Claverie, J.-M., 2013. Genome of Phaeocystis globosa virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. Proceedings of the National Academy of Sciences 110, 10800–10805. https://doi.org/10.1073/pnas.1303251110
Shapiro, B., Rambaut, A., Drummond, A.J., 2005. Choosing Appropriate Substitution Models for the Phylogenetic Analysis of Protein-Coding Sequences. Molecular Biology and Evolution 23, 7–9. https://doi.org/10.1093/molbev/msj021
Sheets, R.L., Pandey, R., Klement, V., Grant, C.K., Roy-burman, P., 1992. Biologically selected recombinants between Feline Leukemia Virus (FeLV) subgroup a and an endogenous FeLV element. Virology 190, 849–855. https://doi.org/10.1016/0042-6822(92)90924-e
Shoguchi, E., Shinzato, C., Kawashima, T., Gyoja, F., Mungpakdee, S., Koyanagi, R., Takeuchi, T., Hisata, K., Tanaka, M., Fujiwara, M., Hamada, M., Seidi, A., Fujie, M., Usami, T., Goto, H., Yamasaki, S., Arakaki, N., Suzuki, Y., Sugano, S., Toyoda, A., Kuroki, Y., Fujiyama, A., Medina, M., Coffroth, M.A., Bhattacharya, D., Satoh, N., 2013. Draft Assembly of the Symbiodinium minutum Nuclear Genome Reveals Dinoflagellate Gene Structure. Current Biology 23, 1399–1408. https://doi.org/10.1016/j.cub.2013.05.062
Stimson, J., Sakai, K., Sembali, H., 2002. Interspecific comparison of the symbiotic relationship in corals with high and low rates of bleaching-induced mortality. Coral Reefs 21, 409–421. https://doi.org/10.1007/s00338-002-0264-3
Wakefield, T.S., Kempf, S.C., 2001. Development of Host- and Symbiont-Specific Monoclonal Antibodies and Confirmation of the Origin of the Symbiosome Membrane in a Cnidarian–Dinoflagellate Symbiosis. The Biological Bulletin 200, 127–143. https://doi.org/10.2307/1543306
Wolf, Y.I., Rogozin, I.B., Grishin, N.V., Koonin, E.V., 2002. Genome trees and the tree of life. Trends in Genetics 18, 472–479. https://doi.org/10.1016/S0168-9525(02)02744-0
Yutin, N., Colson, P., Raoult, D., Koonin, E.V., 2013a. Mimiviridae: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family. Virology Journal 10, 106. https://doi.org/10.1186/1743-422x-10-106

Yutin, N., Kapitonov, V.V., Koonin, E.V., 2015a. A new family of hybrid virophages from an animal gut metagenome. Biology Direct 10. https://doi.org/10.1186/s13062-015-0054-9

Yutin, N., Raoult, D., Koonin, E.V., 2013b. Virophages, polintons, and transpovirons: a complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. Virol. J. 10, 158. https://doi.org/10.1186/1743-422X-10-158

Yutin, N., Shevchenko, S., Kapitonov, V., Krupovic, M., Koonin, E.V., 2015b. A novel group of diverse Polinton-like viruses discovered by metagenome analysis. BMC Biology 13. https://doi.org/10.1186/s12915-015-0207-4

Yutin, N., Wolf, Y.I., Koonin, E.V., 2014. Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. Virology 466–467, 38–52. https://doi.org/10.1016/j.virol.2014.06.032

# Appendix 1A

Published review article discussing some of the ideas presented in this work.
Note: As this is an unaltered published manuscript the figure in it is numbered according to the published work and not by thesis chapter. Only page numbers have been added for ease of navigation and reference. Bold monospace font at the bottom centre of each page of the article follow page numbering of the thesis. Variable space font at the top left of all pages past the first is the original unaltered page numbering (starting at 59).

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Virology

# Disentangling the origins of virophages and polintons

Samuel Campbell[1], Amr Aswad[1] and Aris Katzourakis

Virophages and polintons are part of a complex system that also involves eukaryotes, giant viruses, as well as other viruses and transposable elements. Virophages are cosmopolitan, being found in environments ranging from the Amazon River to Antarctic hypersaline lakes, while polintons are found in many single celled and multicellular eukaryotes. Virophages and polintons have a shared ancestry, but their exact origins are unknown and obscured by antiquity and extensive horizontal gene transfer (HGT). Paleovirology can help disentangle the complicated gene flow between these two, as well as their giant viral and eukaryotic hosts. We outline the evidence and theoretical support for polintons being descended from viruses and not vice versa. In order to disentangle the natural history of polintons and virophages, we suggest that there is much to be gained by embracing rigorous metagenomics and evolutionary analyses. Methods from paleovirology will play a pivotal role in unravelling ancient relationships, HGT and patterns of cross-species transmission.

**Address**
Department of Zoology, University of Oxford, Oxford OX1 3PS, UK

Corresponding author: Katzourakis, Aris (aris.katzourakis@zoo.ox.ac.uk)
[1] These authors contributed equally to this work.

**Current Opinion in Virology** 2017, **25**:59–65

This review comes from a themed issue on **Paleovirology**

Edited by **Aris Katzourakis**

For a complete overview see the Issue and the Editorial

Available online 9th August 2017

http://dx.doi.org/10.1016/j.coviro.2017.07.011



## Introduction

The inhabitants of the microscopic world of giant viruses, their eukaryotic hosts, virophages, and various transposable elements (TEs) interact in a tangled web of interrelatedness, parasitism and coevolution. Giant viruses, can reach up to 1500 nm wide and harbour genomes over a megabase long. They are part of the nucleocytoplasmic large DNA viruses group (NCLDVs), that share a complement of six 'core genes' [1–3], and includes families such as the *Poxviridae* and *Phycodnaviridae*. Several of the largest NCLDVs contain genes once considered exclusively cellular which has been taken as evidence in favour of a reduced-cell explanation for their gigantism [4]. However, genome expansion from smaller viruses via horizontal gene transfer (HGT) and genomic duplication is more consistent with the phylogenetic and genomic evidence, such as the presence of recently acquired host translational machinery in the Klosneuviruses [5••]. Furthermore Poxviruses are known to adaptively expand their genomes rapidly by up to 10%, albeit only transiently [6].

Giant virus genomes have been colonised by TEs [7,8], and phylogenies reconstructed from different genes reveal conflicting topologies indicating an extensive history of HGT. This complexity is reflective of their ecology, where giant viruses interact with both their hosts and associated satellite viruses known as virophages. All viruses are obligate parasites, but satellite viruses take parasitism a step further in that they require coinfection by a helper virus to replicate. The first virophage to be discovered, dubbed 'Sputnik', is associated with the giant virus *Acanthamoeba castellanii mamavirus* (ACMV) [9]. Like other giant viruses (and many NCLDVs), ACMV creates dense cytoplasmic structures known as viral factories. The term virophage is derived from bacteriophage, as Sputnik hijacks the replication machinery of these viral factories, rather than the native replication machinery of the host cell, and has a negative impact on replication of the helper virus [9].

Virophages are dsDNA viruses, but they do not appear to be closely related to other dsDNA viruses. Their morphogenesis genes, such as their pox-like A32 packaging proteins, place them in the diverse PRD1-adenovirus lineage, but some possess unrelated genes such as retroviral-like integrases, and they also contain genes of unknown function, some of which have no known homologues [9–11]. Interestingly however, there are remarkable similarities between virophages and a family of DNA TEs initially named mavericks but now more widely referred to as polintons [12]. Polintons are an atypical group of large TEs that are able to self-replicate using their own polymerase and integrase genes, hence the name 'Pol-Int-on' [13–15]. Although there are larger and smaller examples, most polintons are ~15–22 kbp in length [16,17•] which is comparable to that of virophage genomes. While virophages and polintons are extremely diverse, both at the nucleotide level and in terms of their genomic content, they nonetheless share multiple genes indicative of shared ancestry [12,17•,18].

The interactions between all the players in this tangled microscopic world are far from resolved, and the nature of the relationship between virophages and polintons is elusive. The clear indication of shared ancestry is set against a backdrop of frequent HGT that complicates

attempts to identify distinct lineages, and the antiquity of their evolutionary interactions obscures the processes that shaped their combined natural histories [16,18,19]. However, because polintons are anciently integrated into host genomes, their relative evolutionary stability (in terms of mutation rate and frequency of HGT events) allows us to use techniques developed in paleovirology for the study of ancient viruses. As the field progresses and more data become available, we will be able to integrate observations from phylogenetic reconstruction and genomics. This will be informative for both virophages and polintons separately, but also in combined analyses that will reveal aspects of their shared evolutionary history. The synthesis of such analyses can be contextualised with additional evidence from NCLDV evolutionary reconstruction and that of their eukaryotic hosts.

This paleovirological approach has previously been effective in resolving similarly complex evolutionary histories, such as that of a retrovirus with mammalian origins that has also integrated into farmed poultry, a herpesvirus, and a poxvirus [20]. Similarly, a paleovirological approach was deployed to identify multiple gene transfer events between an unusual retroviral superantigen gene and three different herpesviruses, via two separate mammalian hosts [21]. In this paper, we review the prevailing evidence and opinions about the origins of virophages and polintons with a paleovirological perspective. We consider the extent to which different aspects of their relatedness can be explained by their evolutionary interactions with giant viruses and eukaryotic hosts.

## Virophages

Since the discovery of Sputnik, other virophage families have been discovered, primarily in metagenomic data [12,22–26,27•,28•,29•,30•]. Most virophages appear to have circular double stranded DNA genomes ranging from ~17 to 29 kbp, but the presence, order, and orientation of most genes are not well conserved between proposed families. Only a fraction of virophages have been experimentally validated and associated with a giant virus. Sputnik has been observed *in vivo*, and its presence in viral factories results in dysfunctional and abortive ACMV virions. This reduces the yield of ACMV by up to 70% and decreases host cell lysis threefold [9]. Although distantly related to Sputnik, mavirus similarly inhibits its host giant virus *Cafeteria roenbergensis* virus (CroV) and increases *C. roenbergensis* survival [12]. The exact mechanisms behind how Sputnik and mavirus hijack the viral factories are unknown, and since most virophages are only known from metagenomic data, it is not certain how many of them rescue the eukaryotic host. In contrast, the Zamilon virophage, a close relative of Sputnik, does not inhibit APMV reproduction or cell lysis [26]. It is conceivable that like the subviral satellite hepatitis D [31], a virophage could lead to increased morbidity and mortality in the eukaryotic host. It is also unclear how other players in the system, such as TEs, affect the outcome of the interaction between virophages, giant viruses, and eukaryotes.

## Polintons

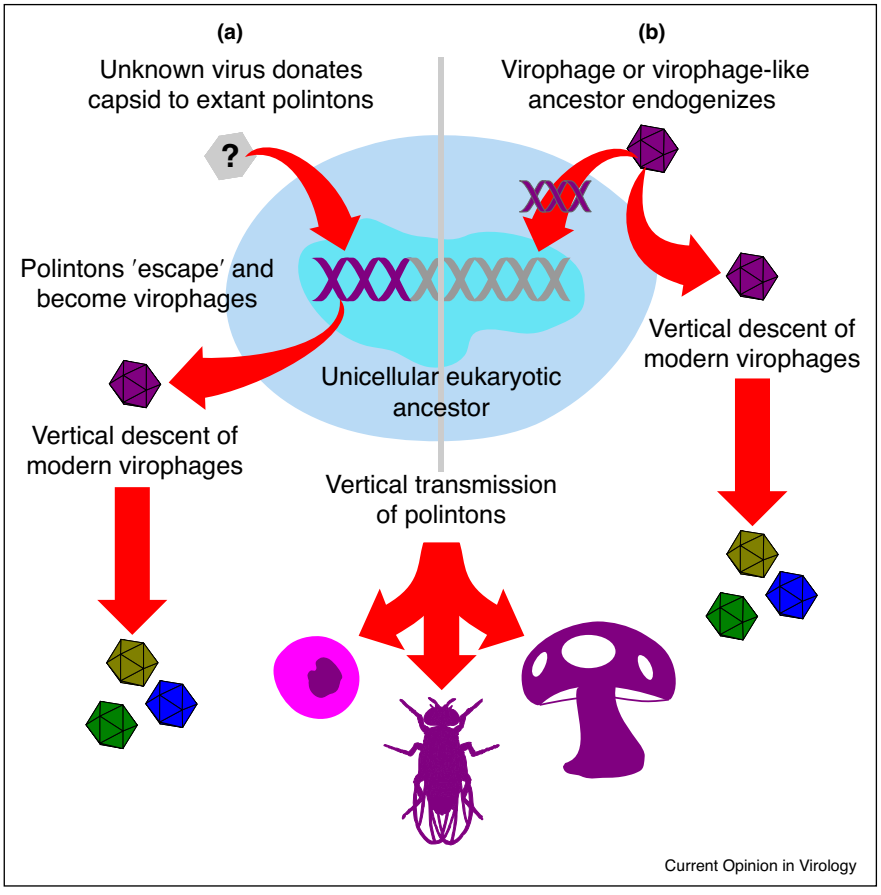
While giant viruses and virophages only infect single celled eukaryotes, polintons and polinton-like elements (PLEs) have been described in diverse single celled and multicellular eukaryotes, from cnidarians to water moulds. Given how distantly related some of the hosts are, and the notable absence of polintons in mammals and terrestrial plants [16,17•,32], their taxonomic distribution suggests frequent novel integrations, with extensive HGT between different kingdoms/super groups, and the repeated loss in descendant lineages of ancient integrations. Indeed, phylogenetic analysis indicates a mix of both vertical and horizontal transmission [16]. This pattern is not unlike that of endogenous retrovirus phylogenies, which contain both horizontally and vertically transmitted insertions [33]. As paleovirology focuses on such questions of gene flow, it could be used to quantify the relative extent of these two processes via a statistical comparison of polinton and host phylogenies.

## Virophages ⇄ Polintons

While convergent evolution and HGT could explain some of their similarities, extensive phylogenetic and genomic evidence linking virophages and polintons makes a compelling case for the hypothesis of shared ancestry, for which there is broad consensus [12,18,19,32,34]. However, this leads to questions on the direction of gene flow and whether the ancestor of both entities was originally a TE that resembles contemporary polintons, or if a viral ancestor gave rise to virophages and their genomic relatives. Conclusive evidence for either scenario has yet to be unearthed, and arguments for both have been proposed (Figure 1). In the TE-first scenario, virophages are the descendants of 'escaped' polintons (Figure 1a) [10]. An argument against this is that because virophages are dependent on co-infection by giant viruses, while polintons are not, the loss of dependency following integration is more readily explained than its arising *de novo* in a DNA transposon [12,18]. Furthermore, there are seven genes in mavirus with similarity to polinton genes, compared to only three potential homologs with other virophages like Sputnik. This suggests that mavirus may be more closely related to polintons than it is to other virophages, indicating that gene flow went from virus to TE, and that a virus is ancestral to both groups (Figure 1b) [12,18].

Furthermore, there is experimental evidence of gene flow from a virophage to its host, which supports the virophage-first hypothesis. The spontaneous integration of mavirus into *C. roenbergensis* has been observed *in vitro*, where they remain quiescent until coinfection by CroV results in the production and eventual release of mavirus

**Figure 1**




Two proposed scenarios for the origins of polintons. **(a)** An unknown virus (grey icosahedron) infected a cell and recombined with pre-existing polintons. The polintons then escaped as virophages (purple icosahedron). **(b)** A virophage (purple icosahedron) integrated into the genome of a cell and was then transmitted vertically, as a standard EVE, diverging along with its hosts. In both cases polintons subsequently descend primarily vertically as do virophages.

[35••]. Unlike with normal mavirus infection, this does not seem to protect the individual cell initially infected with CroV, but it does protect the wider population from lysis, suggesting that endogenous virophages can act as antiviral defence systems in clonal protist populations [35••].

This co-opting of endogenous viral elements (EVEs) for defence is known as EVE-derived immunity (EDI), and has been described in over a dozen cases in multicellular organisms [36]. As well as experimental observations, transcriptionally active genes of putative virophage origin, dubbed virophage like elements (VLEs), are incorporated in the genome of the algae *Bigelowiella natans* [37••]. No known giant virus has been observed parasitising *B. natans*, so any defensive function is speculative. Nevertheless, VLEs could be indicative of a historical confrontation with an ancient giant virus [37••], and techniques from paleovirology are well-placed to investigate this. VLEs are in essence EVEs, and their presence gives us a clue as to what sort of viruses ancestral *B. natans* had to contend with. It would be interesting to use functional paleovirology [38] to reconstitute virophages from VLEs to test their host giant virus range. Although they are the only two reported cases of virophage/VLE integration in eukaryotes, the cases of *B. natans* and *C. roenbergensis* provide evidence of an evolutionary scenario in which polintons arose from integrated virophages that were intitially maintained as EDIs [35••,37••]. However, even if some polintons arose from EVEs that conferred a benefit as EDIs, others could have randomly integrated without substantial effect on the host, and drifted or hitchhiked to fixation [39].

Paleovirological discoveries of additional VLEs combined with evolutionary analysis of these integrated viral sequences can differentiate these two scenarios. With this amount of HGT, the questions must be examined at the level of individual genes. Designating 'core genes' in virophages [19,40] will be a challenge, because the

discovery of new taxa may whittle down their universality. Nonetheless, identifying widely conserved genes will facilitate answering specific questions, such as mapping the connections between polintons and virophages. For instance, integrase is going to be crucial as it is frequently found in both, even if we cannot include those virophages where it is absent.

## Networks and trees

Different papers have presented phylogenetic trees with varying levels of disagreement, depending on which genes are used and which taxa are being sampled — even the monophyly of polintons and virophages is not consistent [12,19,29•,32,41]. The matter is complicated further by the fact that some virophages have multiple open reading frames (ORFs) with no homologues in sequence databases [11]. Therefore, despite a plausible evolutionary scenario for a 'virus first' model being demonstrated (supported with data and experimental evidence [35••,37••]), the sequence of events in this chicken-and-egg scenario is unknown.

Reconstructing the evolutionary history of virophages will be central to disentangling the complex history of HGT between virophages, giant viruses, polintons, and their hosts. Their relatively small size and ability to integrate mean that virophages could facilitate HGT between giant viruses and their host eukaryotes. Genes taken from a giant virus by a virophage that is subsequently donated to a eukaryote could be viewed as 'vicarious EVEs' [18]. Indeed, genes of likely NCLDV origin have been discovered in land plants, despite there being no known exogenous NCLDVs that infect them [42,43•]. The large size of the giant virus genomes would make integration of the entire viral genome without disrupting cellular functions difficult, which may explain why no giant viral EVEs have been identified. Nevertheless, large viral integrations can and do occur, such as herpesviruses in primates [44–46], phycodnaviruses in algae [47,48], as well as endogenized phycodnavirus segments [49,50].

One way to make sense of the tangled relationships between NCLDVs, eukaryotes, virophages and bacteriophages is by drawing gene networks [10,19,34]. While these accurately capture a general pattern of HGT between these entities, they provide limited insight into the underlying phylogenetic history, and direction of transfer. Furthermore, these similarity-based approaches do not distinguish genes on the basis of gene function. For example, many virophages contain genes not found in other virophage lineages such as retroviral-like integrase in maviruses. Similarly, Sputnik appears evolutionarily linked to bacterial TEs carrying TV-Pol family genes (e.g. IS605 or SCCmec elements) [51], and has a phage-like bacterial transposase that is unrelated to integrases from mavirus, autonomous polintons, and some PLEs [9]. Thus, in addition to recombination, this indicates convergent evolution that gene network approaches cannot distinguish.

## Origins of virophages: a viral ancestor for virophages and polintons

If polintons are EVEs, descended from an anciently integrated virophage (Figure 1b), this does not answer where virophages came from. A third hypothesis addressing whether virophages or polintons came first puts hypothetical 'polintoviruses' at the origin of not just virophages and polintons, but also NCLDVs, bidnaviruses, and adenoviruses [10,52•]. Based on the presence of ORFs that are predicted to form the double jelly roll fold seen in virophage capsid proteins, it has been suggested that many elements classified as non-viral TEs (i.e. polintons) are actually *bona fide* viruses [53]. In this hypothesis, tectiviruses (bacteriophages infecting gram-negative bacteria) came into eukaryotes with the capture of mitochondria [10,52•]. Tectiviruses acquired cysteine protease and integrase from already extant transposons, and subsequently became polintoviruses, capable of existing exogenously and endogenously, while those that lost their capsid gene became intracellular polintons [52•].

The polintovirus hypothesis also suggests that these integrated tectiviruses escaped the nucleus via the acquisition of DNA dependent RNA polymerase and an mRNA capping system, forking into cytoplasmic plasmids via genome reduction and into NCLDVs via the replacement of POLB, and a 10–100 fold (or more) genome expansion [52•]. Additionally, polintoviruses diverged into other exogenous viruses, some of which began parasitising and swapping genes with some NCLDVs [52•].

Possession of a coat-protein encoding gene, as well as the ability to form virions are the defining features of a *bona fide* virus. Polintoviruses have never been observed, and thus far, there is no indication that the putative capsid ORFs of some polintons are functional. Nonetheless, it is likely that the ancestor of both polintons and virophages was a virus, and future work remains to demonstrate what it may have resembled and to what extent it is also the ancestor of other lineages.

## In depth gene-by-gene paleovirological analysis will unravel the natural history of polintons and virophages

Regardless of the direction of transfer, polintons are endogenous relatives of virophages. As such, they can be studied under a paleovirology framework in the same manner as EVEs. Over the last several decades, retroviruses have been systematically investigated using robust phylogenetics and genome-wide statistical analyses. This has revealed insights ranging from uncovering ancient patterns of cross-species transmissions across

vertebrates [33,54–56], to identifying the many ways that retroviral genes have been co-opted by their hosts for a beneficial function [36,57,58]. Moreover, the relationship between extant virophages and polintons is analogous to that between contemporary and ancient retroviruses. For instance, like polintons, the diversity and range of ancient retroviruses far exceeds that of their contemporary cousins [54]. Such paleovirological studies have enabled us to re-evaluate our understanding of groups like lentiviruses and foamy viruses, as well as identify the broad and influential effect they have on host genomes [36,58,59,60••]. By applying the same paleovirology principles, we will be able to reach detailed and statistically robust conclusions about specific groups of polintons, virophages, and/or the genes contained within them.

Central to the study of viruses in paleovirology is the adherence to certain standards of data analysis and rigorous evidence, since most research depends on third-party genomic data collected under different conditions and assembled to varying degrees of accuracy. The recognition that metagenomics and paleovirology have an overlapping set of methodologies and goals is giving rise to novel approaches in viral discovery and analysis [46,61]. Most virophages identified thus far only have bioinformatic representation, having been mined from large-scale metagenomic studies. While these sequences are invaluable, we must be cautious and apply the same level of scrutiny as we would with endogenous viruses. This includes offering supportive evidence for valid assembly as well as distinguishing between integrations and 'free-living' viruses.

## Conclusions

There is unlikely to be a unifying theory that can at once explain the lack of synteny between virophages and giant viruses, between different lineages of virophages, and between virophages and polintons. Although a relationship between polintons and virophages is clear, there remains a large gap that separates the groups in terms of phylogenetic similarity, genomic organisation, and gene content. Therefore while their ancestor was likely viral, the nature and identity of this virus remains to be determined. A more piecemeal approach that can explain parts of the puzzle with less uncertainty will inevitably develop as we identify new lineages of virophages and polintons. Moreover, increased taxonomic sampling will also reveal transitional viral fossils that could help disentangle the entwined natural histories of genes and genomes, by combining the information from phylogenetic reconstruction and analysis of genomic organisation.

The application of techniques that synthesise analyses of both EVEs and viral genomes will help reconstruct very ancient evolutionary events, and paleovirology is well equipped to deal with remarkably large timescales. For example, such hybrid approaches have been used to trace the origin of retroviruses to the seas of the early Palaeozoic Era more than 450 million years ago [60••]. However, we must also be prepared to recognise that if the history of these entities extends far beyond such timescales — possibly even preceding the last universal cellular ancestor — then the answers to some (or all) of the questions posed here may be out of reach.

## Conflicts of interest

We wish to confirm that there are no known conflicts of interest associated with this publication. There has been no significant financial support for this work that could have influenced its outcome.

## References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest


1. Iyer LM, Balaji S, Koonin EV, Aravind L: **Evolutionary genomics of nucleo-cytoplasmic large DNA viruses**. *Virus Res* 2006, **117**:156-184.
2. Yutin N, Wolf Y, Raoult D, Koonin EV: **Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution**. *Virol J* 2009, **6**.
3. Yutin N, Koonin EV: **Pandoraviruses are highly derived phycodnaviruses**. *Biol Direct* 2013, **8**.
4. Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie JM: **The 1.2-megabase genome sequence of Mimivirus**. *Science* 2004, **306**:1344-1350.
5. •• Schulz F, Yutin N, Ivanova NN, Ortega DR, Lee TK, Vierheilig J, Daims H, Horn M, Wagner M, Jensen GJ *et al.*: **Giant viruses with an expanded complement of translation system components**. *Science* 2017, **356**:82-85.

Describes a novel family of giant viruses with the most complete set of translational machinery genes yet seen in viruses.

6. Elde NC, Child SJ, Eickbush MT, Kitzman JO, Rogers KS, Shendure J, Geballe AP, Malik HS: **Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses**. *Cell* 2012, **150**:831-841.
7. Desnues C, La Scola B, Yutin N, Fournous G, Robert C, Azza S, Jardot P, Monteil S, Campocasso A, Koonin EV *et al.*: **Provirophages and transpovirons as the diverse mobilome of giant viruses**. *PNAS* 2012, **109**:18078-18083.
8. Sun C, Feschotte C, Wu Z, Mueller RL: **DNA transposons have colonized the genome of the giant virus *Pandoravirus salinus***. *BMC Biol* 2015, **13**:38.
9. La Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, Fournous G, Merchat M, Suzan-Monti M, Forterre P, Koonin EV *et al.*: **The virophage as a unique parasite of the giant mimivirus**. *Nature* 2008, **455**:100-104.
10. Krupovic M, Koonin EV: **Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution**. *Nat Rev Microbiol* 2015, **13**:105-115.
11. Sun S, La Scola B, Bowman VD, Ryan CM, Whitelegge JP, Raoult D, Rossmann MG: **Structural studies of the Sputnik virophage**. *J Virol* 2010, **84**:894-897.
12. Fischer MG, Suttle CA: **A virophage at the origin of large DNA transposons**. *Science* 2011, **332**:231-234.
13. Feschotte C, Pritham EJ: **Non-mammalian c-integrases are encoded by giant transposable elements**. *Trends Genet* 2005, **21**:551-552.

14. Pritham EJ, Putliwala T, Feschotte C: **Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses**. *Gene* 2007, **390**:3-17.

15. Kapitonov VV, Jurka J: **Self-synthesizing DNA transposons in eukaryotes**. *PNAS* 2006, **103**:4540-4545.

16. Haapa-Paananen S, Wahlberg N, Savilahti H: **Phylogenetic analysis of Maverick/Polinton giant transposons across organisms**. *Mol Phylogenet Evol* 2014, **78**:271-274.

17. • Bao W, Kojima KK, Kohany O: **Repbase Update, a database of repetitive elements in eukaryotic genomes**. *Mob DNA* 2015, **6**:11.
The primary public repository of polinton sequences.

18. Katzourakis A, Aswad A: **The origins of giant viruses, virophages and their relatives in host genomes**. *BMC Biol* 2014, **12**:51.

19. Yutin N, Raoult D, Koonin EV: **Virophages, polintons, and transpovirons: a complex evolutionary network of diverse selfish genetic elements with different reproduction strategies**. *Virol J* 2013, **10**:158.

20. Niewiadomska AM, Gifford RJ: **The extraordinary evolutionary history of the reticuloendotheliosis viruses**. *PLoS Biol* 2013, **11**: e1001642.

21. Aswad A, Katzourakis A: **Convergent capture of retroviral superantigens by mammalian herpesviruses**. *Nat Commun* 2015, **6**:8299.

22. Yau S, Lauro FM, DeMaere MZ, Brown MV, Thomas T, Raftery MJ, Andrews-Pfannkoch C, Lewis M, Hoffman JM, Gibson JA *et al.*: **Virophage control of antarctic algal host–virus dynamics**. *Proc Natl Acad Sci U S A* 2011, **108**:6163-6168.

23. Zhou J, Zhang W, Yan S, Xiao J, Zhang Y, Li B, Pan Y, Wang Y: **Diversity of virophages in metagenomic data sets**. *J Virol* 2013, **87**:4225-4236.

24. Gaia M, Pagnier I, Campocasso A, Fournous G, Raoult D, La Scola B: **Broad spectrum of mimiviridae virophage allows its isolation using a mimivirus reporter**. *PLoS ONE* 2013, **8**:e61912.

25. Santini S, Jeudy S, Bartoli J, Poirot O, Lescot M, Abergel C, Barbe V, Wommack KE, Noordeloos AA, Brussaard CP *et al.*: **Genome of *Phaeocystis globosa* virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes**. *PNAS* 2013, **110**:10800-10805.

26. Gaia M, Benamar S, Boughalmi M, Pagnier I, Croce O, Colson P, Raoult D, La Scola B: **Zamilon, a novel virophage with mimiviridae host specificity**. *PLoS ONE* 2014, **9**:e94923.

27. • Bellas CM, Anesio AM, Barker G: **Analysis of virus genomes from glacial environments reveals novel virus groups with unusual host interactions**. *Front Microbiol* 2015, **6**:656.
Describes a metagenomically identified virophage from Greenland glacier showing homology in one of its genes to *Pithovirus sibericum*, a giant virus extracted from Siberian permafrost.

28. • Zhou J, Sun D, Childers A, McDermott TR, Wang Y, Liles MR: **Three novel virophage genomes discovered from yellowstone lake metagenomes**. *J Virol* 2015, **89**:1278-1285.
This paper expands on the number of virophages found in Yellowstone Lake, further indicating that multiple species of virophage can co-exist in a single environment.

29. • Gong C, Zhang W, Zhou X, Wang H, Sun G, Xiao J, Pan Y, Yan S, Wang Y: **Novel virophages discovered in a freshwater lake in China**. *Front Microbiol* 2016, **7**:5.
Metagenomically identifies a novel group of virophages via their major capsid proteins that are most closely related most Yellowstone Lake virophages as well Organic Lake Virophage, with some novel ORFs showing sequence similarity to (and possible HGT with) phycodnavirus ORFs.

30. • Oh S, Yoo D, Liu WT: **Metagenomics reveals a novel virophage population in a Tibetan mountain lake**. *Microbes Environ* 2016, **31**:173-177.
This paper metagenomically identifies numerous virophage like sequences closely related Organic Lake Virophage and Yellowstone Lake virophages as well as conclusively identifying a novel full virophage genome.

31. Fattovich G, Giustina G, Christensen E, Pantalena M, Zagni I, Realdi G, Schalm SW: **Influence of hepatitis delta virus infection on morbidity and mortality in compensated cirrhosis type B. The European Concerted Action on Viral Hepatitis (Eurohep)**. *Gut* 2000, **46**:420-426.

32. Yutin N, Shevchenko S, Kapitonov V, Krupovic M, Koonin EV: **A novel group of diverse Polinton-like viruses discovered by metagenome analysis**. *BMC Biol* 2015, **13**:95.

33. Hayward A, Grabherr M, Jern P: **Broad-scale phylogenomics provides insights into retrovirus-host evolution**. *PNAS* 2013, **110**:20146-20151.

34. Iranzo J, Krupovic M, Koonin EV: **The double-stranded DNA virosphere as a modular hierarchical network of gene sharing**. *MBio* 2016, **7**:e00978-e1016.

35. •• Fischer MG, Hackl T: **Host genome integration and giant virus-induced reactivation of the virophage mavirus**. *Nature* 2016, **540**:288-291.
Describes the population protective *de novo* integration of mavirus into *C. roenbergensis*, a notes the similarity between the newly integrated pro-virophages and polintons.

36. Aswad A, Katzourakis A: **Paleovirology and virally derived immunity**. *Trends Ecol Evol* 2012, **27**:627-636.

37. •• Blanc G, Gallot-Lavallée L, Maumus F: **Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses**. *PNAS* 2015, **112**:E5318-E5326.
First report of transcriptionally active virophage like elements being found in a eukaryotic genome, presenting evidence for the integration of virophages into cellular genomes.

38. Katzourakis A: **Paleovirology: inferring viral evolution from host genome sequence data**. *Philos Trans R Soc B* 2013, **368** 20120493.

39. Katzourakis A, Gifford RJ: **Endogenous viral elements in animal genomes**. *PLoS Genet* 2010, **6**:e1001191.

40. Krupovic M, Kuhn JH, Fischer MG: **A classification system for virophages and satellite viruses**. *Arch Virol* 2016, **161**:233-247.

41. Yutin N, Kapitonov VV, Koonin EV: **A new family of hybrid virophages from an animal gut metagenome**. *Biol Direct* 2015, **10**:19.

42. Maumus F, Epert A, Nogué F, Blanc G: **Plant genomes enclose footprints of past infections by giant virus relatives**. *Nat Commun* 2014, **5**:4268.

43. • Mushegian A, Shipunov A, Elena SF: **Changes in the composition of the RNA virome mark evolutionary transitions in green plants**. *BMC Biol* 2016, **14**:68.
This paper indicates that there was a transition to smaller viruses during the colonisation of land by plants which may give clues as to why known giant viruses and virophages only infect single celled eukaryotes.

44. Luppi M, Marasca R, Barozzi P, Ferrari S, Ceccherini-Nelli L, Batoni G, Merelli E, Torelli G: **Three cases of human herpesvirus-6 latent infection: integration of viral genome in peripheral blood mononuclear cell DNA**. *J Med Virol* 1993, **40**:44-52.

45. Morissette G, Flamand L: **Herpesviruses and chromosomal integration**. *J Virol* 2010, **84**:12100-12109.

46. Aswad A, Katzourakis A: **The first endogenous herpesvirus, identified in the tarsier genome, and novel sequences from primate rhadinoviruses and lymphocryptoviruses**. *PLoS Genet* 2014, **10**:e1004332.

47. Delaroque N, Maier I, Knippers R, Müller DG: **Persistent virus integration into the genome of its algal host, *Ectocarpus siliculosus* (Phaeophyceae)**. *J Gen Virol* 1999, **80(Pt 6)**:1367-1370.

48. Meints RH, Ivey RG, Lee AM, Choi TJ: **Identification of two virus integration sites in the brown alga feldmannia chromosome**. *J Virol* 2008, **82**:1407-1413.

49. Blanc G, Duncan G, Agarkova I, Borodovsky M, Gurnon J, Kuo A, Lindquist E, Lucas S, Pangilinan J, Polle J *et al.*: **The Chlorella variabilis NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex**. *Plant Cell* 2010, **22**:2943-2955.

50. Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, Mayer C, Miller J, Monier A, Salamov A *et al.*: **Pan genome of the phytoplankton Emiliania underpins its global distribution**. *Nature* 2013, **499**:209-213.

51. Iyer LM, Abhiman S, Aravind L: **A new family of polymerases related to superfamily A DNA polymerases and T7-like DNA-dependent RNA polymerases**. *Biol Direct* 2008, **3**:39.

52. • Koonin EV, Krupovic M, Yutin N: **Evolution of double-stranded DNA viruses of eukaryotes: from bacteriophages to transposons to giant viruses**. *Ann N Y Acad Sci* 2015, **1341**:10-24.
Proposes an evolutionary relationship between hypothetical polinto-viruses and adenoviruses, bidnaviruses, NCLDVs, polintons, tectivirus bacteriophages, various plasmids, and virophages.

53. Krupovic M, Bamford DH, Koonin EV: **Conservation of major and minor jelly-roll capsid proteins in Polinton (Maverick) transposons suggests that they are bona fide viruses**. *Biol Direct* 2014, **9**:6.

54. Hayward A, Cornwallis CK, Jern P: **Pan-vertebrate comparative genomics unmasks retrovirus macroevolution**. *PNAS* 2015, **112**:464-469.

55. Cui J, Tachedjian G, Wang LF: **Bats and rodents shape mammalian retroviral phylogeny**. *Sci Rep* 2015, **5**:16561.

56. Diehl WE, Patel N, Halm K, Johnson WE: **Tracking interspecies transmission and long-term evolution of an ancient retrovirus using the genomes of modern mammals**. *Elife* 2016, **5**:e12704.

57. Dupressoir A, Vernochet C, Bawa O, Harper F, Pierron G, Opolon P, Heidmann T: **Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene**. *PNAS* 2009, **106**:12127-12132.

58. Chuong EB, Elde NC, Feschotte C: **Regulatory evolution of innate immunity through co-option of endogenous retroviruses**. *Science* 2016, **351**:1083-1087.

59. Katzourakis A, Gifford RJ, Tristem M, Gilbert MT, Pybus OG: **Macroevolution of complex retroviruses**. *Science* 2009, **325**:1512.

60. •• Aiewsakun P, Katzourakis A: **Marine origin of retroviruses in the early Palaeozoic Era**. *Nat Commun* 2017, **8**:13954.
Describes for the first time the origin of retroviruses >450 million years ago demonstrating the utility of paleovirology in deciphering the origins of ancient lineages.

61. Houldcroft CJ, Breuer J: **Tales from the crypt and coral reef: the successes and challenges of identifying new herpesviruses using metagenomics**. *Front Microbiol* 2015, **6**:188.

# Chapter 2
# Footprints of Archaic HLA Variants in Modern Native Populations Suggest Distinct HLA-A Selection Pressures

**Footprints of archaic HLA variants in modern native populations suggest distinct HLA-A selection pressures**

Samuel Campbell [1,2], Nicolaas C. Kist [1,2], Daniel Lunn [3], Astrid K.N. Iversen [1*]

[1] Nuffield Department of Clinical Neurosciences, Division of Clinical Neurology, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, United Kingdom.

[2] Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, United Kingdom.

[3] Department of Statistics, University of Oxford, St Giles', Oxford OX1 3LB, Oxford, United Kingdom.

**ABSTRACT**

Archaic HLA alleles have been identified in the genomes of three Neanderthals and one Denisovan, and HLA-B*73:01 has been proposed to be of archaic origin. Here we estimated the average relative and actual archaic HLA frequencies worldwide in modern native populations and found that distinct selective forces appear to act on HLA-A, HLA-B, and HLA-C. The HLA-A frequency was several times higher than that of HLA-B and HLA-C, suggesting that additional selective forces act on HLA-A. We hypothesise that one such selective force could be the need to control human endogenous retrovirus (HERV) expression. Combinations of archaic HLA variants in modern native populations clustered according to geographic location and often with populations with known shared ancestry. North-to-south clines were observed in Europe for Neanderthal and HLA-B*73:01-recombinants, and in China for Neanderthal, Denisovan, and HLA-B*73:01-recombinants. The composition of archaic HLA variants was distinct for each population. Our results underscore that ancient combinations of pathogens have selected for distinct HLA frequencies in modern native populations and that these still carry a substantial local survival advantage.

**AUTHORSHIP NOTE**

Statistical analysis was done with the assistance of the University of Oxford Department of Statistics consultancy service.

## INTRODUCTION

The genes of the human leucocyte antigen (HLA) are some of the most polymorphic genes in the genome, and thousands of different alleles have been identified at the three primary HLA class I loci; HLA-A (>6000), HLA-B (>7500), and HLA-C (>6000) (Robinson et al., 2015). This polymorphism is characterized by a high level of sequence variation in the peptide-binding groove of HLA molecules and an enhanced rate of non-synonymous (NS) mutations (Hughes and Nei, 1988; Parham, 1988; Reche and Reinherz, 2003). Each HLA variant has specific binding motifs that allow for binding of only a subset of the intracellular pathogen-derived peptides, or tumour neo-antigens, produced by the conserved cellular antigen-processing machinery.

HLA class I molecules are central to the regulation of both adaptive and innate immune responses. HLA class I present intracellular ‘non-self’-derived peptides to cytotoxic T lymphocytes (CTL) and upon recognition of the foreign peptide or tumour neo-antigen, CTLs kill the infected cell. Some HLA molecules also function as ligands for specific natural killer (NK) cell receptors. NK cells constantly contact other cells and receive a balance of signals from activating and inhibitory receptors on the NK cell surface. The combined signal, and downregulation, or loss, of HLA class I presentation on infected cells or tumour cells, result in NK cell lysis of the abnormal cell.

Many HLA polymorphisms are ancient and allele lineages can predate species divergence and be retained across multiple species. The HLA system (Klein, 1987) and ABO histoblood group (Segurel et al., 2012) are the best described

forms of transspecies-polymorphism in humans. The degree of HLA-A and HLA-B polymorphism is related to pathogen diversity, primarily virus diversity (Prugnolle et al., 2005), and selection pressures imposed by viruses have been estimated to account for ~30% of all amino acid changes in the human proteome (Enard et al., 2016). In contrast to HLA-A and HLA-B, HLA-C plays a role in reproduction and the interplay between paternal and maternal HLA-C variants on the trophoblast and maternal NK-cells are essential for successful placentation (Parham and Guethlein, 2018).

HLA-A, HLA-B, and HLA-C are subject to multi-allelic balancing selection which, with recombination, act to preserve functional, beneficial genetic diversity. The exact mechanism of balancing selection is under debate, but three main processes have been proposed: overdominance (i.e., heterozygote advantage) (Doherty and Zinkernagel, 1975), frequency-dependent selection (i.e., rare allele advantage) (Bodmer, 1972), and fluctuating selection over time and space (Hill, 1991; Key et al., 2014). These processes are not mutually exclusive and might interact with each other. As modern human populations dispersed around the globe and encountered novel diverse local environments, positive directional selection of newly occurring, or previously neutral, mutations in HLA and recombination resulted in population, and/or regional, adaptation (Quintana-Murci, 2019).

Beneficial HLA variation can also be acquired through admixture and adaptive introgression of HLA alleles between modern human populations (Patin et al., 2017) and between modern human populations and archaic humans (Abi-

Rached et al., 2011; Fu et al., 2015). Such interbreeding events between different human lineages appear to be frequent as genomic analyses of three modern humans from the time when Neanderthals lived in Eurasia (~40,000 years before present (YBP) or earlier) revealed that one (33%) had a Neanderthal ancestor four to six generations back in the family tree (Fu et al., 2015). Likewise, reports suggest that interbreeding and adaptive introgression between archaic humans occurred in regions where the two overlapped (**Figure 2.1**). For example, in the genome of two of the six individuals (33%) from Denisova cave, Altai mountains, interbreeding (Meyer et al., 2012; Slon et al., 2018) and introgression of the HLA region (Meyer et al., 2012) has been documented. Traces of interbreeding between Denisovans and a superarchaic human lineage (possibly *Homo erectus*) has also been observed (Prufer et al., 2014).

The earliest modern human fossils outside of Africa were discovered in Misliya Cave, Israel, and are dated to around 177,000 to 194,000 YBP (Hershkovitz et al., 2018). However, this early migration seems not to have made a lasting impact and the major dispersal of modern humans out-of-Africa appears to have occurred only ~60,000 YBP (Reich, 2018). The ancestors to all modern humans outside of Africa mated with Neanderthals in the Levant. As modern humans ventured further into Eurasia about ~45,000 to 35,000 YBP, interbreeding occurred with local Neanderthal (Fu et al., 2015) and Denisovan populations (Jacobs et al., 2019). The latter belonged to at least three distinct lineages that predominated in Siberia (Altai mountains) and North East Asia; South and East Asia and parts of Oceania; and Papua New Guinea (Jacobs et al., 2019; Reich, 2018) (**Figure 2.1**). Because these archaic humans had already adapted to

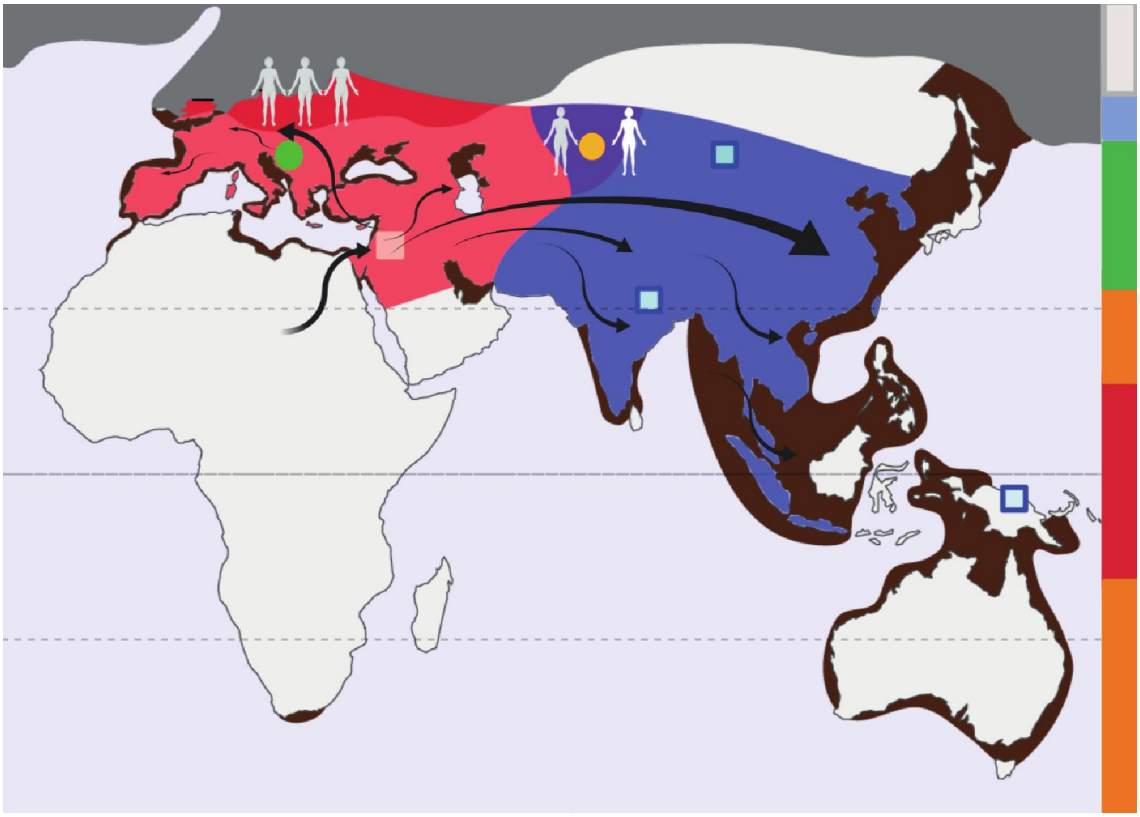


*Figure 2.1* Outline of the global Neanderthal and Denisovan distribution and the main modern human migration. The approximate distributions of Neanderthals (red) and Denisovans (blue) are adapted from (Mafessoni, 2019). The ice sheet (dark grey) shown is the estimated maximum glacial extent during one of the several glacial maximums (Hughes, 1985) that occurred during the ~700,000 years the ancestors to Neanderthals and Denisovans and later each human form lived in Eurasia. The dark brown land areas indicate now submerged territory which were above water during early human migrations (e.g., Sundaland in south east Asia, Doggerland in Europe, and the region between North Africa and Spain). These might have served as habitats for archaic humans during some of this period, and may have provided land-bridges to what are now islands (e.g., Japan, Taiwan, and other islands in south east Asia). Arrows indicate the main modern human migration out-of-Africa ~60,000 YBP and the subsequent dispersal throughout Eurasia. The green dot signifies Vindija Cave, Croatia, where remains of the three Neanderthals used in this study were found (Green et al., 2010). The orange dot shows Denisova Cave in the Altai Mountains, Russia, where the genomic sequences from one Denisovan (Reich et al., 2010), and one Neanderthal originate (Prufer et al., 2014). The red square signifies the Levant where Neanderthals and the ancestors to all modern humans in Eurasia interbred. The light blue squares show the approximate position of interbreeding between modern humans and three distinct Denisovan lineages. Colour bar (right) illustrate estimated differences in temperature, from red (hottest) to blue (coldest). Heavy dashed line is the equator. Light dashed lines are the tropics of Cancer and Capricorn

Eurasian pathogens for ~700,000 to ~450,000 years, adaptive introgression of archaic HLA variants likely provided modern humans with pre-adapted HLA variants that were crucial to their survival in the new environment.

Here we investigated the frequency and distribution of 20 archaic HLA variants from four Neanderthals and one Denisovan in 150 global 'modern native populations' (MNP) from which HLA-A, HLA-B, and HLA-C data were available. These HLA variants comprise less than 0.1% of all known HLA class I alleles. These analyses were extended to include the proposed archaic HLA-B*73:01 allele and recombinants carrying parts of this gene (Abi-Rached et al., 2011). HLA-B*73:01 is a unique major histocompatibility complex (MHC)-BII allele more closely related to chimpanzee and gorilla MHC-B alleles than to other modern human B alleles.

We found that distinct selective pressures acted on HLA-A, HLA-B, and HLA-C. Furthermore, we examined how MNP clustered based on the combination of archaic HLA frequencies, which is an indicator of the shared, or divergent, population immune responsiveness conveyed by these HLA variants. Although other factors, e.g., cultural transitions such as the advent of agriculture ~12,000 YBP exposed humans to novel pathogens, our results highlight that ancient shared and distinct selective pressures act on humans in different regions of the world.

## RESULTS

### The average relative frequency of archaic HLA variants vary geographically

We focussed on the ten Neanderthal and eight Denisovan HLA variants identified by Abi-Rached et al. (2011) in the genome sequences of three related female Neanderthals from Vindija Cave, Croatia (Green et al. 2010), and one female Denisovan from Denisova Cave in the Altai Mountains, Russia (Reich et al., 2010; Abi-Rached et al. 2011) (**Figure 2.1**, **Table S2.1**). The Neanderthal samples date

from ~44,500-38,000 YBP and the Denisovan from ~50,000-30,000 YBP. Some HLA variants, e.g., HLA-A*26 and HLA-A*66, were reported to be equally likely and were both included in our analyses, whereas possible archaic HLA variants identified only through linkage disequilibrium (LD) analyses were not. We HLA typed a high-resolution genome from a Neanderthal woman from the Denisova Cave, Altai Mountains (Prufer et al., 2014) (**Methods**). The fossil could not be dated directly, but was found in a >50,000 years old cave layer (Prufer et al., 2014), and sequence analysis has dated the sample to ~120,000 YBP (Slon et al., 2018). Despite the separation of time and space between the samples the four HLA variants we identified were similar to those found in the Vindija Neanderthal DNA samples either directly (HLA-B*51:01 and HLA-C*16:01) or through LD decay analyses (HLA-A*24:02 and HLA-A*31:01) (Abi-Rached et al., 2011). The latter HLA variants were therefore included as Neanderthal HLA variants together with HLA-A*24:03 as our analyses did not clearly distinguish between HLA-A*24:02 and HLA-A*24:03 (**Table S2.1**).

To avoid confounding our analyses by the impact of well-documented large-scale migrations and admixture between modern populations during the last ~500 years, we limited the examined populations to 150 MNP from which HLA-A, HLA-B, and HLA-C data were available **(Methods)**. The populations were grouped into 11 geographic regions using hand-curated database associated location coordinates; Africa (AF), North Africa (NAF), Europe (EU), Western Asia (WA), South Asia (SA), North East Asia (NEA), South East Asia (SEA), Oceania (OC), Australia (AU), North America (NAM), and South America (SAM) **(Methods**).

To estimate the global distribution of archaic HLA variants, we examined the relative frequency of Neanderthal and Denisovan HLA alleles in each region (**Figure 2.2a, b**). Eleven of the twelve Neanderthal HLA variants were observed in all or most regions including Africa (95%). The exception was HLA-B*07:03, which was only found in 23 people in the WA and EU. (**Figure 2.2a**). In contrast, only five of the eight Denisovan HLA alleles were global or present in most regions (56%); one of these, HLA A*02:01, was also identified in the Neanderthal sequences. Three HLA variants were predominately Asian with the highest frequencies in SEA, and an extremely rare HLA variant, HLA-B15:58 was only found in SEA (**Figure 2.2b**). Only three Denisovan HLA variants were observed in Africa. These results mirror the accepted view that much of the Neanderthal DNA in modern humans came from interbreeding events that took place ~60,000 to 40,000 YBP in the Levant and that only some modern human ancestors interbred with Denisovans as they ventured into Asia thousands of years later (Reich, 2018).

The proposed archaic HLA-B*73:01 allele was observed primarily in WA (**Figure 2.2c**, blue square), but was also found in AF, EU, NEA, and SA. Some human HLA alleles acquired fragments of HLA-B*73:01 through recombination either in archaic or modern humans (**Figure 2.2c, d, e, S2.2**). These recombinants all carry a methionine at position -21 (-21M) in the leader sequence whereas most HLA-B variants carry threonine in this position (-21T). The leader sequence is processed into peptides in the cell and a peptide is produced which, if it contains -21M, can bind to HLA-E. This binding ensures correct folding of HLA-E and results in HLA-E expression on the cell surface. HLA-E is a ligand for NK cell receptor CD94:NKGA and the amount of HLA-E on the cell surface affects the

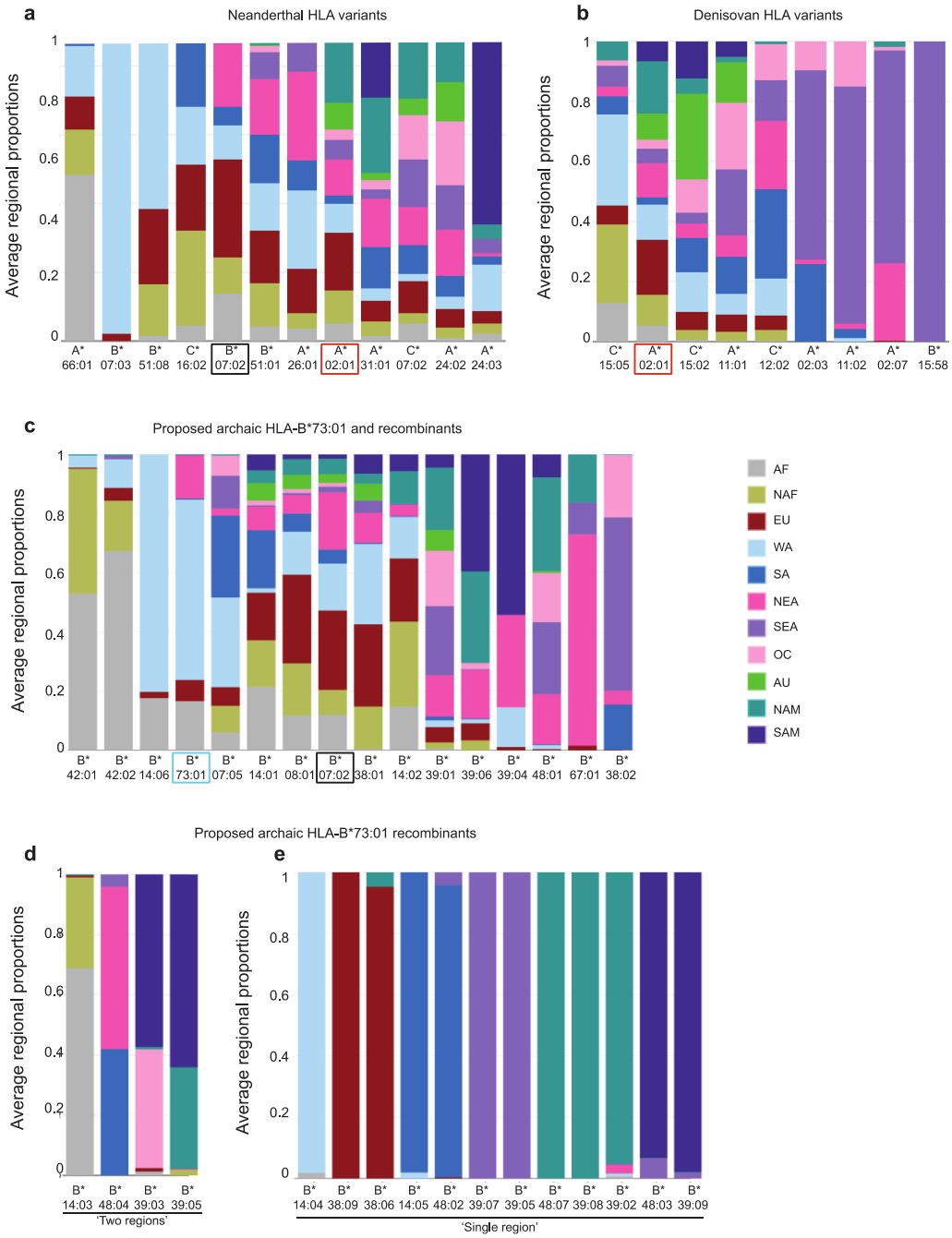



*Figure 2.2* The average relative frequency of archaic HLA variants. **a**, **b**. The average relative frequency of Neanderthal (**a**) and Denisovan (**b**) HLA variants worldwide. The black box highlights HLA-B*07:02, which also carry parts of HLA-B*73:01 (see **c**); Red box highlights HLA-A*02:01, the only HLA variants shared between the Neanderthals and the Denisovan in this study. **c**, **d**, **e**. The average relative frequency of HLA-B*73:01 (blue box) and recombinants between HLA-B*73:01 and other HLA alleles worldwide. The black box highlight HLA-B*07:02, which is also found in the Neanderthal genomes. Geographical regions were colour coded as outlined in the figure. AF = Africa, NAF = North Africa, EU = Europe, WA = Western Asia, SA = South Asia, NEA = North East Asia, SEA = South East Asia, OC = Oceania, AU = Australia, NAM = North America, and SAM = South America.

education and functionality of the NK cell repertoire in the host. The functionality of NK cells affect their ability to combat diverse pathogens.

The HLA-B*73:01 recombinants could be divided into three main groups according to their global distribution patterns. Group 1 was found in three or more regions at relative high frequency (**Figure 2.2c**); this group could be further subdivided into those found primarily in AF, NAF, EU, and WA, those found in almost all regions, and some found predominately in NEA, SEA, OC, NAM, and SAM. Group 2 was found mostly in only two regions (**Figure 2.2d**). Of the four HLA variants in this group, three were found in adjacent geographical locations (AF/NAF, SA/NEA, NAM/SAM) whereas one was found almost only in OC and SAM. This is striking and in line with the observation of some Austronesian ancestry in SAM (Moreno-Mayar et al., 2018). Group 3 consisted of twelve HLA variants almost exclusively present in one geographical location; the only regions without this exclusive local adaptation pattern were AF, NAF, NEA, and OC (**Figure 2.2e**).

In combination, these results highlight the wide dispersal of most Neanderthal and some Denisovan HLA variants, the local selection in SEA, NEA, and OC of some Denisovan HLA-A* variants and one rare HLA-B variant, and the variable selective pressures on HLA-B*73:01 and recombinants thereof worldwide. These patterns likely reflect the variable selective pressures from combinations of ubiquitous and local pathogens and the degree to which the HLA variants can present peptides from vulnerable parts of these pathogens and the extent to which the NK response participates in combatting them.

**Distinct selective forces act on HLA-A, HLA-B, and HLA-C evolution**

We next examined the Neanderthal and Denisovan HLA-A, HLA-B, and HLA-C frequency differences between populations within each geographical region by

summing the frequencies of each HLA group (**Figure 2.3a, b**). This analysis strongly suggests that distinct selective factors have acted on each group.

The Neanderthal HLA-A alleles were found worldwide. The combined HLA-A frequencies in non-African populations generally varied between 0.25 and 0.75, with few populations carrying less and some up to 0.95. The percentage of individuals with one of these alleles therefore were between 50-100% in most populations (**Figure 2.3a**). The combined HLA-A frequency was highest in SEA-Taiwanese indigenous populations and typically high in most of NEA, OC, and NAM and SAM. Some NAM and SAM populations, e.g., the Arizona Gala River Amerindian and the Venezuela Perja Mountain Bari populations, had frequencies comparable to those found in indigenous Taiwanese. WA, SA, and SEA-Chinese populations, excluding SEA-Tibetans, displayed the lowest frequencies, while the EU frequencies were intermediate at around 0.4. A similar pattern was found for the Denisovan HLA-A frequencies, although the maximum and overall allele frequencies were lower (**Figure 2.3b**). Denisovan HLAs in western Eurasian populations (e.g. Europe, Western Asia, and North Africa) likely represent genetic overlap with Neanderthals from their closely shared heritage, long-term regional cohabitation (**Figure 2.1**), and later back migrations by modern humans, as unlike Neanderthals, Denisovans are thought to have admixed only with eastern Eurasian populations. For the Denisovan HLA variants, the highest frequencies were observed in SEA-Chinese populations, not the SEA-Taiwanese populations, in contrast to the Neanderthal HLA-A pattern. Likewise, the frequencies in NEA populations were relatively low and comparable to the frequencies found in WA and SA. Only a few of the Oceanian populations, and

none of the NAM and SAM populations, had frequencies that exceeded 0.5. This result suggests that the selective forces working on HLA-A are and were to a large extent similar in all populations, and to a lesser extent local.

In contrast to the Neanderthal HLA-A distribution pattern, Neanderthal HLA-B were largely absent in several Eurasian regions (**Figure 2.3a**). The HLA-B allele frequencies were typically ~0.20 in EU and less than 0.30 in WA although significant variation was observed in this region. The frequency in all other populations was less than 0.2, with stable frequencies in NEA and highly variable frequencies in SEA and NAM. Neanderthal HLA-B variants were absent in most SEA-Taiwanese, OC, and SAM populations, despite the same populations having the highest Neanderthal HLA-A frequencies. The Denisovan HLA-B variant, HLA-B*15:58, has only been identified in a few individuals worldwide, primarily of SEA-Chinese ancestry, suggesting that this HLA variant is largely irrelevant to the survival of modern humans, for example it might be a local adaptation to ancient and possibly extinct pathogens in the Altai mountains which was lost through genetic drift (**Figure 2.3b**). In combination, these results suggest that the selective factors working on HLA-B are and were predominately local.

Neanderthal HLA-C variants were present throughout all regions and were observed in almost all populations (**Figure 2.3a**). The frequencies generally ranged from 0.1-0.2, with some SEA, OC, and NAM and SAM populations displaying higher frequencies, although rarely more than 0.3. Three populations had extreme frequencies, e.g., the SAM Yucpa population. A similar pattern was found for Denisovan HLA-C variants, although for these, the EU frequencies were

typically lower than in the rest of Eurasia, in contrast to the Neanderthal HLA-C pattern (**Figure 2.3b**). The frequencies were also low in NAM and absent or low in SAM, except for the extreme frequency of 0.49 in SAM North Chimila Amerindians. While Neanderthal HLA-C was not observed, or found at low frequencies in AU, the Denisovan HLA-C frequencies were relatively high in all AU populations. The Denisovan HLA-C variants were also absent in most African populations. These patterns underscore that Asian, Oceanian and Australian populations interbred with Denisovans, while this was not generally the case for the ancestors of the EU populations. This result highlights that similar to HLA-A, the selective forces working on HLA-C were and are to a large extent similar in all population and to a lesser extent local.

The combined frequency of HLA-B*73:01 and recombinants thereof were between 0.1 and 0.2 in most populations worldwide (**Figure 2.3c**). In contrast to the Neanderthal and Denisovan HLA frequencies, the combined frequency of HLA-B*73:01 and recombinants were similar to or higher in Africa than in most Asian populations. This might indicate that HLA-B*73:01 and recombinants thereof were present in modern humans when they migrated out of Africa, or that these HLA variants were present in the common ancestor to Neanderthals/Denisovans and modern humans. These results also suggest that it is beneficial to the survival of populations worldwide that part of the population carry NK cells educated through an augmented interaction between HLA-E and the CD94:NKGA receptor on NK cells.

Some of the differences in archaic HLA-A, HLA-B, and HLA-C frequencies between regions and within regions likely reflect the population bottlenecks that occurred

during the migration of humans with introgressed Neanderthal HLAs from the Levant to South America and adaptation of isolated populations to environmental pathogens. The similar selective pressures affecting HLA-C are due to its crucial role in placentation and the shared variables of human anatomy and physiology

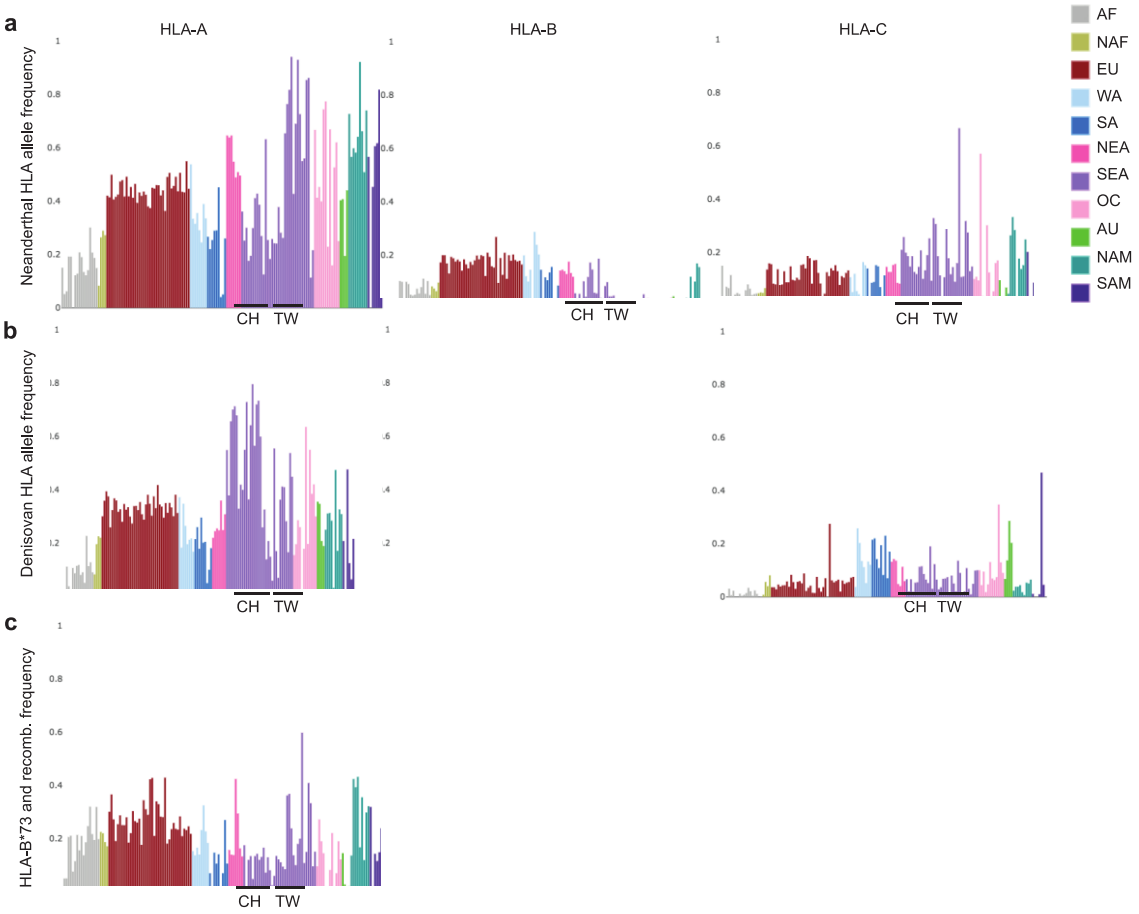



*Figure 2.3* Worldwide archaic HLA-A, HLA-B, HLA-C frequencies. **a**. The sum of Neanderthal HLA-A, HLA-B, and HLA-C variants in each modern native population. Distinct differences were observed between SEA-Chinese and SEA-Taiwanese populations as the latter displayed the highest HLA-A frequencies, both had very low levels, if any, of HLA-B, while the HLA-C frequencies were comparable. **b**. The sum of Denisovan HLA-A, HLA-B, and HLA-C variants in each modern native population. Distinct differences were observed between SEA-Chinese and SEA-Taiwanese populations. In contrast to the results in **a**, SEA-Chinese populations displayed the highest combined HLA-A frequencies worldwide, while the SEA-Taiwanese populations varied in HLA-A frequency and displayed the same pattern as OC populations. The Denisovan HLA-B variant was only found in two SEA-Chinese individuals. The overall level of HLA-C was comparable in both populations. **c**. The sum of HLA-B*73:01 and recombinants thereof in in each modern native population. The overall level was higher than for the Neanderthal HLA-B variants and although frequencies were higher in AF and EU populations, these HLA variants were mostly present at a frequency >0.1 and were only absent in a few populations. CH denotes China, TW denotes Taiwan. Geographical regions were colour coded as outlined in the figure; AF = Africa, NAF = North Africa, EU = Europe, WA = Western Asia, SA = South Asia, NEA = North East Asia, SEA = South East Asia, OC = Oceania, AU = Australia, NAM = North America, SAM = South America.

(Parham and Guethlein, 2018). Poor placentation can result in miscarriage, preterm labour, foetal growth restriction, stillbirth, and potentially fatal preeclampsia and eclampsia in the pregnant woman. This interplay has been studied in detail in the SAM Yucpa population where the abundance of Neanderthal HLA-C*07:02 has induced changes in the killer-cell immunoglobulin-like receptor (KIR) repertoire to optimise the chances of a successful pregnancy (Parham and Guethlein, 2018).

**The distinct relationships between HLA-B frequency and latitude**

Our results suggest that the combination of selective forces acting on HLA-A were different from those on HLA-B (**Figures 2.4**). Previously, pathogen-driven multi-allelic balancing selection has been hypothesized to explain the polymorphism in the HLA complex as the diversity of HLA-A and HLA-B genes correlate significantly with pathogen diversity, defined as the number of disease agents in a country (Prugnolle et al., 2005). Pathogen diversity has in turn been shown to be negatively associated with the distance from the Equator, although the local environments have modifying effects, with, e.g., temperature, seasonal dry extremes, and frost-free climate increasing pathogen prevalence (Cashdan, 2014; Stephens et al., 2016). As 30,000 to 120,000 YBP encapsulates several climate changes and glacial maxima, it defies a concise accurate description. However, it is fair to propose that in general the Neanderthals in Croatia lived in a temperate Mediterranean climate and the Neanderthals and Denisovans in the Altai mountains in a temperate mountain zone. Because of this, the HLA variants we investigate have all been adapted to combat pathogens prevalent in temperate or colder climates.

In line with these proposed HLA adaptations to pathogens that thrive in a

temperate climate, we found a significant positive relationship between the amount of Neanderthal HLA-B variants and distance from the Equator, ($R_p^2$ =

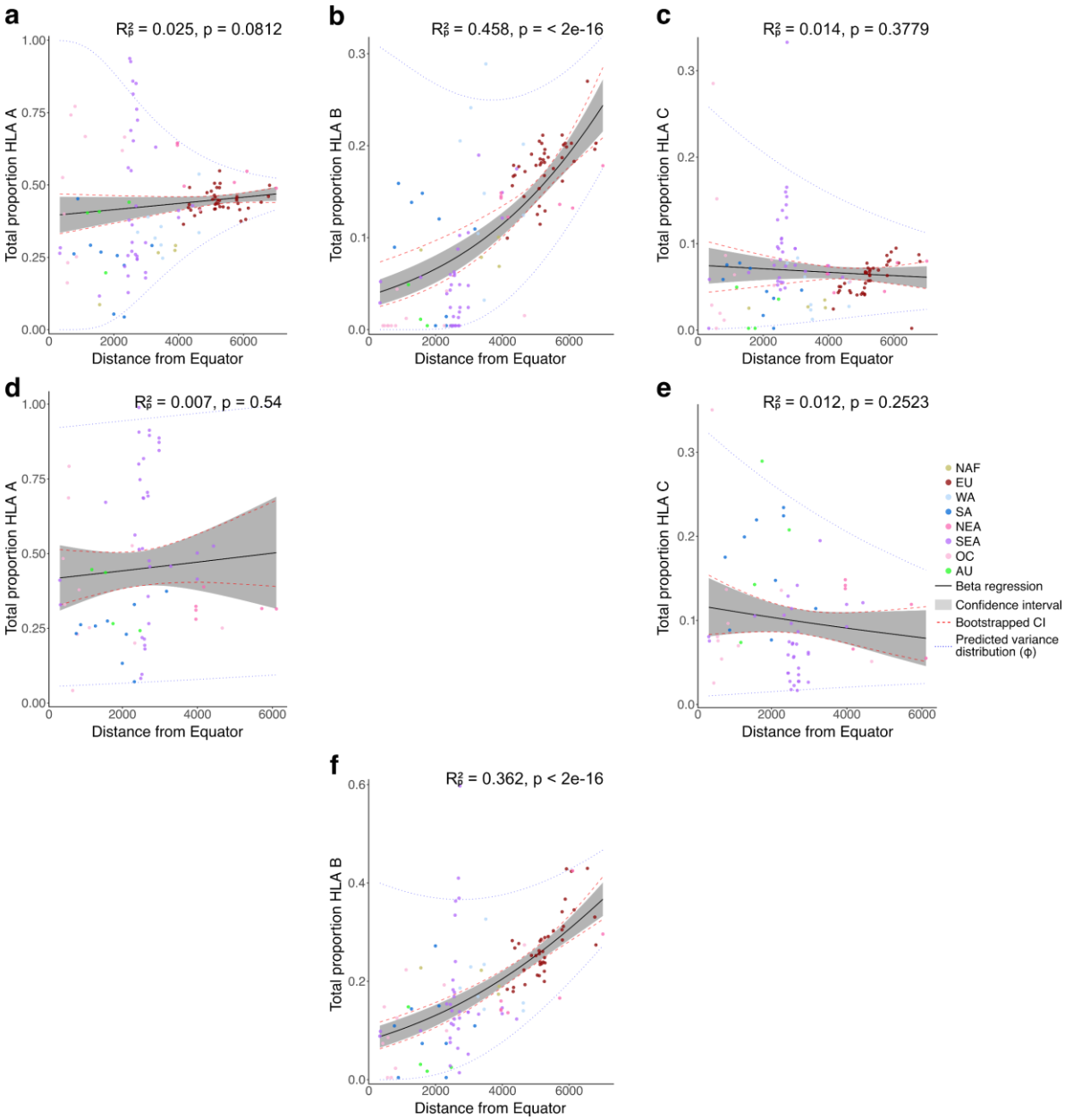



*Figure 2.4* Distance from the Equator versus archaic HLA frequencies in modern native populations. **a**, **b**, **c** Neanderthal HLA A*, B*, and C* respectively. **d**, **e** Denisovan HLA A* and C* respectively. **f** HLA-B*73:01 and recombinants thereof. Regressions are beta regressions (see methods) of the distance from the Equator against the total amount of each respective HLA in a population. Grey shaded area is the confidence interval. Red dashed lines are the upper (97.5%) and lower (2.5%) quantiles of nonparametric bootstraps with 10,000 replicates. Blue dotted lines are for visualising the variance distribution and are the upper (97.5%) and lower (2.5%) quantiles of the predicted beta distribution. p-values are for the mean model (i.e. what one would see in a normal linear regression). φ p-values for the variance model were significant for non-Denisovans (see **Table S2.5**). $R_p^2$ is pseudo R squared, a global measure of explained variation. Geographic regions are colour coded as outlined in the figure. NAF = North Africa, EU = Europe, WA = Western Asia, SA = South Asia, NEA = North East Asia, SEA = South East Asia, OC = Oceania, AU = Australia.

0.46, p < 2e-16) (**Figure 2.4b**). A similar pattern was found for HLA-B*73:01 and recombinants thereof, although the associations were weaker ($R_p^2$ = .36, p = 2.2e-8) (**Figure 2.4f**). Allowing for variable dispersion to be modelled separately from the mean showed a statistically significant negative relationship between distance from the Equator and the precision parameter (φ) for the non-Denisovan HLAs, i.e. there was significantly less variability in the total amount of archaic HLAs for populations further away from the Equator (**Figure 2.4ac & f**). This was true even if there was no significant link between the mean amount and distance from the Equator (see **Table S2.5** for φ p-values).

**Multiple regression of distance and pathogen diversity**

Next we carried out a multiple regression adding distance from the Levant (Manot cave) and intracellular pathogen richness (the number of pathogens listed in a country) (**Methods**) to our models. We looked at viruses, intracellular bacteria (both obligate and facultative), and both viruses and intracellular bacteria pooled together (**Table 2.1**, see **TABLE S2.6** for φ p-values). For each pathogen set, the amount of Neanderthal HLA A in a population was significantly correlated with both distance from Manot cave and the Equator, but was not correlated with pathogen richness. For Denisovans, HLA A was significantly correlated with distance from the Equator when regressing total pathogen and bacteria richness but not viruses. Denisovan HLA A was correlated with pathogen richness and was not correlated with distance from Manot cave for all pathogen datasets. For both Neanderthal HLA B and HLA-B*73:01 and recombinants thereof, the amount of archaic HLAs in a population was significantly correlated with both distance from the Equator and pathogen richness for all pathogen datasets. HLA B was also significantly correlated with distance from

*Table 2.1* Pseudo R squared and p-values for beta regression analyses of the relationship between Archaic HLA frequency, distance from Manot cave, distance from the equator, and pathogen species richness

| | HLA A | HLA B | HLA C |
|---|---|---|---|
| **Neanderthal HLAs** | | | |
| $R_p^2$ | 16% | 49% | <1% |
| Pathogens | 0.36$^{NS}$ | 9.05e-05*** | 0.56$^{NS}$ |
| Dist.Equator | 4.85e-06*** | 1.18-15*** | 0.21$^{NS}$ |
| Dist.Manot Cave | 6.74e-04*** | 5.85e-03** | ***6.44e-04****** |
| $R_p^2$ | 17% | 53% | <1% |
| Bacteria | 0.40$^{NS}$ | 0.02* | 0.45$^{NS}$ |
| Dist.Equator | 1.76e-04*** | 2.42e-14*** | 0.21$^{NS}$ |
| Dist.Manot Cave | 1.27e-03** | 1.25e-03** | ***5.39e-04****** |
| $R_p^2$ | 15% | 49% | <1% |
| Viruses | 0.52$^{NS}$ | 1.06e-05*** | 0.71$^{NS}$ |
| Dist.Equator | 6.14e-07*** | 2.46e-16*** | 0.21$^{NS}$ |
| Dist.Manot Cave | 4.11e-04*** | 2.63e-03*** | ***4.24e-04****** |
| **Denisovan HLAs** | | | |
| $R_p^2$ | 12% | | 4% |
| Pathogens | 6.05e-3** | | 0.10$^{NS}$ |
| Dist.Equator | 0.04* | | 0.169$^{NS}$ |
| Dist.Manot Cave | 0.48$^{NS}$ | | 0.59$^{NS}$ |
| $R_p^2$ | 18% | | 2% |
| Bacteria | 7.10e-05*** | | 0.26$^{NS}$ |
| Dist.Equator | 1.18e-03** | | 0.13$^{NS}$ |
| Dist.Manot Cave | 0.89$^{NS}$ | | 0.93$^{NS}$ |
| $R_p^2$ | 7% | | 6% |
| Viruses | 0.03* | | 0.06$^{NS}$ |
| Dist.Equator | 0.19$^{NS}$ | | 0.26$^{NS}$ |
| Dist.Manot Cave | 0.17$^{NS}$ | | 0.48$^{NS}$ |
| **B*73:01 Recombinants** | | | |
| $R_p^2$ | | 38% | |
| Pathogens | | 5.83e-07*** | |
| Dist.Equator | | 4.68e-16*** | |
| Dist.Manot Cave | | ***0.03**** | |
| $R_p^2$ | | 36% | |
| Bacteria | | 3.74e-07*** | |
| Dist.Equator | | < 2e-16*** | |
| Dist.Manot Cave | | ***0.01**** | |
| $R_p^2$ | | 39% | |
| Viruses | | 3.27e-06*** | |
| Dist.Equator | | 4.98e-15*** | |
| Dist.Manot Cave | | 0.07$^{NS}$ | |

Dist.Manot Cave is distance from Manot Cave. Dist.Equator is the distance from the equator (i.e. a measure of absolute latitude). Pathogens is species richness of intracellular pathogens, both bacteria and viruses, in a country from which population was drawn. Bacteria and Viruses denote their respective individual species richness. Bold italicized text denotes values that were not significant when accounting for spatial autocorrelation. See **Table S2.6** for φ p-values $R_p^2$ is the pseudo R squared value, a global measure of variation explained by the model for beta regressions, akin to $R^2$ for linear models. p-values for asymptotic Wald tests: *** < 0.001; ** < 0.01; * < 0.05; NS, non-significant.

Manot cave for all pathogen datasets except B*73:01 compared to Virus richness. Neanderthal HLA C was correlated with distance from Manot cave while Denisovan HLA C was not. Neither Neanderthal nor Denisovan HLA C was correlated with distance from the Equator or pathogen richness.

For Neanderthal HLA A vs all pathogen sets and Denisovan HLA A vs total intracellular pathogens and intracellular bacteria only, distance from the Equator was significantly correlated with the amount of archaic HLA in a population in the multiple regression while it was not in the simple (single predictor) regression. This can be indicative of multicollinearity, where the predictor variables are not only correlated with the response variable but also highly correlated with each other. Testing this, the distance from the Equator is correlated with total, bacterial, and viral pathogen richness. Briefly, the mechanism for how this affects the significance is that in the simple regression, the not significant predictor (distance from the Equator) is significantly associated with an omitted predictor (for example, pathogen richness). Distance from the Equator then partially reflects the effect of the omitted pathogen richness in addition to its own effect. When pathogen richness is added to the multiple regression, distance from the Equator no longer captures the partial effects of pathogen richness and now better reflects the "true" effect of distance from the Equator. There are other confounding factors, such as local microclimate, which the model does not capture which further complicate the matter. The variance inflation factor (VIF) is a widely used measure of the degree multicollinearity, and the general rule of thumb is that a VIF over 5 requires further investigation while a VIF over 10 is a sign of severe multicollinearity and requires correction (Fox and Weisberg, 2019; Kim, 2019). All predictors in all of our models

had a VIF under 3, and the majority were under 2.

Earlier studies looking at similar data, e.g. Prugnolle et al. (2005), made no efforts to control for non-independence of the data points. Indeed, there is considerable variation in the extent to which analyses across populations within a species account for gene flow and phylogenetic non-independence, and the majority completely ignore both (Meirmans, 2012; Stone et al., 2011). Nevertheless, we have endeavoured to control these factors here. Most prescribed methods of controlling for these effects require gene sequences, phylogenetic data, and other factors that do not exist for the data in this study. For example, Felsenstein's (2002) "intraspecific contrasts method" requires the observed phenotype, the optimum phenotype, a scalar describing incremental per generation movement of the phenotype towards its local optimum, and the pairwise migration rates between all populations. This is perhaps useful for simulations but for real data where most of these things are unknowable, proxies need to be found. Fortunately, spatial autocorrelation can be used as a proxy to control for genetic structure and the homogenising influence of gene flow (Diniz-Filho et al., 2013; Epperson, 2005; Epperson and Li, 1996; Furstenau and Cartwright, 2016; Lao et al., 2008; Novembre et al., 2008; Sokal and Jacquez, 1991; Sokal and Wartenberg, 1983; Urban, 2011). Our data exhibited statistically significant levels of spatial autocorrelation as determined by Moran's I Autocorrelation Index. This is unsurprising given Tobler's (1970) first law of geography which states "everything is related to everything else, but near things are more related than distant things." We generated distance based Moran's eigenvector maps (dbMEM) (**Methods**) in order to account for this spatial autocorrelation (Borcard et al., 2004; Borcard and

Legendre, 2002; Dray et al., 2006; Legendre and Legendre, 2012) as this and related techniques have been regarded as the best approach to spatial regression (Diniz-Filho et al., 2009; Legendre and Legendre, 2012). Incorporating spatial autocorrelation as a predictor made the distance from Manot cave non-significant for Neanderthal HLA-C and B*73:01, but did not otherwise negate the significance of our other predictors.

The significant correlations between proportions of archaic HLAs A and B regressed against distance from the Equator plus intracellular pathogen diversity suggest that the selective pressure from exogenous pathogens vary with temperature, in general agreement with earlier work (McMichael, 1993; Prugnolle et al., 2005b; Quach and Quintana-Murci, 2017; Quintana-Murci, 2019; Riley and Olerup, 1992). These correlations exist even when controlling for distance from likely areas of initial introgression and spatial autocorrelation. These results also concur with the finding that the selective pressure on the HLA B gene is ~3 and 16 times stronger than on HLA-A and HLA-C, respectively (Prugnolle et al., 2005). In combination, these results strongly suggest that additional, so far unidentified, shared and stable selective forces act to maintain the high archaic HLA-A frequencies in Eurasians (**Figure 2.3**).

**Why are the selective factors acting on HLA-A different from HLA-B?**

Both HLA-A and HLA-B polymorphism have been demonstrated to be driven by pathogen diversity, primarily by virus diversity (Enard et al., 2016; Prugnolle et al., 2005), but strong selective factors acting on HLA-A, but not HLA-B, have not been described. Here we hypothesise that additional selective pressures acting

primarily on HLA-A could be imposed by the need to continuously suppress damaging HERV replication to prevent, e.g., tumour formation (**Figure 2.5**).

HERVs are typically relics of ancient germ line infections along the last ~100 million years of primate evolution and comprise >8% of the human genome (Jha et al., 2011; Wildschutte et al., 2016). HERV-K (HML-2) proviruses are an exception to the antiquity of most HERVs with ~120 human-specific insertions estimated to have occurred during the last 150,000 - 2 million years (Jha et al., 2011). Eleven human-specific HERV-K insertions have been observed in the genomes of the five archaic humans examined here (Agoni et al., 2012; Marchi

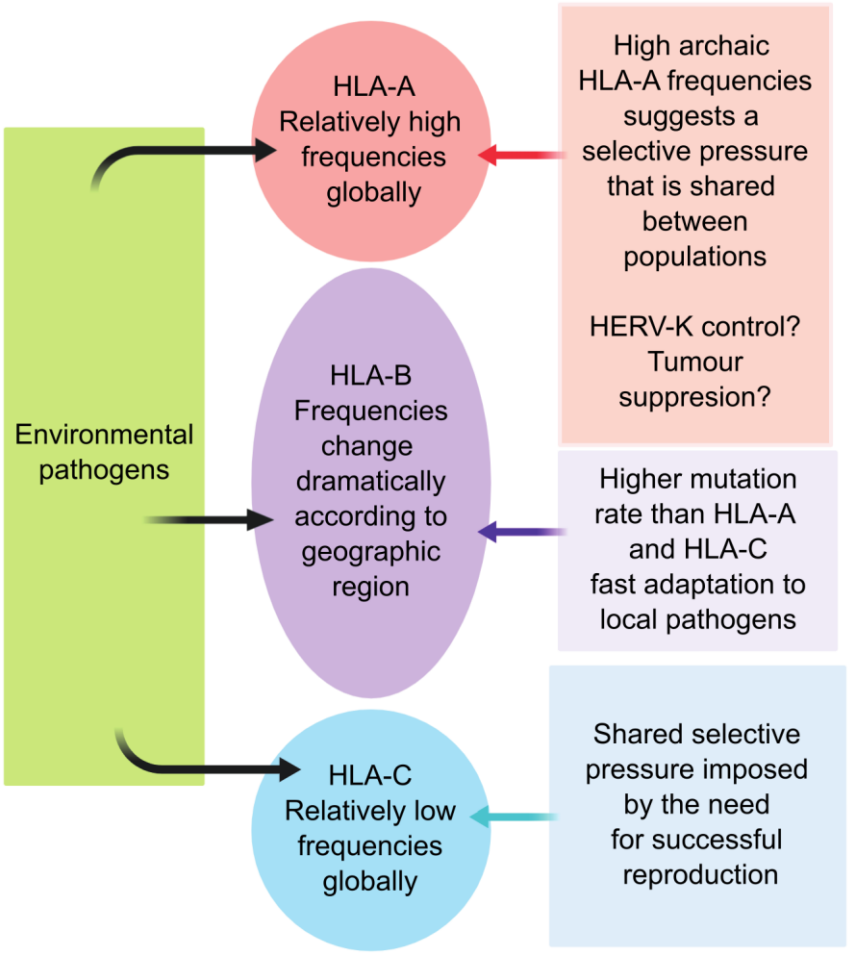



Figure 2.5 Outline of proposed selective pressures working on archaic HLA-A, HLA-B, and HLA-C. Different combinations of selective forces appear to be acting on different HLA types, resulting in different rates of retention and replacement in global populations.

et al., 2013; Wildschutte et al., 2016). Interbreeding with Neanderthals and Denisovans might have introduced novel, so far unidentified retroviral elements into Eurasians, might have provided populations with HERV-K insertions that they had otherwise lost through genetic drift, or may have increased the dose of certain HERV-K insertions explaining the significant differences in HERV-K insertion frequencies in global populations (Wildschutte et al., 2016).

Both intermediate and final products of endogenous retroelement replication are potentially immunogenic, and their protein products are recognizable by T cell receptors (TCRs), because of incomplete epigenetic silencing during thymic TCR selection. Beneficial HERV sequences and HERV activation can participate in shaping and modulating the adaptive and innate immune response (Grandi and Tramontano, 2018; Jern and Coffin, 2008; Kassiotis and Stoye, 2016). HERVs with IFN-γ-inducible enhancer sequences are found adjacent to IFN-γ-induced genes and are involved in the regulation of essential immune functions (Chuong et al., 2016). Replication intermediates of endogenous retroelements result in induction of an interferon response and cellular activation, which increases immune reactivity to unrelated antigens (Zeng et al., 2014). This heightened reactivity has been proposed to act as an ‘intrinsic adjuvant’ for immune responses to poorly immunogenic targets (Kassiotis and Stoye, 2016; Zeng et al., 2014).

However, damaging HERV activation contributes to the development of distinct inflammatory, autoimmune, and neoplastic diseases (Kassiotis, 2014; Volkman and Stetson, 2014). Infection with a range of exogenous, often ubiquitous, viruses can cause HERV activation, e.g., Epstein-Barr virus (EBV), cytomegalovirus

(CMV), herpes simplex virus 1 (HSV1), Kaposi's sarcoma-associated herpesvirus (KSHV), human immunodeficiency virus type 1 (HIV-1), human T-lymphotropic virus 1 (HTLV-1), and influenza A virus (IVA) (Chen et al., 2019). Many of these viruses cause chronic infections and, e.g., HIV-1 and CMV, selectively down-regulate HLA-A expression more than that of HLA-B and HLA-C (Rajapaksa et al., 2012; Zimmermann et al., 2019). Cellular transformation in cancer typically results in alterations of the epigenetic landscape and loss of repression of endogenous retroelements and can be targets for anti-tumour immunity (Cherkasova et al., 2013; Downey et al., 2015; Salmons et al., 2014; Szpakowski et al., 2009). Several groups of HERVs that are not frequently expressed in healthy cells, particularly the HERV-K group, show substantial expression in various cancers (Kassiotis, 2014). As regional pathogens and commensals in combination modulate the likelihood of HERV activation, selective pressure to adapt to these microorganisms add additional local selective pressure on the HLA region.

The archaic HLA-A variants might be selected for by a combination of selective pressures from exogenous regional pathogens, particularly viruses, and endogenous selective pressures to avoid excessive and damaging HERV expression (outlined in **Figure 2.6**). The offspring of a Neanderthal and a modern human would have a survival advantage when exposed to exogenous pathogens and would be able to destroy transformed cells and maintain healthy tissues.

To investigate if HLA-A played a greater role than HLA-B and HLA-C in the adaptive immune responses towards cells with inappropriate levels of HERV activation, we compiled a list of CTL responses against HERV (**Table S2.3**). We

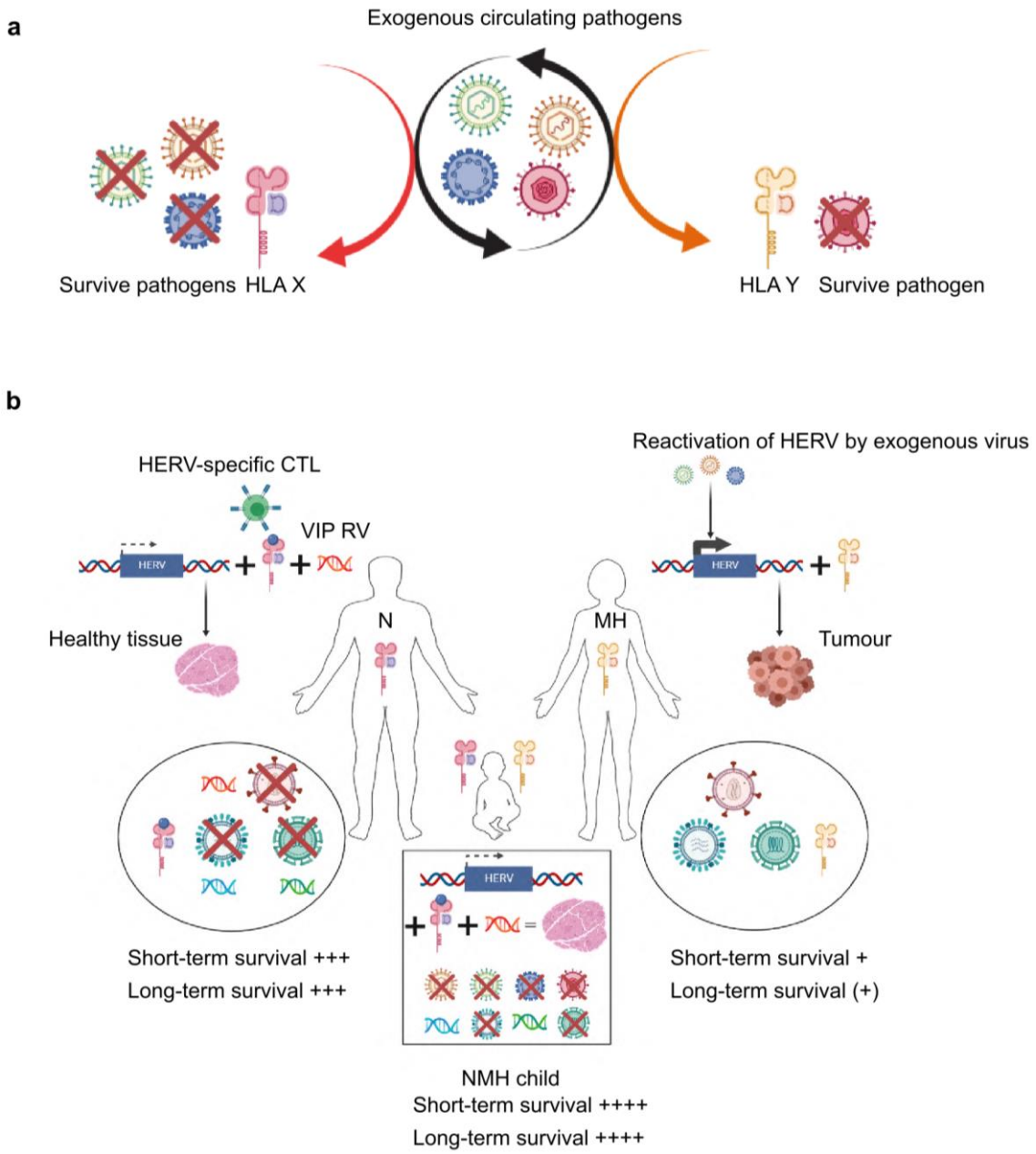



Figure 2.6 Outline of the modern human survival advantage from interbreeding with Neanderthals. **a**. Schematic illustration showing four circulating viruses and the impact of carrying two different HLA variants on surviving the infectious challenges. An individual with HLA X is likely to survive an infection with three of the four locally, circulating viruses as infection will be combatted successfully (signified by red crosses over the pathogens). In contrast, an individual without HLA X, but with HLA Y, will survive infection with only one of the four viruses. **b**. Neanderthals had lived in parts of Eurasia for ~700,000 years and were highly adapted to their habitat. Genomic analyses have demonstrated that modern humans have introgressed both beneficial archaic HLA variants and gene segments enriched for proteins that interact with viruses (VIPs). The most significant introgression is associated with VIPs that interact with HIV-1, or by extension, retroviruses (RV). The introgressed archaic HLAs and VIPs in combination are likely to protect against local pathogens and activation of human endogenous retroviruses (HERV), both directly through pathogen and HERV-specific cytotoxic T cell responses (CTL), and indirectly as successfully combatting exogenous viral challenges reduces the likelihood of HERV activation. A reduction of HERV activation reduces the risk of tumour formation. N= Neanderthal, MH = modern human, NMH = Neanderthal and modern human offspring.

found that all the CTL HERV-specific responses were restricted by HLA-A variants, primarily HLA-A*02:01, HLA-A*03:01, and HLA-A*11, HLA variants found in the five archaic humans in this study (Abi-Rached et al., 2011). This pattern was consistent regardless of whether we investigated the HERV-specific response in cancer, autoimmune or infectious diseases. The dominance of HLA-A restricted HERV-specific CTL responses is in stark contrast to, e.g., the immune response against HIV-1 where almost all protective responses are restricted by HLA-B variants (Goulder and Walker, 2012). Although this only provides indirect support of our hypothesis, it is nonetheless a striking finding.

Our hypothesis is also indirectly supported by a recent study demonstrating that virus-associated selective pressures have resulted in adaptive introgression of segments of Neanderthal ancestry enriched for proteins that interact with viruses (VIPs) (Guirimand et al., 2015), especially VIPs that interact with RNA viruses, and in particular VIPs associated with HIV-1, IVA, and hepatitis C virus (HCV). The strength of the association was strongest for HIV-associated VIPs (Enard and Petrov, 2018), and as HIV-1 is unlikely to have infected archaic and ancient humans, selection was likely driven by other retroviruses. The VIPs were overrepresented within genes with the biological Gene Ontology function annotations ‘immune effector process’, ‘virion attachment to cell’, and ‘viral genome replication’. These results suggest that the introgressed VIPs conferred a measure of protection against viruses, exogenous and endogenous, that were found in the novel Eurasian environment, or transmitted from Neanderthals to modern humans, and were driven into modern humans by positive directional selection (Enard and Petrov, 2018).

**The combination of archaic HLA variants suggest shared modern native population admixture**

Our analyses of archaic HLA-A, HLA-B, and HLA-C variants demonstrated that these alleles varied between populations and that the selective pattern for each HLA class I gene was distinct. Therefore, we next asked how MNP clustered based on their combination of the archaic HLA allele frequencies. Although we tried to account for more recent migration and admixture events by using MNP, these, in turn, reflect older migration and admixture events that we cannot control for. Nevertheless, adaptation to infectious diseases has throughout space and time favoured selection of the most beneficial HLA variants in a population in a given location. Because of this strong selective pressure, the distance between the populations combined HLA profiles might reflect both shared and distinct ancient selective pressures between populations, and suggest common or different ancestral founder populations. How far back in time these ancestral founder populations existed is not possible to determine.

We investigated the clustering patterns of MNP using a dissimilarity distance matrix of the combined HLA-A, HLA-B, and HLA-C frequencies and non-metric multi-dimensional scaling (NMDS). NMDS clusters populations based on minimising the distances between their dissimilarities and projects this high dimensional space onto a lower dimensional representation. It is similar to classical MDS and principal component analysis (PCA) but is a non-parametric rank-based method that is relatively robust to both data which do not have a known/identifiable distribution and to non-linear relationships between the lower dimensional projected distances and the calculated dissimilarity measure. It is less sensitive to zero-inflation than

PCA and less susceptible to problems such as the arch effect due to its lack of assumed linearity. Like other ordinations, NMDS does not require assumptions about an underlying selective model as the ordinations in the reduced space simply aim at summarizing the HLA allele variability, and can therefore reveal any kind of genetic structuring, including clines (Jombart et al., 2009).

We first performed separate analyses of how the MNP clustered using the twelve Neanderthal and nine Denisovan HLA variants. Using the Neanderthal HLA variants (**Figure 2.7a**), SEA-Taiwanese and OC populations were widely dispersed at one extreme of NMDS-1 (ordination axis created by NMDS akin to a principle component in PCA), with WA, NAF, and relatively tightly clustered EU populations were at the other. Most SA, NEA-Korean, NEA-Tuvan, and some SEA populations, including Chinese populations, were found around the centre. NEA-Japanese populations clustered tightly and exhibited a similar distance from the centre in NMDS-1 as the EU populations. The SEA-Chinese populations were slightly more dispersed than the EU populations despite having ~10% less land area, although the current Chinese population is roughly double that of Europe. OC-Papau New Guinea and SEA-Taiwan were both more dispersed than the EU populations despite those countries being ~22 and ~280 times smaller by land area and ~1300 and ~85 times smaller by population respectively (not counting the Taiwanese Han population). Indeed, the NMDS distances between the most extreme populations in OC were ~3 times that in EU populations and twice that in the non-Taiwanese SEA populations. Therefore, the observed distributions of the points is unlikely to be due solely to the geographic proximity or density of the sample sites or the size of the modern population.

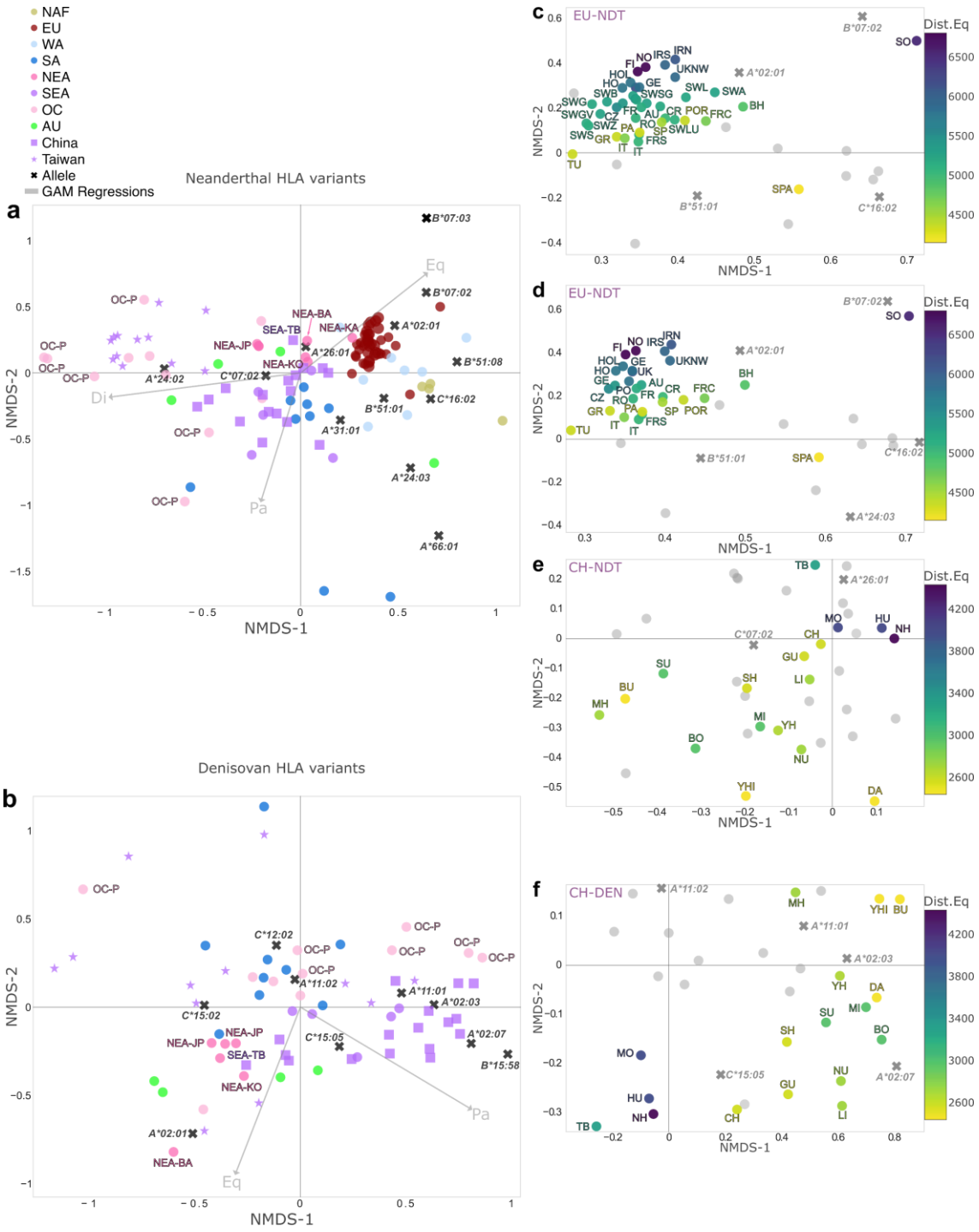



*Figure 2.7* **a**, **b**. NMDS analyses of modern native populations based on their combination of Neanderthal (**a**) or Denisovan (**b**) HLA variants. The plot was coloured according to geographical region. Grey arrows are statistically significant fitted generalised additive models (GAM) of the individual predictors used in the beta regressions. Xs are alleles and indicate how the allele is separating the data points in higher dimensional space.
**c, d**. The EU Neanderthal populations in a, with Swiss populations (**c**) and refit without Swiss populations (**d**). Both coloured according to their distance from the Equator. A North-to-South cline is observed in both cases, indicative of HLA adaption.
**e**, **f**. The SEA-Chinese populations in a and b are coloured according to their distance from the equator. The cline here is less clear than with the EU-NDT populations but there is still a clear division between the northern and southern populations.
Coloured roman text denotes populations. Black/grey italics text denotes alleles. Light grey text denotes the predictors fitted by generalised additive model. (Abbreviations continued on next page).

*Figure 2.7 (cont.)* Geographic regions are colour coded as outlined in the figure. NAF = North Africa, EU = Europe, WA = Western Asia, SA = South Asia, NEA = North East Asia, SEA = South East Asia, OC = Oceania, AU = Australia, NDT = Neanderthal, DEN = Denisovan
**a, b. NEA-BA = Bering Island Aleuts,** NEA-KA = Karelia, NEA-JP = Japan, NEA-KO = Korea, OC-P = Papua New Guinea
**c,d,** AU = Austria, BH = Bosnia and Herzegovina, CR = Croatia, CZ = Czech Republic, FI = Finland, FR = France, FRC = France Corsica, FRS = France South-East, GE = Germany, GR = Greece, HO = Holland, HOL = Holland Leiden, IRN = Ireland North, IRS = Ireland South [sic] (Republic of Ireland), IT = Italy, NO = Norway, PO = Poland, POR = Portugal, PA = Portugal Azores Terceira Island, RO = Romania, SO = Scotland Orkney Islands, SP = Spain, SPA = Spain Andalusia Gypsies [sic] (Roma), SWA = Switzerland Aargau-Solothurn, SWB = Switzerland Basel, Switzerland Bern = SWBE, SWGV = Switzerland Geneva, SWG = Switzerland Graubunden, SWLA = Switaerland Lausanne, SWL = Switzerland Luzern, SWLU = Switzerland Lugano, SWS = Switzerland Sion, SWSG = Switzerland St. Gallen, SWZ = Switzerland Zurich, TU = Turkey, UK = United Kingdom, UKNW = Wales
**e,f.** BO = Bouyie, BU = Bulang, CH = Canton Han, DA = Dai, GU = Guangzhou, HU = Hui, LI = Lisu, MH = Meizhou Han, MI = Miao, MO = Inner Mongolia,YHI = Yunnan Hani, NH = North Han, NU = Nu, SH = South Han, SU = Shui, TB = Tibet, YH = Yunnan Han

In contrast to the Neanderthal NMDS, when using the Denisovan HLA variants, the SEA-Taiwanese populations were more widely dispersed with large overlaps with other non-Taiwanese SEA, NEA, and SA populations (**Figure 2.7b**). The Oceanian non-Papua New Guinea populations grouped together, with the Philippines as an outlier, whereas the Papua New Guinea populations were dispersed across the entire ordination space on NMDS-1. The NEA-Japanese populations were more dispersed than in the Neanderthal NMDS but still clustered relatively tightly. They were closer to the NEA-Korean and NEA-Tuvan populations than in the Neanderthal NMD, while the NEA-Bering Island Aleuts were much further away from the other NEA populations. The core of the SA populations was more broadly dispersed than in the Neanderthal NMDS but there were also fewer extreme outliers.

An NMDS of HLA-B*73:01 and recombinants thereof (**Figure 2.8a**) was broadly similar to the Neanderthal NMDS. However, in the Neanderthal NMDS, the EU,

OC/Taiwanese, and SEA clusters were broadly separated by NMDS-1. In the HLA-B*73:01 recombinant NMDS the EU populations were separated from the OC/Taiwanese populations and the SEA by NMDS-1. The OC/Taiwanese populations were similar to SEA populations on NMDS-1 and only formed a separate cluster from them on NMDS-2. Similar to both the Neanderthal and Denisovan NMDS plots, populations from OC and SEA-TW both clustered and displayed the greatest distance between populations, although there were some extreme outliers.

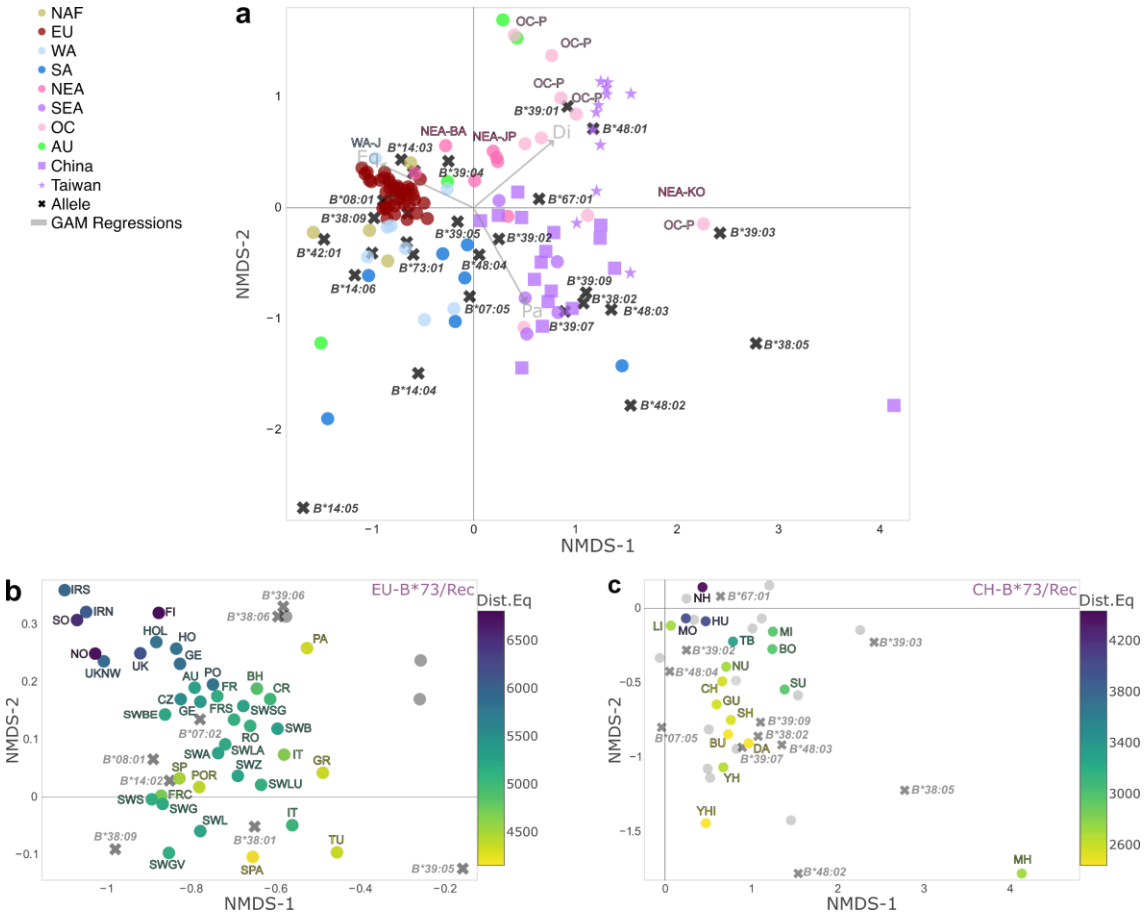



*Figure 2.8* NMDS analyses of the clustering of modern native populations using their combined composition of HLA-B*73:01 and recombinants thereof. **a**. NMDS analyses of modern native populations based on their combination of HLA-B*73:01 and recombinants. The plot was coloured according to geographical region. **b**. The EU populations were coloured according to latitude. A North-to-South cline is observed, indicative of HLA adaption. Populations abbreviated as in Figure 7. **c**. The SEA-Chinese populations were coloured according to latitude. A North-to-South cline is observed, indicative of HLA adaption. Populations abbreviated as in Figure 2.7.

Our results concur with other genetic data (Lazaridis et al., 2014; Raghavan et al., 2014; Friedlaender et al., 2008; Moreno-Mayar et al., 2018; Posth et al., 2018). For example, in the Neanderthal and HLA-B*73:01 analyses, the NEA-Bering Island Aleuts (an indigenous people of North America and the Bering Sea), were placed relatively closely to European populations which might reflect the shared ancestry from an ancestral 'Ancient North Eurasian ghost-population' which no longer exists in an unmixed form. While this population initially was inferred only through statistical reconstruction (Lazaridis et al., 2014), direct evidence was obtained from bones from one ~25,000 years old boy in South-East Siberia (Raghavan et al., 2014). Part of that population migrated west and contributed to Europeans (Moreno-Mayar et al., 2018). Others migrated East across Siberia ~25,000 YBP. They contributed to at least one of the East Asian populations that crossed the Bering Land bridge and gave rise to populations in North and South America. .

Populations from Oceania and Taiwan display a uniquely dispersed pattern. Although major parts of Oceania were settled ~50,000-30,000 YBP, the populations were relatively isolated for ~25,000 years and each had a small effective population size, which, together with distinct selective pressures, were likely to result in a high degree of identity-by-descent within each population. The dispersed distribution pattern of the OC populations, especially the OC-Papua New Guinea, and SEA-Taiwanese populations reflect this and concur with genetic and statistical data demonstrating that the most divergent populations were found on large islands in the Pacific with topographical complexity, while the populations internally were the most homogeneous (Friedlaender et al.,

2008). Likewise, the clustering of OC-Papua New Guinea with some SEA-Taiwanese populations in the Neanderthal and B*73:01 recombinants analyses likely reflects the reported weak Taiwanese genetic signature in Melanesia. The stronger clustering in the Denisovan NMDS between some OC-Papua New Guinea and SEA populations, especially Chinese populations, are in line with the stronger link to groups in East Asia (Friedlaender et al., 2008). Our results also reflect the large genetic distance between Melanesian and Polynesian populations (Friedlaender et al., 2008).

It is important to underscore that our NMDS results only cluster the modern native populations according to the combined archaic HLA allele distances between them. These HLA alleles only comprise a small number of each population's HLA profile. Furthermore, the Neanderthal and Denisovan HLA variants were obtained from individuals in temperate or colder climate zones. Because of this, they likely only partly reflect the HLA variants in all Neanderthals, and probably not those HLA variants that predominated in Denisovans from subtropical and tropical climate zones (**Figure 2.1**). The overall HLA allele variability within each population is likely greater than the HLA variability between them when considering only archaic HLA variants. Nevertheless, it is striking that NMDS-1 and NMDS-2 in all three analyses largely separate MNP according to geographic region. Some of the apparent overlap is due to the unusual categorization found in the AFD, e.g. SEA-Taiwanese indigenous populations are distinctly Oceanian rather than Southeast Asian, and NEA-Karelia is part of European Russia directly adjacent to Finland. Moreover, it is surprising that our results to a large extent suggest shared ancestry between populations that concur with other genetic analyses of the peopling of the Americas

and the genetic variability in Oceania.

**North-to-South clines provide evidence for local adaptation of archaic HLA variants**

Clines (relatively smooth gradients in allele frequencies at selected loci) have long been considered to be indirect evidence for local adaptation to environmental heterogeneity (Futuyma, 2005; Haldane, 1948). We hypothesized that differences in temperature from North to South would favour different combinations of pathogens that in turn would impose distinct selective pressures on the HLA region. We focussed these analyses on Europe and China. Although China by itself is not one of the geographical regions designated by the AFD (González-Galarza et al., 2015), it is appropriate in these analyses to consider this country as such as it covers a large land area that currently can be divided from north to south into six temperature zones: cold-temperate, Qinghai-Tibet Plateau Temperate Zone, temperate, warm-temperate, subtropical, and tropical, as well as a wealth of local climate regions due to differences in precipitation. Likewise, Europe covers a large land area with eight climate regions: subarctic, tundra and highland, humid continental, marine, humid subtropical, Mediterranean, and semiarid. Although these temperature zones do not accurately reflect the fluctuating climate over the last ~650,000 years in Europe and ~450,000 years in China, the different distances from the Equator will result in temperature differences related to latitude. As the regions encompassing Europe and China were inhabited by Neanderthals and Denisovans, respectively, interbreeding with these highly adapted archaic humans likely took place as modern humans ventured into their territory.

Using the latitude information associated with the MNP, we found a North-to-South cline within Europe in the Neanderthal (**Figure 2.7c**) and HLA-B*73:01 HLA NMDS analyses (**Figure 2.8b**). The populations from different regions of Switzerland clustered with many different European populations rather than with each other. The genetic heterogeneity of Switzerland has been observed previously and the genetic similarity with other European populations often does not cluster together with linguistic boundaries (Buhler et al., 2012). To investigate, we removed the Swiss populations and the North-to-South cline was slightly more uniform, but was largely unaffected (**Figure 2.7d**). These results suggest that strong selective pressures acted on the HLA region to select for different combinations of Neanderthal HLA variants, as well as HLA-B*73:01 recombinants, in Europe.

In contrast, when considering the SEA-Chinese populations, we observed only a faint cline in the Neanderthal (**Figure 2.7e**) and Denisovan (**Figure 2.7f**) NMDS, although there was still a clear division between the northern and southern populations. The North-to-South cline was readily apparent in the HLA-B*73:01 recombinants NMDS (**Figure 2.8c**). This suggest that the selective pressure exerted by exogenous pathogens in different regional climates affected the Neanderthal and Denisovan HLA variants less than the HLA-B*73:01-related HLA variants. This result likely reflects that HLA-B variants are more granularly adapted to pathogens in specific regions. The HLA-B*73:01-related dataset was larger and entirely composed of HLA-B variants which are much more affected by pathogen diversity (and therefore climate), while the Neanderthal and Denisovan datasets comprised HLA-A, -B, and -C variants, so this result is to be expected.

Neutral effects can potentially confound the interpretation of these results. Gene flow will homogenise populations (i.e. erase clines) if unopposed by other factors, and random genetic drift on its own will not produce consistent directional genetic changes, although it can produce temporary clines in very small populations (Futuyma, 2005). However, local genetic drift combined with isolation by distance (IBD), i.e. the spatial autocorrelation discussed earlier, can also lead to clines and "arch" or "horseshoe" effects in ordinations (Meirmans, 2012; Novembre and Stephens, 2008). Another possible complicating factor is the potential for linkage disequilibrium, which can influence the frequency of one allele in a population if it is linked to another allele that is more strongly selected for. HLA alleles exhibit linkage disequilibrium (Ahmad et al., 2003) and moreover HLAs work in combination, such as HLA-B*73:01 recombinants interactions with HLA-E, which could potentially cause the frequency of one allele to rise, even if not directly selected for. In order to examine these confounding factors we looked at the amount of each archaic HLA type individually within populations. If the clines observed were due solely to neutral processes such as spatial autocorrelation, we would expect to see uniform clines across all HLA subsets such as the decrease in both neutral genetic diversity and genetic diversity for all HLA types with distance from Africa (Prugnolle et al., 2005b, 2005a). We do not observe this in our results (**Figures 2.9, 2.10**). Clines of neutral or advantageous genes linked to another gene that is being selected for/against have specific characteristics (Barton, 1979; Bürger, 2017) that we also do not observe in our results. As an additional comparison, we investigated the global distribution of the transspecies-polymorphism ABO and Rhesus blood group systems in populations worldwide (**Figure 2.9e**). These blood groups play a role in host immune responsiveness

as, e.g., antibody cross-reaction can be observed between certain pathogens and the A and B glycoprotein antigens and may affect the antibody response to these pathogens, and incorporation of histoblood group structures as part of the viral envelope can affect viral transmission rates between host and recipient (Seymour et al., 2004). Thus, the high within and between population diversity could be a result of natural selection on individuals in specific environments. Unlike with the archaic alleles, we did not observe any clines in the NMDS of blood groups, although the NMDS was able to cluster populations by region.

In combination, these results suggest that different combinations of pathogens thrived across the temperature gradients in Europe and China, and that

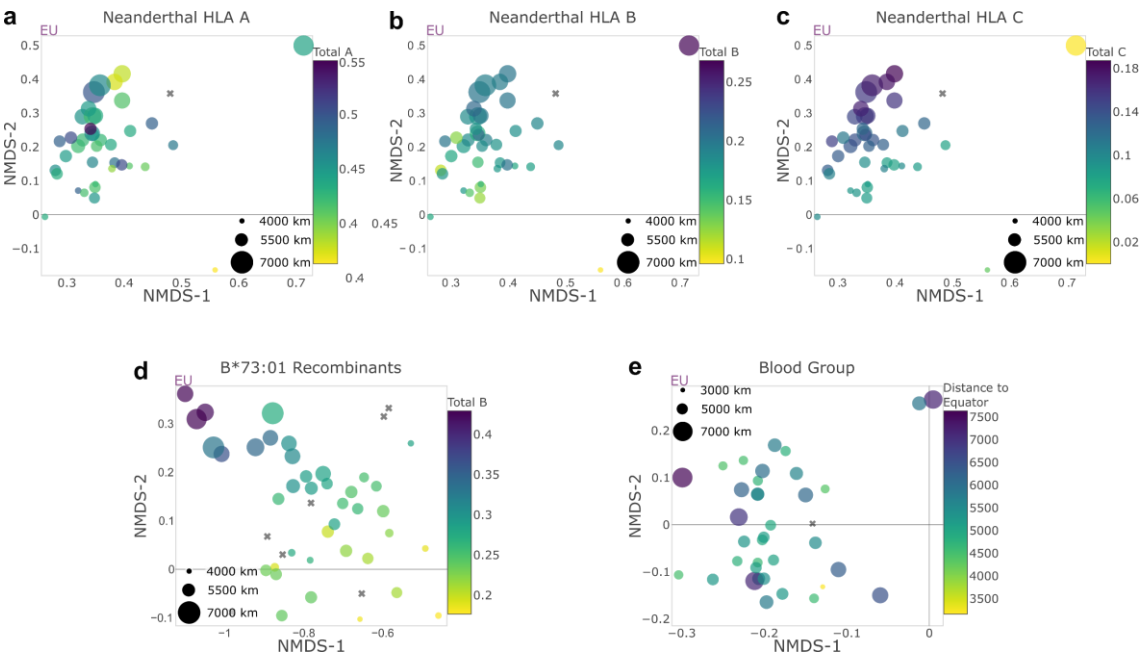



*Figure 2.9* NMDS analyses of modern native populations based on their combination of Neanderthal HLA variants (**a-c**), B*73:01 recombinants HLA variants (**d**) or ABO and Rhesus blood group (**e**). Colour scale for **a-e** is the total amount of the archaic HLA-A, -B, or -C. Colour scale for **e** is distance from the Equator. Point sizes are all scaled by distance from the Equator as shown in the sub-legends (black circles and text) in each image. All plots are focusing on the European cluster of the NMDS. For Neanderthal HLA-A variants (**a**), no cline is present. For Neanderthal HLA-B (**b**) HLA-C (**c**) and the B*73:01 Recombinants HLA variants the cline for the amount of archaic HLA largely tracks the cline for distance from the Equator. For the ABO and Rhesus blood group NMDS (**e**), there is no cline for distance from the Equator.

adaptation to regional pathogens selected for specific HLA combinations. Whether this adaptation took place in Neanderthals in Europe and Denisovans in China and was introgressed during sequential interbreeding between groups of local archaic humans and modern humans, or whether this occurred later only in modern humans, is unknown.

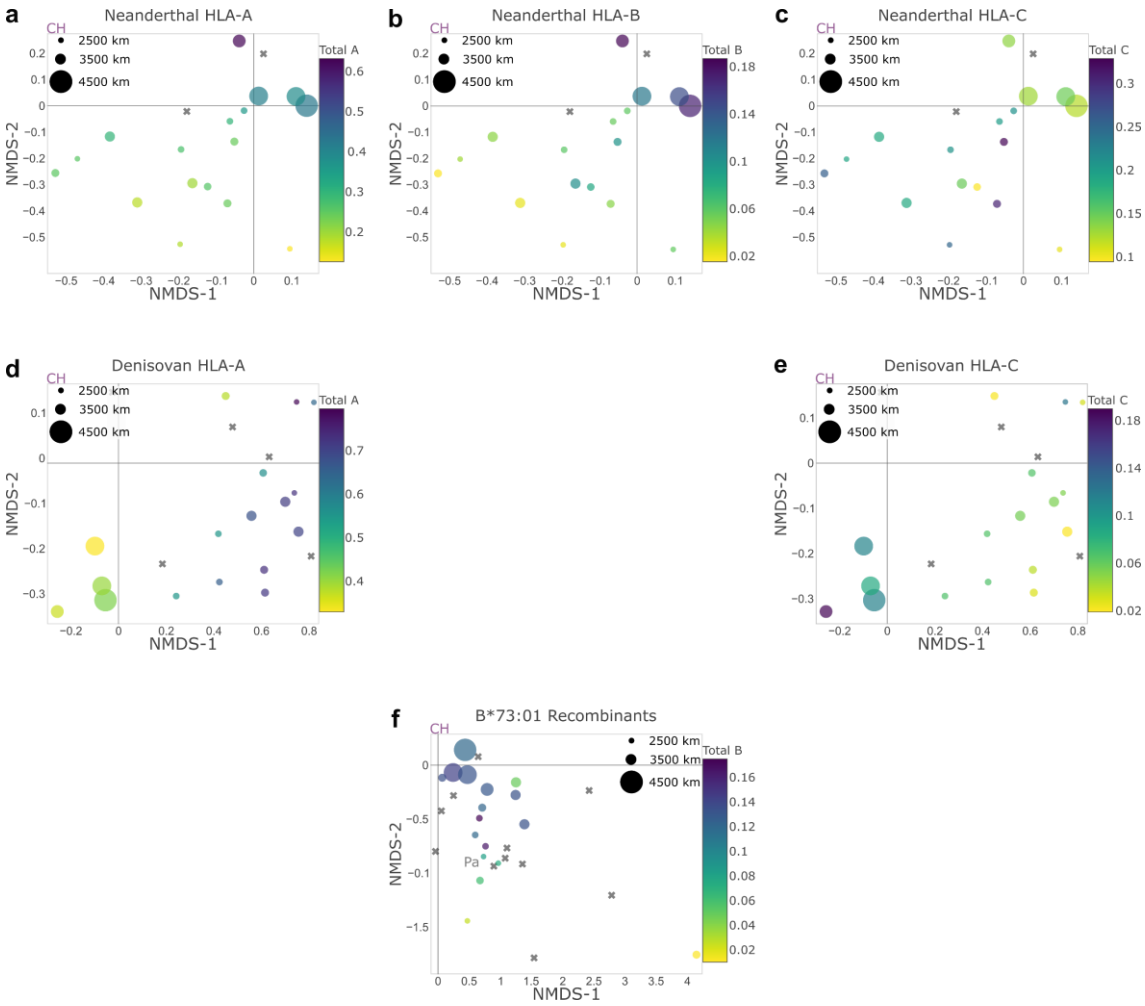



*Figure 2.10* NMDS analyses of modern native populations based on their combination of Neanderthal (**a-c**), Denisovan (**d,e**) or B*73:01 recombinants HLA variants (**f**). Points are coloured according to the total amount archaic HLA-A, -B, or -C. Point sizes are all scaled by distance from the Equator as shown in the sub-legends (black circles and text) in each image. All plots are focusing on the Chinese cluster of the NMDS. For Neanderthal HLA-A (**a**) and HLA-B (**b**) variants a faint cline mirrors the cline for distance from the Equator. For Neanderthal HLA-C (**c**), no cline is present. For Denisovan HLA-A (**d**) a cline runs in the opposite direction to that of distance from the Equator. For Denisovan HLA-C (**e**) a cline is present which tracks distance from the Equator. For HLA B*73:01 Recombinants (**f**) a cline is present which tracks distance from the Equator.

To summarise, we observed a robust significant relationship between a population's archaic HLA variant frequencies and their geographical location, even when accounting for neutral factors such as spatial autocorrelation, despite these HLA variants comprising less than 0.1% of all known HLA variants. This suggests these archaic HLA variants are especially beneficial for survival in Eurasia, North Africa, and Oceania, regions previously inhabited by archaic humans for ~650,000 to ~450,000 years. The cause of this population level adaptation to local regions worldwide might in part be due to the combination of exogenous pathogens and the HERV-K insertion pattern in each population.

## DISCUSSION

Here we investigated archaic HLA-A, HLA-B, and HLA-C allele frequencies in modern human native populations. We discovered that at least one of these HLA alleles are found in more than 50% of all Eurasians alive today, ~60,000 years after our common ancestors left Africa, despite comprising < 0.1% of known HLA alleles (Gonzalez-Galarza et al., 2015). Moreover, our analyses revealed that local selective pressures acted on HLA-A and HLA-B and that distinct selective pressures were related to the geographical location of the examined populations. Additional selective pressures appeared to stabilise the frequencies of HLA-A and HLA-C. HLA-C is known to evolve more slowly than HLA-A and HLA-B because of the interplay between HLA-C ligands and NK cells during placentation and the need for successful reproduction to sustain a population. Selective pressures that act on HLA-A, but not HLA-B have to the best of our knowledge not been described previously. Here we hypothesized that HLA-A, in addition to presenting epitopes from exogenous pathogens, in some way act to ensure the long-term survival of the individual, and that one possibility could be that HLA-A presentation is important for controlling endogenous retrovirus expression. While low-level HERV activation can be beneficial as it increases immune reactivity to unrelated poorly immunogenic targets, extensive activation can be damaging and needs to be controlled by the immune system (Grandi and Tramontano, 2018; Jern and Coffin, 2008; Kassiotis and Stoye, 2016; Zeng et al., 2014). This proposition does not rule out that additional selective pressures could act on HLA-A, but not HLA-B. Exploratory bioinformatics analyses indicate that the amino acid composition of epitopes presented by HLA-A is more varied than in epitopes presented by HLA-B and HLA-C, but these analyses did not include the composition of HERV epitopes

(Pierini and Lenz, 2018). The selective pressure imposed by the need to control HERV expression would be shared across all populations, explaining the high frequency of archaic HLA-A variants across populations. As the HLA-A frequencies are several times higher than those of HLA-C, the proposed selective pressure imposed by HERVs to avoid cancer and premature death of the individual appear to be stronger than that imposed on HLA-C by the need for successful reproduction.

We found that the NMDS clustering of modern native populations based on their combination of archaic HLA variants were in line with their geographic location and results from genetic studies regarding their shared ancestry. It is unknown to what extent this reflects sequential interbreeding with highly adapted archaic humans, interbreeding and subsequent adaptation in modern humans, or a combination of both. Studies have demonstrated that HLA adaptation can occur relatively fast, particularly with regards to HLA-B, as, e.g. new HLA-B alleles are a characteristic of Amerindian populations in South America, Central America, and the southwestern part of North America (Parham et al., 1997) and have evolved over the -15,000 years as modern humans ventured into and spread throughout the Americas (Skoglund and Reich, 2016).

This result suggests that these archaic HLA variants are still beneficial to the survival of the populations and therefore that some of the pathogen-associated selective pressures in the different environments are ancient. While gradual climate transformations allow for gradual adaptation of resident populations to novel pathogens, the current rapid climate changes will not. Old pathogens, e.g.,

dengue virus, malaria and other mosquito-borne viruses, are currently spreading into new areas (McMichael et al., 2003). The outcome of infection by old pathogens that spread in populations with HLA variants maladapted to combat them might be different from that in populations where these pathogens are common. For example, before vaccination against measles was possible and antibiotics available against secondary bacterial infections, the mortality rate was relatively low in Western populations, whereas it could be >40% in isolated populations without prior exposure to the virus (Shanks et al., 2014).

Our results suggest that although humans are able to survive in extremely different habitats across the globe, we also carry an inherent ability to combat different combinations of pathogens. As old pathogens spread to new regions, populations might be more or less maladapted to combat the new infectious challenges. Our results suggest that further studies might be able to predict to what extent populations have similar or different immune responses to pathogens they have not previously encountered.

**Acknowledgements**

We are forever indebted to our co-author, mentor, and dear friend Dr Daniel Lunn, who sadly passed away prior to the acceptance of this publication. We thank the sampled populations and the researchers whose publicly available data made this work possible. We are grateful to Dr Ben Lambert and Professor Gill McVean for helpful discussions.

## METHODS

### Creation of the modern native population dataset

To estimate the impact of admixture with archaic hominins on modern native human populations we downloaded all 123,626 listed allele frequencies from the Allele Frequency Database (AFD) (Gonzalez-Galarza et al., 2015), and consolidated the frequencies on a per population basis. Any alleles over four digits were trimmed down to four digits and combined, e.g., "Norway BMDR" A*01:01:01 and A*01:01:02 both get trimmed to HLA-A*01:01 and their frequencies are added together. 2 digit alleles (e.g., HLA-A*01) were discarded. Manual correction of location coordinates was carried out as required based on the database information about the sampled region or the papers describing the population (e.g., changing latitude and longitude to place "Germany Pop 8" in Ulm, Germany rather than Xudun, Somalia). Some populations without submitted coordinates had obvious locations based on the name: e.g., Switzerland Bern. For less precisely named populations (e.g., Japan pop 17) we used the geographic centre of population for the named region when available, otherwise we used the geographic centroid for the country. Populations without a recoverable location were discarded. Populations were then restricted to "modern native populations" which excluded modern colonial populations (e.g., people of European descent in the Americas and Australia), as well as inherently mixed populations (e.g., Mestizo populations). Other systemically biased populations such as "China Han HIV negative" were also removed. Our final database of modern native populations contained 150 worldwide populations comprising a total of 204,748 people (see **Table S2.4** for full list). We used this database to

extract HLA variants found in Neanderthal, Denisovan, and/or HLA-B*73:01 and recombinants thereof.

### Creation of the pathogen richness dataset

The Epidemiology module of the Global Infectious Diseases and Epidemiology Network tracks the global and country-specific status of 337 generic infectious diseases in 220 countries compiled from health ministries, military agencies, dedicated Internet lists, peer-review journals, standard texts, and data presented at major conferences (Yu and Edberg, 2005). As no standardised prevalence data for all pathogens exist, we created a presence absence/matrix of endemic pathogens on a per country basis from these data (Retrieved from www.gideononline.com on 05-06-2020). Pathogens that were present in all countries examined were removed, as were smallpox and SARS-CoV-1 as they were listed as absent in all countries. Pandemic SARS-CoV-2, colloquially known as covid-19, was not present in all countries at the time of database access but was manually removed due to the expectation that it soon would be (this prediction has unfortunately since been borne out). Because the HLA-A, -B, and -C, are primarily involved with the presentation of peptides derived from intracellular pathogens we filtered the presence/absence matrix to viruses and intracellular bacteria (both facultative and obligate) only.

### HLA typing of a high-resolution Neanderthal genome from the Altai mountains

We performed HLA typing of the high-resolution genome sequences from a ~120,000 years old (Slon et al., 2018) Neanderthal woman from the Denisova Cave, Altai Mountains (Prufer et al., 2014) using Optitype, an HLA genotype

algorithm based on integer linear programming (Szolek et al., 2014).

### Regression Analysis

Linear regressions were carried out using the R package "stats" version 3.6.1 (R Core Team, 2018). Beta regressions were carried out using the R package "betareg" version 3.1-4 (Cribari-Neto and Zeileis, 2010).

We initially performed linear regression analysis of all the comparisons in this study such as those comparing the various distance metrics vs. proportions of archaic alleles in populations (**Figure 2.11**). Although these linear models showed significant correlations in some cases, e.g. **Figure 2.11b**, these results are unreliable for a number of reasons. When examining the results of these models they often predicted values less than 0 for the dependent variable. As the dependent variable is a proportion, this is a nonsensical result. Formal diagnostics of the linear models also gave reason to reject them. The linear models were tested for heteroskedasticity using the Breusch-Pagan test and for

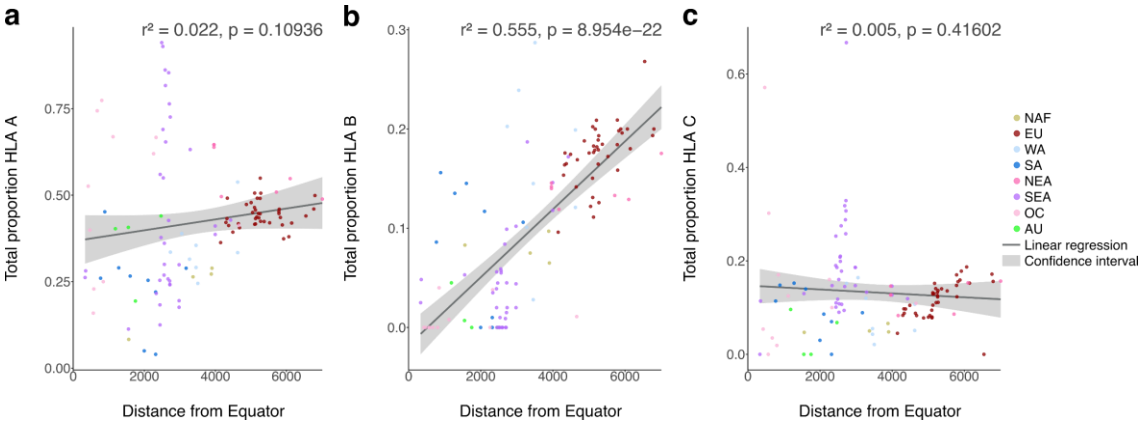



*Figure 2.11* A representative example of initial linear regressions carried out. These are linear regressions comparing distance from the equator vs. the total proportion of Neanderthal HLAs in populations. Note that in **b** both the linear regression and confidence interval go below 0. Geographic regions are colour coded as outlined in the figure. NAF = North Africa, EU = Europe, WA = Western Asia, SA = South Asia, NEA = North East Asia, SEA = South East Asia, OC = Oceania, AU = Australia.

normality using the Shapiro-Wilk test, and in general failed both of these tests. For example, conducting the Breusch-Pagan test on the linear model used in **Figure 2.11** a gave a p-value of 2.56e-05, thus rejecting the null hypothesis of homoscedasticity. Similarly, conducting the Shapiro-Wilk test gave a p-value 1.55e-05, rejecting the null hypothesis of normality. Linear regression analysis assumes both normality and homoscedasticity.

These are known problems with linear regressions of proportional data. The dependent variables used in these analyses are proportions bounded by zero and one ("the unit interval"), which runs contrary to the assumption of a theoretically unbounded continuous dependent variable assumed by linear regression (Douma and Weedon, 2019). Bounded proportion data tend to display more variation around the mean and less towards the lower and upper limits of the unit interval, making them naturally heteroskedastic (Kieschnick and McCullough, 2003). The distributions of proportions are also typically asymmetric, confounding Gaussian-based (normal) inference and hypothesis testing (Cribari-Neto and Zeileis, 2010; Ferrari and Cribari-Neto, 2004). This can, for example, cause linear regression analyses to predict impossible estimates outside the unit interval (e.g. **Figure 2.11b**).

Ferrari and Cribari-Neto (2004) proposed the beta regression model for dealing with data that are continuous values in the standard unit interval such as those found with proportions. The model is naturally heteroskedastic and can accommodate asymmetries (Cribari-Neto and Zeileis, 2010). For a full treatment of the mathematics behind the beta regression model see Ferrari and Cribari-

Neto (2004). Briefly, the beta density is parameterised in terms of both the mean μ and the precision parameter φ (accounting for the precision of the data), and based on varying combinations of these parameter values, the density can assume a variety of shapes, making the model very flexible (Cribari-Neto and Zeileis, 2010). The linear predictor is linked to the mean response through a link function, and when the logit link is used it allows interpretation of the regression parameters as an odds ratio (Ferrari and Cribari-Neto, 2004). Simas, Barreto-Souza, and Rocha (2010) developed a more general variant of the beta regression model called the variable dispersion beta regression model, in which the φ parameter is allowed to vary rather than being assumed to be constant. This variability of φ better accounts for heteroskedasticity.

As the beta regression model is tailored to deal with a dependent variable on the unit interval, we used this method. A drawback of the beta regression model is that it operates on the open interval (0,1), meaning that it excludes the boundary points of 0 and 1, which the data we use here include. There are several ways of addressing this problem. A number of extensions of the beta regression model variably titled "zero-inflated beta regression", "zero-augmented beta regression", and "zero-and-one augmented beta regression" have been developed which allow for the analysis of data containing zeros and ones (Douma and Weedon, 2019; Liu and Kong, 2015; Ospina and Ferrari, 2012; Wright et al., 2017). Underlying these extensions however is the assumption that the data generating mechanism involves two separate but linked stochastic processes, a Bernoulli process for generating non-zero, non-one values and then a subsequent beta distribution for the actual proportions of the non-zero non-one values. The other

option is rescaling the dependent variable from the closed interval [0,1] the to the open interval (0,1) via a variety of different methods. Cribari-Neto and Zeileis (2010), the authors of the R package "betareg" we utilised in this study, recommend the transformation described by Smithson and Verkuilen (2006):

$$y' = \frac{y * (n-1) + 0.5}{n} \qquad \text{Equation 2.1}$$

where $y$ is the dependent variable and $n$ is the sample size. As there is no reason to believe the assumptions required by "zero-and-one augmented beta regression" are valid for this study, we transformed the data as recommended via Equation 2.1.

Beta regression models with explanatory variables and either equidispersion (constant φ) and variable dispersion (φ is allowed to vary) were examined and compared to the null model (no explanatory variable) and each other. Likelihood ratio tests and the Akaike information criterion (AIC) showed that both types of model with explanatory variables were preferable to the null model, and that the models with variable dispersion were generally preferable to those with equidispersion. All mean link functions available in the betareg package were compared, namely "logit", "probit", "cloglog", "cauchit", "log", and "loglog". The logit link function was used as the best compromise between good AIC and interpretability of the results.

Beta regression models do not have a normal $R^2$ value. Instead, Ferrari and Cribari-Neto (2004) suggested a "pseudo $R^2$" ($R_p^2$) as an equivalent metric which is the squared correlation between the linear predictor for the mean and the link-transformed response. *Nota bene* that the term "pseudo $R^2$" has multiple distinct

meanings in statistics e.g. "McFadden pseudo $R^2$" or "Cox and Snell pseudo $R^2$" which are not directly comparable to the $R^2$ of ordinary least square models and cannot be interpreted as the proportion of the variability in the dependent variable that is explained by model. In contrast to these, the pseudo $R^2$ used here is a global measure of explained variation.

Similarly, the betareg library has no infrastructure equivalent to base R's predict method for determining confidence intervals in standard linear models. That is, there is nothing analogous to `predict.lm(..., se.fit = TRUE)` for beta regressions. This function would correspond to the residual variance of the beta regression, which is assumed to be constant in a linear model (Zeileis, 2016). The variance in a beta regression pertains to the predicted variance of the response rather than variance of the predicted mean. Zeileis (2016) suggests two alternative methods for equivalents to confidence intervals for standard linear models: either using the R package lsmeans (now emmeans) which computes estimated marginal means (also called least-squares means) or else using the bootstrap (Davison and Hinkley, 1997). We show the results of both these methods which are broadly in agreement with each other. We used the R packages "emmeans" version 1.5.4 (Lenth, 2021) as implemented in "ggeffects" version 1.0.1 (Lüdecke, 2018) to compute confidence intervals from estimated marginal means. We used the R package "boot" version 1.3-25 (Canty and Ripley, 2020) to carry out ordinary nonparametric bootstraps with 10,000 replicates for each model from which upper (0.975) and lower (0.025) quantiles were then calculated. Finally, we showed the 2.5% and 97.5% quantiles of the predicted beta distribution as a means of visualising the variance model's

distribution.

The R package "car", version 3.0-10 (Fox and Weisberg, 2019) was used to compute the variance inflation factors for the multiple regression predictors in order to ascertain the effects of multicollinearity.

**Spatial Autocorrelation**

The Moran's I Autocorrelation Indices of each dataset were calculated using the `Moran.I()` function of the R package "ape" version 5.4-1 (Paradis et al., 2004). The weighting matrix for the `Moran.I()` function was constructed using the geodist() function of the R package "geodist" version version 0.0.7 (Padgham and Sumner, 2021) using the "geodesic" measure for distance which takes into account that the Earth is not a perfect sphere when measuring between coordinates. Results are significantly autocorrelated if their value of Moran's I Autocorrelation Index was significantly ($p < 0.05$) different from the expected value under the null hypothesis of no autocorrelation.

**Distance Based Moran's Eigenvector Maps**

dbMEMs (Borcard et al., 2004; Borcard and Legendre, 2002; Dray et al., 2006) were generated using the `dbmem()` function of the R package "adespatial" version 0.3-13 (Dray et al., 2021). For technical details see Legendre and Legendre (2012). Briefly, pairwise geodesic distances between the coordinates were computed as before. Unlike a normal Principal Coordinates of Neighbour analysis from which dbMEM is derived, a threshold value is calculated as four times the maximum edge length in the minimum spanning tree connecting all the points. The

diagonal value of the distance matrix is set to this threshold indicating sites are not connected to themselves. The truncated distance matrix is then submitted to Principle Coordinate Analysis producing orthogonal spatial eigenfunctions which are ordered according to the amount of variation they capture and can be used in regression models as regular explanatory variables.

**Non-metric Multidimensional Scaling**

NMDS was carried out using the `metaMDS()` function of the R package "vegan" version 2.5-7 (Oksanen et al., 2020). A matrix of the individual allele frequencies per population was used as the input data. Stress is the measure of the degree to which distance between samples in the lower dimensional space corresponds to the actual higher dimensional distance. Stress was determined for *k* (number of dimensions) 1 to 10 by getting average stress of 20 replicates for value of *k* and was at acceptable levels for $k > 1$, decreasing as the number of dimensions went up. $k = 2$ was selected for ease of representation. A minimum of 1000 random starts was used in search of a stable solution. Environmental variables (e.g. pathogen richness or distance from the Equator) were fit onto the NMDS ordinations using vegan's `envirofit()` function. `envirofit()` uses a generalised additive model (GAM) to fit each environmental variable separately, with the environmental variable as the dependent variable explained by the ordination scores.

**Images**

Plots were created using the R package "plotly" version 4.9.3 (Sievert, 2020) and exported as interactive html files using the R package "htmlwidgets" version

1.5.3 (Vaidyanathan et al., 2020). Figures were then constructed in Inkscape version 1.0.1 (Inkscape Project, 2017).

# Supplementary Tables

## Table S2.1. Functional features of the Neanderthal and Denisovan HLA variants

| N or D | HLA | Signal peptide | HLA E | KIR ligand | NK cell | Additional notes |
|---|---|---|---|---|---|---|
| **N and D** | A*02:01 | -21M | Refold | - | CD94:NKGA | Posttranscriptional regulation by INF-γ |
| **N** | A*24:02 | -21M | Refold | Bw4 | KIR3DL1<br>CD94:NKGA | High surface expression |
| **N** | A*24:03 | -21M | Refold | Bw4 | KIR3DL1<br>CD94:NKGA | |
| **N** | A*26:01 | -21M | Refold | - | CD94:NKGA | Medium to high surface expression |
| **N** | A*31:01 | -21M | Refold | - | CD94:NKGA | |
| **N** | A*66:01 | -21M | Refold | - | CD94:NKGA | |
| **N** (B*73) | B*07:02 | -21M | Refold | - | CD94:NKGA | Posttranscriptional regulation by INF-γ<br>LD with HLA C*07:02 in Europe<br>LD with HLA C*15:05* in Asia<br>-21M because of gene conversion with HLA B*73 |
| **N** | B*07:03 | -21M | Refold | - | CD94:NKGA | Posttranscriptional regulation by INF-γ ?<br>-21M because of gene conversion with HLA B*73?<br>Possibly not a real allele but the result of sequencing errors? |
| **N** | B*51:01 | -21T | - | Bw4 | KIR3DL1 | |
| **N** | B*51:08 | -21T | - | Bw4 | KIR3DL1 | |
| **N** | C*07:02 | -21A | Refold | C1 | KIR2DS2<br>KIR2DL2/3 | Low surface expression<br>LD with B*07:02 or B*08:01 in Europe |
| **N** | C*16:02 | -21M | Refold | C2 | KIR2DL1<br>KIR2DS1**<br>CD94:NKGA | Medium surface expression |
| **D** | A*02:03 | -21M | Refold | - | CD94:NKGA | |
| **D** | A*02:07 | -21M | Refold | - | CD94:NKGA | |
| **D** | A*11:01 | -21M | Refold | A3/A11 motif | KIR3DL2<br>CD94:NKGA | |
| **D** | A*11:02 | -21M | Refold | A3/A11 motif | KIR3DL2<br>CD94:NKGA | |
| **D** | B*15:58 | -21T | - | NA | | Extremely rare (nearest modern HLA is B*46, a result of B*15:01 and C*01:02 mini conversion |
| **D** | C*12:02 | -21M | Refold | C1 | KIR2DS2<br>KIR2DL2/3<br>CD94:NKGA | Medium to high surface expression |
| **D** | C*15:02 | -21M | Refold | C2 | KIR2DL1<br>KIR2DS1**<br>CD94:NKGA | |
| **D** | C*15:05 | -21M | Refold | C2 | KIR2DL1<br>KIR2DS1**<br>CD94:NKGA | |

N signify a Neanderthal HLA variant, D a Denisovan HLA variant
* Very rare combination in Europeans and relatively rare in Asians, because of a bias towards coupling of HLA B-Bw6 motif and HLA C-C1 epitope. This bias is not found in Africans.

** Suggested to be an adaptive introgression from archaic humans in Eurasia

**Table S2.2 HLA-B*73 and recombinants thereof all carry -21M in the signal peptide:**

Table S2.4 HLA-B*73 and recombinants thereof. Italics indicates allele is also found in Neanderthals. Bold highlights HLA B*73:01.

| | | | |
|---|---|---|---|
| HLA-B*07:02 | HLA-B*07:05 | HLA-B*08:01 | HLA-B*14:01 |
| HLA-B*14:02 | HLA-B*14:03 | HLA-B*14:04 | HLA-B*14:05 |
| HLA-B*14:06 | HLA-B*38:01 | HLA-B*38:02 | HLA-B*38:05 |
| HLA-B*38:06 | HLA-B*38:09 | HLA-B*39:01 | HLA-B*39:02 |
| HLA-B*39:03 | HLA-B*39:04 | HLA-B*39:05 | HLA-B*39:06 |
| HLA-B*39:07 | HLA-B*39:08 | HLA-B*39:09 | HLA-B*42:01 |
| HLA-B*42:02 | HLA-B*48:01 | HLA-B*48:02 | HLA-B*48:03 |
| HLA-B*48:04 | HLA-B*48:07 | HLA-B*67:01 | **HLA-B*73:01** |

**Table S2.3: HLA restriction of cytotoxic T cell response to HERV**

| HLA-restricted HERV-specific responses in cancer | HLA-A | HLA-B | HLA-C | Notes |
|---|---|---|---|---|
| HERV-E Envelope | HLA-A*02:01 | | | Clear cell Kidney Cancer (Cherkasova et al., 2016) |
| HERV-E 4 peptides | HLA-A*11 | | | Renal cell carcinoma (Takahashi et al., 2008) |
| HERV-H<br><br>10 peptides from envelope | HLA-A*02:01 | | | On Xp22.3 Gastrointestinal cancers (Mullins and Linnebacher, 2012) |
| HERV-H | H LA-A*02:01 (2 cell lines with HLA info, two without ) | | | Ovarian cancer (Rycaj et al., 2015) |
| HERV-K-MEL | HLA-A*02:01 | | | Melanoma (Schiavetti et al., 2002) |
| HERV-K | HLA-A*02:01 | | | Breast cancer (Wang-Johanning et al., 2008) |
| HERV-K | HLA-A*02:01 | | | Seminoma (Rakoff-Nahoum et al., 2006) |
| HERV-W Envelope | HLA-A*02:01 | | | Potential autoimmune response (Tu et al., 2017) |
| HERV-K | HLA-A*02:01 | | | HIV-infected children (vertically infected)(Tandon et al., 2011) |

| HLA-restricted HERV-specific responses in HIV-1 infection | HLA-A | HLA-B | HLA-C | Additional information |
|---|---|---|---|---|
| HERV-K (HML2) | HLA-A*03:01 | (HLA-B51, the response is cross-reactive with HIV-1 and SIV infected target cells) | | HIV-infection (Jones et al., 2012) |
| HERV-K | HLA-A*02:01 | | | HIV-1 infection (Garrison et al., 2007) |

## Table S2.4: Modern Native Populations from the Allele Frequency Database

Table S2.4 List of populations from the AFD used in these analyses

| | | | | |
|---|---|---|---|---|
| Australia Cape York Peninsula Aborigine | Australia Groote Eylandt Aborigine | Australia Kimberly Aborigine | Australia Yuendumu Aborigine | Brazil Terena |
| Brazil Vale do Ribeira Quilombos | Cameroon Baka Pygmy | Cameroon Bakola Pygmy | Cameroon Bamileke | Cameroon Beti |
| Cameroon Sawa | Central African Republic Mbenzele Pygmy | Chile Easter Island | China Canton Han | China Guangdong Province Meizhou Han |
| China Guangzhou | China Guizhou Province Bouyei | China Guizhou Province Miao pop 2 | China Guizhou Province Shui | China Inner Mongolia Region |
| China North Han | China Qinghai Province Hui | China South Han | China Southwest Dai | China Tibet Region Tibetan |
| China Yunnan Bulang | China Yunnan Hani | China Yunnan Province Han | China Yunnan Province Lisu | China Yunnan Province Nu |
| Colombia North Wiwa El Encanto | Columbia North Chimila Amerindians | Czech Republic NMDR | England North West | Finland |
| France Corsica Island | France Southeast | Georgia Tibilisi | Georgia Tibilisi Kurd | Germany DKMS - Austria minority |
| Germany DKMS - Bosnia and Herzegovina minority | Germany DKMS - Croatia minority | Germany DKMS - France minority | Germany DKMS - Greece minority | Germany DKMS - Italy minority |
| Germany DKMS - Netherlands minority | Germany DKMS - Portugal minority | Germany DKMS - Romania minority | Germany DKMS - Spain minority | Germany DKMS - Turkey minority |
| Germany DKMS - United Kingdom minority | Germany pop 6 | Germany pop 8 | Ghana Ga-Adangbe | Hong Kong Chinese BMDR |
| India Andhra Pradesh Telugu Speaking | India Khandesh Region Pawra | India Mumbai Maratha | India New Delhi | India Tamil Nadu |
| India Tamil Nadu Nadar | India West Bhil | India West Coast Parsi | Iran Baloch | Ireland Northern |
| Ireland South | Israel Arab Druze | Israel Ashkenazi and Non Ashkenazi Jews | Italy pop 5 | Japan Central |
| Japan pop 16 | Japan pop 3 | Jordan Amman | Kenya | Kenya Luo |
| Kenya Nandi | Malaysia Peninsular Malay | Mali Bandiagara | Mexico Chihuahua Tarahumara | Mexico Oaxaca Mixe |

| | | | | |
|---|---|---|---|---|
| Mexico Oaxaca Mixtec | Mexico Oaxaca Zapotec | Morocco Atlantic Coast Chaouya | Morocco Nador Metalsa pop 2 | Netherlands Leiden |
| New Caledonia | New Zealand Maori with Full Ancestry | New Zealand Polynesians with Full Ancestry | Norway BMDR PMID 29684411 | Papua New Guinea East New Britain Rabaul |
| Papua New Guinea Eastern Highlands Goroka Asaro | Papua New Guinea Karimui Plateau Pawaia | Papua New Guinea Madang | Papua New Guinea Wanigela Keapara | Papua New Guinea West Schrader Ranges Haruai |
| Papua New Guinea Wosera Abelam | Philippines Ivatan | Poland DKMS | Portugal Azores Terceira Island | Russia Bering Island Aleut |
| Russia Karelia | Russia Tuva pop 2 | Saudi Arabia Guraiat and Hail | Saudi Arabia pop 5 | Scotland Orkney |
| Senegal Niokholo Mandenka | Singapore Javaneses | Singapore Riau Malay | South Africa Black | South Africa Harvard Zulu |
| South Korea pop 3 | Spain Andalusia Gipsy | Sri Lanka Colombo | Switzerland Aargau-Solothurn | Switzerland Basel |
| Switzerland Bern | Switzerland Geneva pop 2 | Switzerland Graubunden | Switzerland Lausanne | Switzerland Lugano |
| Switzerland Luzern | Switzerland Sion | Switzerland St Gallen | Switzerland Zurich | Taiwan Ami |
| Taiwan Atayal | Taiwan Bunun | Taiwan Paiwan | Taiwan Pazeh | Taiwan Puyuma |
| Taiwan Rukai | Taiwan Saisiat | Taiwan Siraya | Taiwan Tao | Taiwan Taroko |
| Taiwan Tsou | Thailand | Tunisia | Uganda Kampala | Uganda Kampala pop 2 |
| USA Alaska Yupik | USA Arizona Gila River Amerindian | USA Hawaii Okinawa | USA NMDP North American Indian | USA South Dakota Lakota Sioux |
| Venezuela Perja Mountain Bari | Venezuela Sierra de Perija Yucpa | Vietnam Hanoi Kinh pop 2 | Zambia Lusaka | Zimbabwe Harare Shona |

## Table S2.5: φ p-values for Simple (Single Predictor) Beta Regression Analyses

*Table S2.5* Pseudo R squared and φ p-values for beta regression analyses of the relationship between Archaic HLA frequency and distance from the Equator

| | HLA A | HLA B | HLA C |
|---|---|---|---|
| **Neanderthal HLAs** | | | |
| Dist.Equator | < 2e-16*** | 1.70e-11*** | 1.30e-05*** |
| **Denisovan HLAs** | | | |
| Dist.Equator | .23$^{NS}$ | | 0.09$^{NS}$ |
| **B*73:01 Recombinants** | | | |
| Dist.Equator | | 2.21e-08*** | |

Dist.Manot Cave is distance from Manot Cave. Dist.Equator is the distance from the Equator (i.e. a measure of absolute latitude). Pathogens is species richness of intracellular pathogens, both bacteria and viruses, in a country from which population was drawn. Bacteria and Viruses denote their respective individual species richness. $R_p^2$ is the pseudo R squared value, a global measure of variation explained by the model for beta regressions, akin to $R^2$ for linear models. p-values for asymptotic Wald tests: *** < 0.001; ** < 0.01; * < 0.05; NS, non-significant.

## Table S2.6: φ p-values for Multiple Beta Regression Analyses

*Table S2.5* Pseudo R squared and φ p-values for beta regression analyses of the relationship between Archaic HLA frequency, distance from Manot cave, distance from the Equator, and pathogen species richness

| | HLA A | HLA B | HLA C |
|---|---|---|---|
| **Neanderthal HLAs** | | | |
| $R_p^2$ | 16% | 49% | <1% |
| Pathogens | 0.04* | 8.75e-04*** | $0.12^{NS}$ |
| Dist.Equator | $0.22^{NS}$ | $0.33^{NS}$ | 5.27e-06*** |
| Dist.Manot Cave | < 2e-16*** | < 2e-16*** | 0.05* |
| $R_p^2$ | 17% | 53% | <1% |
| Bacteria | $0.18^{NS}$ | 0.02* | $0.06^{NS}$ |
| Dist.Equator | $0.12^{NS}$ | $0.42^{NS}$ | 5.62e-07*** |
| Dist.Manot Cave | < 2e-16*** | 5.43e-15*** | $0.08^{NS}$ |
| $R_p^2$ | 15% | 49% | <1% |
| Viruses | 0.01* | 4.84e-04 | $0.24^{NS}$ |
| Dist.Equator | $0.46^{NS}$ | $0.48^{NS}$ | 8.11e-06*** |
| Dist.Manot Cave | < 2e-16*** | < 2e-16*** | $0.06^{NS}$ |
| **Denisovan HLAs** | | | |
| $R_p^2$ | 12% | | 4% |
| Pathogens | $0.43^{NS}$ | | $0.42^{NS}$ |
| Dist.Equator | 0.03* | | $0.79^{NS}$ |
| Dist.Manot Cave | 1.09e-04*** | | $0.10^{NS}$ |
| $R_p^2$ | 18% | | 2% |
| Bacteria | $0.9^{NS}$ | | $0.45^{NS}$ |
| Dist.Equator | $0.11^{NS}$ | | $0.92^{NS}$ |
| Dist.Manot Cave | 4.5e-05*** | | $0.05^{NS}$ |
| $R_p^2$ | 7% | | 6% |
| Viruses | $0.31^{NS}$ | | $0.52^{NS}$ |
| Dist.Equator | $0.02^{NS}$ | | $0.57^{NS}$ |
| Dist.Manot Cave | 2.95e-04*** | | $0.20^{NS}$ |
| **B*73:01 Recombinants** | | | |
| $R_p^2$ | | 38% | |
| Pathogens | | $0.68^{NS}$ | |
| Dist.Equator | | 1.62e-03** | |
| Dist.Manot Cave | | 3.89e-04*** | |
| $R_p^2$ | | 36% | |
| Bacteria | | $0.59^{NS}$ | |
| Dist.Equator | | 3.5e-03** | |
| Dist.Manot Cave | | 3.03e-05*** | |
| $R_p^2$ | | 39% | |
| Viruses | | $0.52^{NS}$ | |
| Dist.Equator | | 1.48e-03** | |
| Dist.Manot Cave | | 1.33e-03** | |

Dist.Manot Cave is distance from Manot Cave. Dist.Equator is the distance from the Equator (i.e. a measure of absolute latitude). Pathogens is species richness of intracellular pathogens, both bacteria and viruses, in a country from which population was drawn. Bacteria and Viruses denote their respective individual species richness. $R_p^2$ is the pseudo R squared value, a global measure of variation explained by the model for beta regressions, akin to $R^2$ for linear models. p-values for asymptotic Wald tests: *** < 0.001; ** < 0.01; * < 0.05; NS, non-significant.

## REFERENCES


Abi-Rached, L., Jobin, M.J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., Babrzadeh, F., Gharizadeh, B., Luo, M., Plummer, F.A.*, et al.* (2011). The shaping of modern human immune systems by multiregional admixture with archaic humans. Science *334*, 89-94.

Agoni, L., Golden, a., Guha, C., and Lenz, J. (2012). Neanderthal and Denisovan retroviruses. Current Biology *22*, 437-438.

Ahmad, T., Neville, M., Marshall, S.E., Armuzzi, A., Mulcahy-Hawes, K., Crawshaw, J., Sato, H., Ling, K.-L., Barnardo, M., Goldthorpe, S., Walton, R., Bunce, M., Jewell, D.P., Welsh, K.I., 2003. Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC. Human Molecular Genetics 12, 647–656. https://doi.org/10.1093/hmg/ddg066

Barton, N.H., 1979. Gene flow past a cline. Heredity 43, 333–339. https://doi.org/10.1038/hdy.1979.86

Bodmer, W.F. (1972). Evolutionary significance of the HL-A system. Nature *237*, 139-145 passim.

Borcard, D., Legendre, P., 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. Ecological Modelling 153, 51–68. https://doi.org/10.1016/S0304-3800(01)00501-4

Borcard, D., Legendre, P., Avois-Jacquet, C., Tuomisto, H., 2004. Dissecting the spatial structure of ecological data at multiple scales. Ecology 85, 1826–1832. https://doi.org/10.1890/03-3111

Buhler, S., Nunes, J.M., Nicoloso, G., Tiercy, J.M., and Sanchez-Mazas, A. (2012). The heterogeneous HLA genetic makeup of the Swiss population. PloS one *7*, e41400.

Bürger, R., 2017. Two-locus clines on the real line with a step environment. Theoretical Population Biology 117, 1–22. https://doi.org/10.1016/j.tpb.2017.08.002

Canty, A., Ripley, B.D., 2020. boot: Bootstrap R (S-Plus) Functions.

Cashdan, E. (2014). Biogeography of human infectious diseases: a global historical analysis. PloS one *9*, e106752.

Chen, J., Foroozesh, M., and Qin, Z. (2019). Transactivation of human endogenous retroviruses by tumor viruses and their functions in virus-associated malignancies. Oncogenesis *8*, 6.

Cherkasova, E., Weisman, Q., and Childs, R.W. (2013). Endogenous retroviruses as targets for antitumor immunity in renal cell cancer and other tumors. Front Oncol *3*, 243.

Cherkasova, E., Scrivani, C., Doh, S., Weisman, Q., Takahashi, Y., Harashima, N., Yokoyama, H., Srinivasan, R., Linehan, W.M., Lerman, M.I.*, et al.* (2016). Detection of an Immunogenic HERV-E Envelope with Selective Expression in Clear Cell Kidney Cancer. Cancer Res *76*, 2177-2185.

Chuong, E.B., Elde, N.C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science *351*, 1083-1087.

Clarke, K.R., and Warwick, R.M. (2001). Change in marine communities: an approach to statistical analysis and interpretation (Plymouth, United Kingdom).

Cribari-Neto, F., Zeileis, A., 2010. Beta Regression in *R*. J. Stat. Soft. 34. https://doi.org/10.18637/jss.v034.i02

Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and Their Applications. Cambridge University Press, Cambridge.

Diniz-Filho, J.A.F., Soares, T.N., Lima, J.S., Dobrovolski, R., Landeiro, V.L., de Campos Telles, M.P., Rangel, T.F., Bini, L.M., 2013. Mantel test in population genetics. Genet Mol Biol 36, 475–485. https://doi.org/10.1590/S1415-47572013000400002

Doherty, P.C., and Zinkernagel, R.M. (1975). Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. Nature *256*, 50-52.

Douma, J.C., Weedon, J.T., 2019. Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. Methods Ecol Evol 10, 1412–1430. https://doi.org/10.1111/2041-210X.13234

Downey, R.F., Sullivan, F.J., Wang-Johanning, F., Ambs, S., Giles, F.J., and Glynn, S.A. (2015). Human endogenous retrovirus K and cancer: Innocent bystander or tumorigenic accomplice? Int J Cancer *137*, 1249-1257.

Dray, S., Bauman, D., Blanchet, G., Borcard, D., Clappe, S., Guenard, G., Jombart, T., Larocque, G., Legendre, P., Madi, N., Wagner, H.H., 2021. adespatial: Multivariate Multiscale Spatial Analysis.

Dray, S., Legendre, P., Peres-Neto, P.R., 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecological Modelling 196, 483–493. https://doi.org/10.1016/j.ecolmodel.2006.02.015

Enard, D., Cai, L., Gwennap, C., and Petrov, D.A. (2016). Viruses are a dominant driver of protein adaptation in mammals. Elife *5*.

Enard, D., and Petrov, D.A. (2018). Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. Cell *175*, 360-371 e313.

Epperson, B.K., 2005. Estimating dispersal from short distance spatial autocorrelation. Heredity 95, 7–15. https://doi.org/10.1038/sj.hdy.6800680

Epperson, B.K., Li, T., 1996. Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics. Proceedings of the National Academy of Sciences 93, 10528–10532. https://doi.org/10.1073/pnas.93.19.10528

Fehren-Schmitz, L., Jarman, C.L., Harkins, K.M., Kayser, M., Popp, B.N., and Skoglund, P. (2017). Genetic Ancestry of Rapanui before and after European Contact. Curr Biol *27*, 3209-3215 e3206.

Ferrari, S., Cribari-Neto, F., 2004. Beta Regression for Modelling Rates and Proportions. Journal of Applied Statistics 31, 799–815. https://doi.org/10.1080/0266476042000214501

Fox, J., Weisberg, S., 2019. An R Companion to Applied Regression, Third. ed. Sage, Thousand Oaks CA.

Friedlaender, J.S., Friedlaender, F.R., Reed, F.A., Kidd, K.K., Kidd, J.R., Chambers, G.K., Lea, R.A., Loo, J.H., Koki, G., Hodgson, J.A.*, et al.* (2008). The genetic structure of Pacific Islanders. PLoS genetics *4*, e19.

Furstenau, T.N., Cartwright, R.A., 2016. The effect of the dispersal kernel on isolation-by-distance in a continuous population. PeerJ 4, e1848. https://doi.org/10.7717/peerj.1848

Futuyma, D.J., 2005. Evolution. Sinauer Associates, Sunderland, Mass.

Fu, Q., Hajdinjak, M., Moldovan, O.T., Constantin, S., Mallick, S., Skoglund, P., Patterson, N., Rohland, N., Lazaridis, I., Nickel, B.*, et al.* (2015). An early modern human from Romania with a recent Neanderthal ancestor. Nature *524*, 216-219.

Garrison, K.E., Jones, R.B., Meiklejohn, D.A., Anwar, N., Ndhlovu, L.C., Chapman, J.M., Erickson, A.L., Agrawal, A., Spotts, G., Hecht, F.M., Rakoff-Nahoum, S., Lenz, J., Ostrowski, M.A., Nixon, D.F., 2007. T cell responses to human endogenous retroviruses in HIV-1 infection. PLoS Pathog 3, e165. https://doi.org/10.1371/journal.ppat.0030165

Gonzalez, I., Déjean, S., Martin, P., and Baccini, A. (2008). An R Package to Extend Canonical Correlation Analysis. Journal of Statistical Software *23*.

Gonzalez-Galarza, F.F., Takeshita, L.Y., Santos, E.J., Kempson, F., Maia, M.H., da Silva, A.L., Teles e Silva, A.L., Ghattaoraya, G.S., Alfirevic, A., Jones, A.R.*, et al.* (2015). Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. Nucleic acids research *43*, D784-788.

Goulder, P.J., and Walker, B.D. (2012). HIV and HLA class I: an evolving relationship. Immunity *37*, 426-440.

Grandi, N., and Tramontano, E. (2018). Human Endogenous Retroviruses Are Ancient Acquired Elements Still Shaping Innate Immune Responses. Frontiers in immunology *9*, 2039.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., Hansen, N.F., Durand, E.Y., Malaspinas, A.-S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L.F., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., Pääbo, S., 2010. A draft sequence of the Neandertal genome. Science 328, 710–722. https://doi.org/10.1126/science.1188021

Guirimand, T., Delmotte, S., and Navratil, V. (2015). VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. Nucleic acids research *43*, D583-587.

Haldane, J.B. (1948). The theory of a cline. J Genet *48*, 277-284.

Hershkovitz, I., Weber, G.W., Quam, R., Duval, M., Grun, R., Kinsley, L., Ayalon, A., Bar-Matthews, M., Valladas, H., Mercier, N.*, et al.* (2018). The earliest modern humans outside Africa. Science *359*, 456-459.

Hill, A.V.S. (1991). HLA associations with malaria in Africa: some implications for MHC evolution. In Molecular evolution of the major histocompatibility complex, J. Klein, and D. Klein, eds. (Heidelberg, Berlin: Springer Berlin Heidelberg), pp. 403-420.

Hughes, A.L., and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature *335*, 167-170.

Hughes, T.J. (1985). The Great Cenozoic ice sheet. Palaeography, Palaeoclimatology, Palaeoecology *50*, 9-43.

Inkscape Project, 2017. Inkscape.

Jacobs, G.S., Hudjashov, G., Saag, L., Kusuma, P., Darusallam, C.C., Lawson, D.J., Mondal, M., Pagani, L., Ricaut, F.X., Stoneking, M.*, et al.* (2019). Multiple Deeply Divergent Denisovan Ancestries in Papuans. Cell *177*, 1010-1021 e1032.

Jern, P., and Coffin, J.M. (2008). Effects of retroviruses on host genome function. Annu Rev Genet *42*, 709-732.

Jha, A.R., Nixon, D.F., Rosenberg, M.G., Martin, J.N., Deeks, S.G., Hudson, R.R., Garrison, K.E., and Pillai, S.K. (2011). Human endogenous retrovirus K106 (HERV-K106) was infectious after the emergence of anatomically modern humans. PloS one *6*, e20234.

Jones, R.B., Garrison, K.E., Mujib, S., Mihajlovic, V., Aidarus, N., Hunter, D.V., Martin, E., John, V.M., Zhan, W., Faruk, N.F.*, et al.* (2012). HERV-K-specific T

cells eliminate diverse HIV-1/2 and SIV primary isolates. The Journal of clinical investigation *122*, 4473-4489.

Jombart, T., Pontier, D., and Dufour, A.B. (2009). Genetic markers in the playground of multivariate analysis. Heredity (Edinb) *102*, 330-341.

Kassiotis, G. (2014). Endogenous retroviruses and the development of cancer. Journal of immunology *192*, 1343-1349.

Kassiotis, G., and Stoye, J.P. (2016). Immune responses to endogenous retroelements: taking the bad with the good. Nature reviews Immunology *16*, 207-219.

Key, F.M., Teixeira, J.C., de Filippo, C., and Andres, A.M. (2014). Advantageous diversity maintained by balancing selection in humans. Curr Opin Genet Dev *29*, 45-51.

Kieschnick, R., McCullough, B.D., 2003. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. Statistical Modelling 3, 193–213. https://doi.org/10.1191/1471082X03st053oa

Kim, J.H., 2019. Multicollinearity and misleading statistical results. Korean J Anesthesiol 72, 558–569. https://doi.org/10.4097/kja.19087

Klein, J. (1987). Origin of major histocompatibility complex polymorphism: the trans-species hypothesis. Human immunology *19*, 155-162.

Lao, O., Lu, T.T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L.A., Comas, D., Holmlund, G., Kouvatsi, A., Macek, M., Mollet, I., Parson, W., Palo, J., Ploski, R., Sajantila, A., Tagliabracci, A., Gether, U., Werge, T., Rivadeneira, F., Hofman, A., Uitterlinden, A.G., Gieger, C., Wichmann, H.-E., Rüther, A., Schreiber, S., Becker, C., Nürnberg, P., Nelson, M.R., Krawczak, M., Kayser, M., 2008. Correlation between Genetic and Geographic Structure in Europe. Current Biology 18, 1241–1248. https://doi.org/10.1016/j.cub.2008.07.049

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M.*, et al.* (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature *513*, 409-413.

Legendre, P., Legendre, L., 2012. Numerical ecology, Third English edition. ed, Developments in environmental modelling. Elsevier, Amsterdam.

Lenth, R.V., 2021. emmeans: Estimated Marginal Means, aka Least-Squares Means.

Lever, J., Krzywinski, M., Altman, N., 2017. Principal component analysis. Nature Methods 14, 641–642. https://doi.org/10.1038/nmeth.4346

Liu, F., Kong, Y., 2015. zoib: An R Package for Bayesian Inference for Beta Regression and Zero/One Inflated Beta Regression. The R Journal 7, 34.

https://doi.org/10.32614/RJ-2015-019

Lüdecke, D., 2018. ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. JOSS 3, 772. https://doi.org/10.21105/joss.00772

Mafessoni, F. (2019). Encounters with archaic hominins. Nat Ecol Evol *3*, 14-15.

Marchi, E., Kanapin, A., Byott, M., Magiorkinis, G., and Belshaw, R. (2013). Neanderthal and Denisovan retroviruses in modern humans. Curr Biol *23*, R994-R995.

McMichael, A.J., Campbell-Lendrum, D.H., Corvalán, C.F., Ebi, K.L., Githeko, A.K., Scheraga, J.D., Woodward, A. (Eds.), 2003. Climate change and human health: risks and responses. World Health Organization, Geneva.

McMichael, A., 1993. Natural selection at work on the surface of virus-infected cells. Science 260, 1771–1772. https://doi.org/10.1126/science.8511585

Meirmans, P.G., 2012. The trouble with isolation by distance: NEWS AND VIEWS: OPINION. Molecular Ecology 21, 2839–2846. https://doi.org/10.1111/j.1365-294X.2012.05578.x

Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prufer, K., de Filippo, C.*, et al.* (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science *338*, 222-226.

Moreno-Mayar, J.V., Vinner, L., de Barros Damgaard, P., de la Fuente, C., Chan, J., Spence, J.P., Allentoft, M.E., Vimala, T., Racimo, F., Pinotti, T.*, et al.* (2018). Early human dispersals within the Americas. Science *362*.

Mullins, C.S., and Linnebacher, M. (2012). Endogenous retrovirus sequences as a novel class of tumor-specific antigens: an example of HERV-H env encoding strong CTL epitopes. Cancer Immunol Immunother *61*, 1093-1100.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., Stephens, M., Bustamante, C.D., 2008. Genes mirror geography within Europe. Nature 456, 98–101. https://doi.org/10.1038/nature07331

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H., 2020. vegan: Community Ecology Package.

Ospina, R., Ferrari, S.L.P., 2012. A general class of zero-or-one inflated beta regression models. Computational Statistics & Data Analysis 56, 1609–1623. https://doi.org/10.1016/j.csda.2011.10.005

Padgham, M., Sumner, M.D., 2021. geodist: Fast, Dependency-Free Geodesic Distance Calculations.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: Analyses of Phylogenetics

and Evolution in R language. Bioinformatics 20, 289–290. https://doi.org/10.1093/bioinformatics/btg412

Parham, P. (1988). Function and polymorphism of human leukocyte antigen-A,B,C molecules. Am J Med *85*, 2-5.

Parham, P., Arnett, K.L., Adams, E.J., Little, A.M., Tees, K., Barber, L.D., Marsh, S.G., Ohta, T., Markow, T., and Petzl-Erler, M.L. (1997). Episodic evolution and turnover of HLA-B in the indigenous human populations of the Americas. Tissue antigens *50*, 219-232.

Parham, P., and Guethlein, L.A. (2018). Genetics of Natural Killer Cells in Human Health, Disease, and Survival. Annu Rev Immunol *36*, 519-548.

Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G.H., Barreiro, L.B., Froment, A.*, et al.* (2017). Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. Science *356*, 543-546.

Pierini, F., and Lenz, T.L. (2018). Divergent allele advantage at human MHC genes: signatures of past and ongoing selection. Mol Biol Evol.

Posth, C., Nakatsuka, N., Lazaridis, I., Skoglund, P., Mallick, S., Lamnidis, T.C., Rohland, N., Nagele, K., Adamski, N., Bertolini, E.*, et al.* (2018). Reconstructing the Deep Population History of Central and South America. Cell *175*, 1185-1197 e1122.

Prufer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C.*, et al.* (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43-49.

Prugnolle, F., Manica, A., Charpentier, M., Guegan, J.F., Guernier, V., and Balloux, F. (2005). Pathogen-driven selection and worldwide HLA class I diversity. Curr Biol *15*, 1022-1027.

Quach, H., Quintana-Murci, L., 2017. Living in an adaptive world: Genomic dissection of the genus Homo and its immune response. Journal of Experimental Medicine 214, 877–894. https://doi.org/10.1084/jem.20161942

Quintana-Murci, L. (2019). Human Immunology through the Lens of Evolutionary Genetics. Cell *177*, 184-199.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Raghavan, M., Skoglund, P., Graf, K.E., Metspalu, M., Albrechtsen, A., Moltke, I., Rasmussen, S., Stafford, T.W., Jr., Orlando, L., Metspalu, E.*, et al.* (2014). Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature *505*, 87-91.

Rajapaksa, U.S., Li, D., Peng, Y.-C., McMichael, A.J., Dong, T., Xu, X.-N.,

2012. HLA-B may be more protective against HIV-1 than HLA-A because it resists negative regulatory factor (Nef) mediated down-regulation. Proceedings of the National Academy of Sciences 109, 13353–13358. https://doi.org/10.1073/pnas.1204199109

Rakoff-Nahoum, S., Kuebler, P.J., Heymann, J.J., M, E.S., Ortiz, G.M., G, S.O., Barbour, J.D., Lenz, J., Steinfeld, A.D., and Nixon, D.F. (2006). Detection of T lymphocytes specific for human endogenous retrovirus K (HERV-K) in patients with seminoma. AIDS research and human retroviruses *22*, 52-56.

Reche, P.A., and Reinherz, E.L. (2003). Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms. Journal of molecular biology *331*, 623-641.

Reich, D. (2018). Who we are and how we got here (Oxford University Press).

Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L.F., Maricic, T., Good, J.M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E.E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M.V., Derevianko, A.P., Hublin, J.-J., Kelso, J., Slatkin, M., Pääbo, S., 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468, 1053–1060. https://doi.org/10.1038/nature09710

Riley, E., Olerup, O., 1992. HLA polymorphisms and evolution. Immunology Today 13, 333–335. https://doi.org/10.1016/0167-5699(92)90165-4

Robinson, J., Halliwell, J.A., Hayhurst, J.D., Flicek, P., Parham, P., Marsh, S.G.E., 2015. The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res 43, D423-431. https://doi.org/10.1093/nar/gku1161

Rycaj, K., Plummer, J.B., Yin, B., Li, M., Garza, J., Radvanyi, L., Ramondetta, L.M., Lin, K., Johanning, G.L., Tang, D.G.*, et al.* (2015). Cytotoxicity of human endogenous retrovirus K-specific T cells toward autologous ovarian cancer cells. Clin Cancer Res *21*, 471-483.

Salmons, B., Lawson, J.S., and Gunzburg, W.H. (2014). Recent developments linking retroviruses to human breast cancer: infectious agent, enemy within or both? J Gen Virol *95*, 2589-2593.

Schiavetti, F., Thonnard, J., Colau, D., Boon, T., and Coulie, P.G. (2002). A human endogenous retroviral sequence encoding an antigen recognized on melanoma by cytolytic T lymphocytes. Cancer Res *62*, 5510-5516.

Segurel, L., Thompson, E.E., Flutre, T., Lovstad, J., Venkat, A., Margulis, S.W., Moyse, J., Ross, S., Gamble, K., Sella, G.*, et al.* (2012). The ABO blood group is a trans-species polymorphism in primates. Proceedings of the National Academy of Sciences of the United States of America *109*, 18493-18498.

Seymour, R.M., Allan, M.J., Pomiankowski, A., and Gustafsson, K. (2004). Evolution of the human ABO polymorphism by two complementary selective pressures. Proc Biol Sci *271*, 1065-1072.

Shanks, G.D., Hu, Z., Waller, M., Lee, S.E., Terfa, D., Howard, A., van Heyningen, E., and Brundage, J.F. (2014). Measles epidemics of variable lethality in the early 20th century. Am J Epidemiol *179*, 413-422.

Sievert, C., 2020. Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC.

Simas, A.B., Barreto-Souza, W., Rocha, A.V., 2010. Improved estimators for a general class of beta regression models. Computational Statistics & Data Analysis 54, 348–366. https://doi.org/10.1016/j.csda.2009.08.017

Skoglund, P., and Reich, D. (2016). A genomic view of the peopling of the Americas. Curr Opin Genet Dev *41*, 27-35.

Slon, V., Mafessoni, F., Vernot, B., de Filippo, C., Grote, S., Viola, B., Hajdinjak, M., Peyregne, S., Nagel, S., Brown, S.*, et al.* (2018). The genome of the offspring of a Neanderthal mother and a Denisovan father. Nature *561*, 113-116.

Smithson, M., Verkuilen, J., 2006. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychological Methods 11, 54–71. https://doi.org/10.1037/1082-989X.11.1.54

Sokal, R.R., Jacquez, G.M., 1991. Testing Inferences about Microevolutionary Processes by Means of Spatial Autocorrelation Analysis. Evolution 45, 152. https://doi.org/10.2307/2409490

Sokal, R.R., Wartenberg, D.E., 1983. A Test of Spatial Autocorrelation Analysis Using an Isolation-by-Distance Model. Genetics 105, 219–237.

Stephens, P.R., Altizer, S., Smith, K.F., Alonso Aguirre, A., Brown, J.H., Budischak, S.A., Byers, J.E., Dallas, T.A., Jonathan Davies, T., Drake, J.M.*, et al.* (2016). The macroecology of infectious diseases: a new perspective on global-scale drivers of pathogen distributions and impacts. Ecol Lett *19*, 1159-1171.

Szolek, A., Schubert, B., Mohr, C., Sturm, M., Feldhahn, M., and Kohlbacher, O. (2014). OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics *30*, 3310-3316.

Szpakowski, S., Sun, X., Lage, J.M., Dyer, A., Rubinstein, J., Kowalski, D., Sasaki, C., Costa, J., and Lizardi, P.M. (2009). Loss of epigenetic silencing in tumors preferentially affects primate-specific retroelements. Gene *448*, 151-167.

Takahashi, Y., Harashima, N., Kajigaya, S., Yokoyama, H., Cherkasova, E., McCoy, J.P., Hanada, K., Mena, O., Kurlander, R., Tawab, A.*, et al.* (2008). Regression of human kidney cancer following allogeneic stem cell transplantation is associated with recognition of an HERV-E antigen by T cells. The Journal of clinical investigation *118*, 1099-1109.

Tandon, R., SenGupta, D., Ndhlovu, L.C., Vieira, R.G., Jones, R.B., York, V.A., Vieira, V.A., Sharp, E.R., Wiznia, A.A., Ostrowski, M.A.*, et al.* (2011).

Identification of human endogenous retrovirus-specific T cell responses in vertically HIV-1-infected subjects. Journal of virology *85*, 11526-11531.

Tobler, W.R., 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. Economic Geography 46, 234. https://doi.org/10.2307/143141

Tu, X., Li, S., Zhao, L., Xiao, R., Wang, X., and Zhu, F. (2017). Human leukemia antigen-A*0201-restricted epitopes of human endogenous retrovirus W family envelope (HERV-W env) induce strong cytotoxic T lymphocyte responses. Virol Sin *32*, 280-289.

Urban, M.C., 2011. The evolution of species interactions across natural landscapes: Urban - evolution of species interactions. Ecology Letters 14, 723–732. https://doi.org/10.1111/j.1461-0248.2011.01632.x

Vaidyanathan, R., Xie, Y., Allaire, J.J., Cheng, J., Sievert, C., Russell, K., 2020. htmlwidgets: HTML Widgets for R.

Volkman, H.E., and Stetson, D.B. (2014). The enemy within: endogenous retroelements and autoimmune disease. Nature immunology *15*, 415-422.

Wang-Johanning, F., Radvanyi, L., Rycaj, K., Plummer, J.B., Yan, P., Sastry, K.J., Piyathilake, C.J., Hunt, K.K., and Johanning, G.L. (2008). Human endogenous retrovirus K triggers an antigen-specific immune response in breast cancer patients. Cancer Res *68*, 5869-5877.

Wildschutte, J.H., Williams, Z.H., Montesion, M., Subramanian, R.P., Kidd, J.M., and Coffin, J.M. (2016). Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proceedings of the National Academy of Sciences of the United States of America *113*, E2326-2334.

Wright, W.J., Irvine, K.M., Warren, J.M., Barnett, J.K., 2017. Statistical design and analysis for plant cover studies with multiple sources of observation errors. Methods Ecol Evol 8, 1832–1841. https://doi.org/10.1111/2041-210X.12825

Yu, V.L., Edberg, S.C., 2005. Global Infectious Diseases and Epidemiology Network (GIDEON): A World Wide Web-Based Program for Diagnosis and Informatics in Infectious Diseases. Clinical Infectious Diseases 40, 123–126. https://doi.org/10.1086/426549

Zeileis, A., 2016. variance vs. standard deviation in beta regression. Cross Validated (accessed 12-06-2020).

Zeng, M., Hu, Z., Shi, X., Li, X., Zhan, X., Li, X.D., Wang, J., Choi, J.H., Wang, K.W., Purrington, T.*, et al.* (2014). MAVS, cGAS, and endogenous retroviruses in T-independent B cell responses. Science *346*, 1486-1492.

Zimmermann, C., Kowalewski, D., Bauersfeld, L., Hildenbrand, A., Gerke, C., Schwarzmüller, M., Le-Trilling, V.T.K., Stevanovic, S., Hengel, H., Momburg, F., Halenius, A., 2019. HLA-B locus products resist degradation by the human cytomegalovirus immunoevasin US11. PLOS Pathogens 15, e1008040. https://doi.org/10.1371/journal.ppat.1008040

# Chapter 3
# HIV Subtype Diversity & Human HLA Diversity

# Introduction

The "dominant paradigm" for over 20 years has been that the observed geographic distribution of HIV-1 group M subtypes and circulating recombinant forms (CRFs) is merely an artefact of random founder effects and timing, rather than being a result of adaptive processes (Hu et al., 1999; Rambaut et al., 2001).

## *HIV-1 Group M Subtypes*

HIV-1 Group M is subdivided into subtypes A-K as well as over a hundred CRFs ("Los Alamos National Laboratory HIV Sequence Database," 2019). The subtypes are roughly phylogenetically equidistant from one another and the CRFs stem from co-infection and recombination by different subtypes of the virus within an individual. Some subtypes, such as A and F, are further divided into sub-subtypes, e.g. A1 and A2, where genetically distinct sister clades can be identified within a subtype (Buonaguro et al., 2007; Robertson et al., 2000). Geographically distinct lineages within a subtype such as Ethiopian C or East African D are not phylogenetically equidistant but are still phylogenetically distinguishable (Taylor et al., 2008). Subtypes can also generally be distinguished from one another via subtype-specific positions (SSPs) in the virus' amino acid (AA) sequence ("Los Alamos National Laboratory HIV Sequence Database," 2019; Tenzer et al., 2014), however individual strains from one subtype can mutate towards SSPs of another subtype (Kist et al., 2020). HIV subtype is important not just for tracking the pandemic on a global scale but also has profound clinical effects, affecting

replicative fitness e.g. (Ariën et al., 2005; Njai et al., 2006), disease progression e.g. (Baeten et al., 2007; Kaleebu et al., 2002), viral load e.g. (Fischetti et al., 2004; Sarr et al., 2005), and transmission rates (Hudgens et al., 2002; Kiwanuka et al., 2009; Sarr et al., 2005).

## *The Flawed Paradigm of HIV Subtype Diversity*

Five distinct lineages or subtypes of Group M HIV-1 were reported in 1992 (Myers et al., 1992). Questions as to the origin of subtypes arose as soon as subtypes were identified and their likely importance to future vaccine efforts was also identified early on (Myers, 1994; WHO network for HIV isolation and characterization, 1994). Some researchers were quick to dismiss the possibility of coevolution:

> One possibility may be quickly eliminated: the distribution of subtypes within countries as well as throughout the world shows that there is no host specificity for the HIV-1 subtypes, and therefore that they have not coevolved with particular human host populations, as appears to have happened with human papillomaviruses.
>
> Myers, (1994)

This dismissal appears to have been premature and likely based on limited data. Compared to the hundreds of thousands of HIV sequences available today online ("Los Alamos National Laboratory HIV Sequence Database," 2019), the 1994 physical HIV compendium and accompanying floppy diskettes contained only a few hundred HIV sequences (Myers et al., 1995). Modern analyses show very clear and uneven geographic

distributions of subtypes (Hemelaar et al., 2019, 2006, “Los Alamos National Laboratory HIV Sequence Database,” 2019). The other problematic aspect of this dismissal is comparing HIV with another virus (in this case human papillomaviruses) while stating that HIV, for some unexplained reason, has not been subject to the same host/pathogen adaptive forces seen in the other virus. This is rather strangely not an isolated incident. For example in Pybus and Rambaut (2009), the authors contrast HIV to influenza A, and in the same sentence say that viral adaptation played little part in global subtype diversity but that there is both evidence for the virus having mutated for efficient transmission between humans and that it is currently adapting to class I HLAs. While Myers's 1994 dismissal only states that the reason we see HIV subtypes is definitely not due to host specificity, rather than saying what the reason is, it helped set the stage for what would become the dominant paradigm regarding HIV subtypes.

This paradigm is that the global distribution of subtypes is not due to adaptation or "inherent properties" of the strains, but are rather the chance result of founder effect where a subtype is introduced and established in a population before other subtypes (Hu et al., 1999). When Hu et al. (1999) proposed this over 20 years ago, it was originally formulated in an epidemiological framework, rather than a phylogenetic one. As with Myers, (1994), some claims by Hu et al. (1999) have not aged well. Although studies, e.g. (Foy et al., 1995), already existed that showed links between subtype and phenotypical differences like transmission, Hu et al. (1999) claimed "that it is highly unlikely that a single characteristic such as subtype can account for significant

differences in transmission and disease progression." Nevertheless, numerous subsequent studies have shown that subtypes do indeed have significant phenotypical differences between them (Ariën et al., 2005; Baeten et al., 2007; Fischetti et al., 2004; Hudgens et al., 2002; Kaleebu et al., 2002; Kiwanuka et al., 2009; Njai et al., 2006; Sarr et al., 2005; Tenzer et al., 2014). The issue of timing raised by Hu et al. (1999) was already problematic as evidenced by the HIV epidemic in Thailand. The first HIV-1 infected people in Thailand were detected in the mid-1980s, consisting of two patients with thalassemia and a small number of male commercial sex workers in Bangkok (Bunnell et al., 1999; Smith, 1990; Wangroongsarb et al., 1985). These infections were likely subtype B due to the risk profiles of the infected individuals and the fact that CRF01_AE (subtype E) first appeared in the northern part of the country (Korber et al., 2000; Kuiken, 2000). In the mid-1980s, no HIV-1 was detected in the risk groups where CRF01_AE became prevalent, namely intravenous drug users and soldiers, but in the late 1980s, CRF01_AE experienced explosive growth in these populations (Bunnell et al., 1999; Smith, 1990; Wangroongsarb et al., 1985). CRF01_AE is overwhelmingly (>95%) more common in Thailand (Wirachsilp et al., 2007), despite the fact that subtype B appeared in the country first and should therefore according to the dominant paradigm be the primary subtype of the local epidemic. Other instances where pure subtypes have been gradually overtaken by newly appearing CRFs, such as the epidemics in Malaysia, Brazil, and West Africa, also contravene this paradigm (Lau and Wong, 2013a)

For whatever reason, Hu et al. (1999) did not get nearly as much traction as subsequent efforts proposed under a phylogenetic framework. For example, articles by Peeters and Sharp (2000) and Rambaut et al. (2001) and subsequent derived works such as Rambaut et al. (2004), do not cite Hu et al. (1999), indicating that the authors seem to have arrived at this independently despite proposing a nearly identical idea. Regardless, a review by Peeters and Sharp (2000) looked at subtypes in a phylogenetics framework and proposed that the subtypes likely arose from founder effects, though did not explicitly state that it was due to random chance. The paradigm was refined slightly into its more or less modern form by Rambaut et al. (2001), again under a phylogenetic framework rather than a purely epidemiological one. The source of these chance exportations was narrowed down from western Africa to the Democratic Republic of the Congo (DRC) specifically. Rambaut et al. (2001) do not cite Hu et al. (1999) or Peeters and Sharp (2000), although Peeters is a co-author with Rambaut et al. (2001) and so presumably aware of at least the latter work. The phylogenetic reasoning behind the dominant paradigm is that the global structure of the tree of HIV-1 subtypes has a "starburst" appearance because of the apparently simultaneous appearance of the viral subtypes (Rambaut et al., 2001). This structure is what one would expect with the exponential population growth of an epidemic (Holmes et al., 1995). Many Congolese strains are basal or fall between the established global subtypes (Vidal et al., 2000), and Rambaut et al. (2001) propose this is due entirely to founder effects specifically from chance exportations. Rambaut et al. (2001) reference an earlier paper by Korber et al. (2000) as supporting their point, stating "such founder effects have been proposed to explain the phylogenetically

distinct subtypes B [in the US] and E [CRF01_AE, in Thailand] of HIV-1 group M". However, Korber et al. (2000) does not support the founder effect paradigm for either subtype. For subtype B, the Korber et al. (2000) analysis explicitly states that the subtype B epidemic in the US is unlikely to be the result of a single founder virus, and that it more likely the result of a 5 to 15 year period of subtype B evolution where the virus was circulating but not clinically recognised. For the CRF01_AE (subtype E) epidemic in Thailand, Korber et al. (2000) point out the issues covered earlier concerning the initial appearance of subtype B and its subsequently being supplanted by CRF01_AE. This is contrary to what the dominant paradigm says should occur. Although not directly challenging the primacy of founder effects, McVean (2002) questions the validity of conclusions drawn from a model such as that found in Rambaut et al. (2001) which explicitly ignores recombinant sequences in a virus prone to rampant recombination. Despite these early problems, the initial Rambaut et al. (2001) paper has over 200 direct citations, mostly affirming it, including recent papers e.g. (Archer and Robertson, 2007a; Faria et al., 2019; Ferdinandy et al., 2015; Grant et al., 2020; Roques et al., 2002; Tebit and Arts, 2011). Citations of later review articles restating the founder effect paradigm are even more widespread e.g. Pybus and Rambaut (2009) and Rambaut et al. (2004) with close to 500 and over 700 citations respectively. While subtype diversity being due entirely to founder effect artefacts has become the dominant paradigm in the field, it has changed little in the intervening years as can be seen from the relatively brief sampling presented next.

### *Yearly Examples*

A review article by Thomson et al. (2002) asserts, "Outside central Africa, usually one or two genetic forms are predominant in most countries, which could suggest a founder effect, whereby the earliest genetic form successfully introduced into a population establishes itself as predominant, gaining an initial advantage over other genetic forms arriving later". As detailed earlier, this argument does not match up with observations both at the time in Thailand (Bunnell et al., 1999; Kalish et al., 1994; Smith, 1990; Wangroongsarb et al., 1985) and later ones elsewhere (Lau et al., 2010; Lau and Wong, 2013a). The Thomson et al. (2002) review is also looking at the phylogeny of the subtypes so it seems likely that they should have at least been familiar with the earlier phylogenetic works by Peeters and Sharp (2000) or Rambaut et al. (2001), if not the epidemiological framework of Hu et al. (1999), even if they do not cite any earlier materials with this claim. This could be indicative of early entrenchment of the dominant paradigm, that Thomson et al. (2002) already considered this to be "common knowledge". A research article by Roques et al. (2002) reiterate the random nature of subtypes stating that "no conclusive evidence for any biological correlate with subtypes has been found other than geographical associations" and "the group M subtypes seem to represent little more than founder effects in the pandemic, i.e., the chance exportation of strains from a continuum of diversity in the DRC region to other African regions and to the rest of the world". Even at the time there was evidence that subtype had biological effects, e.g. Foy (1995) showed that subtype CRF01_AE was associated

with higher risk of transmission than subtype B in Thailand. Insofar as any published research can be considered "conclusive evidence", numerous subsequent studies have supported the biological effects of subtype e.g. (Ariën et al., 2005; Baeten et al., 2007; Fischetti et al., 2004; Hudgens et al., 2002; Kaleebu et al., 2002; Kiwanuka et al., 2009; Njai et al., 2006; Sarr et al., 2005; Tenzer et al., 2014). The next year, Yamaguchi et al. (2003) restate the chance exportation founder effect paradigm and assert that "no phenotypic differences have been identified between subtypes", which is again, demonstrably false.

2004 saw the publication of the widely cited review article by Rambaut et al. (2004). In it, the basic premise of the dominant paradigm is rehashed, with some expansions. An earlier work by Vidal et al. (2000) had found a large amount of HIV diversity in the DRC including strains falling between the previously described subtypes, and Rambaut et al. (2004) cite this as evidence that the subtype structure observed globally is due to founder effects and incomplete sampling and/or sampling bias. More sampling and in particular sampling of strains falling between known existing subtypes would by definition change the tree topology, and probably change the average phylogenetic distance between subtypes. However, the relevance of this to why we observe largely monophyletic clusters (and recombinants between them) outside of the DRC is not entirely clear. The fact that subtypes B and D are more closely related to each other than they are to subtype C has little bearing on why subtype C is dominant in southern Africa and subtype B is dominant in North America and Western Europe. Rambaut et

al. (2004) cite a study by Ball et al. (2003) which showed that subtype C viruses have lower fitness when competed against subtype B viruses, and take this as evidence "that the high prevalence of subtype C in sub-Saharan Africa is the result of its chance entry into populations with high rates of partner exchange." However, when the Ball et al. (2003) study is examined more closely, the methodology involved ex vivo testing of the viruses in donor derived primary blood mononuclear cells (PBMC). The study unfortunately does not list the ethnicities of the donors but it appears to have been carried out at Case Western Reserve University which has been about ~50% White American vs. ~5% African American as long as data have been recorded (Case Western Reserve University Institutional Research, 2020), and the United States like most economically developed countries generally has a problem with underrepresentation of ethnic/racial minorities in blood donor populations (Makin et al., 2019). It is therefore highly likely that most if not all of the PBMC used by Ball et al. (2003) came from white blood donors. Given the work shown in this thesis as well as previous studies e.g. (Tenzer et al., 2014), it is perhaps unsurprising that subtype B outcompeted subtype C in PBMC derived from white donors.

The Rambaut et al. (2004) Review also tries to explain why natural selection is not as potent a force between hosts as it is within them. The first argument given is that there is a bottleneck that occurs during inter-host transmission which reduces genetic diversity. The problem with this argument is that it appears to be assuming *a priori* that the transmission event which leads to a genetic bottleneck is an entirely random event

when it is well established that this is not the case: within subtypes CCR5-tropic viruses are preferentially transmitted (Shaw and Hunter, 2012) and HIV subtype affects transmission rates (Hudgens et al., 2002; Kiwanuka et al., 2009; Sarr et al., 2005). As an afterthought the review does at least acknowledge, "Of more debate is whether a bottleneck has a selective component", but does not further address this issue. The second argument Rambaut et al. (2004) give for why natural selection does not substantially impact subtype diversity is the behavioural aspects of HIV transmission, as the virus is predominantly sexually transmitted: "As a result, strains with advantageous mutations could, by chance, find themselves in individuals with low rates of partner exchange and so will not be transmitted far in the population." Human papillomavirus (HPV) is the most common sexually transmitted infection (STI) in the world (Milner, 2020; World Health Organization, 2016), and it is long and well documented that HPV has undergone significant coevolution with specific human host populations (Chen et al., 2018; Ho et al., 1993; Ong et al., 1993; Van Doorslaer, 2013). Indeed, the type of random extinguishing of organisms with advantageous mutations that Rambaut et al. (2004) describe fits under the definition of genetic drift experienced by all evolving organisms, but this issue is not addressed.

Other works in 2004, such as by Holmes (2004), reiterate the hypothesis, "This indicates that subtypes are primarily produced by the chance exportation of some viral lineages to other localities, in Africa and beyond, which then ignite local epidemics". Holmes (2004) further states that there is little evidence that the observed subtype phylogenies

are shaped by fitness differences among the different subtypes, and the same Ball et al. (2003) paper discussed earlier is cited as evidence for chance exportation rather than selection. For the next several years, various research papers e.g. (Archer and Robertson, 2007a), review articles e.g. (Anastassopoulou and Kostrikis, 2006; Galetto and Negroni, 2005), and books e.g. (Volberding, 2008) repeat the dominant paradigm with essentially no change.

Some take the idea of HIV subtypes being due to incomplete sampling a step further and argue that subtypes as phylogenetically discrete entities do not exist at all (Abecasis et al., 2007; Holmes, 2009). Both of these are based on the idea that inter-subtype recombination is common and confounds attempts at drawing neat and tidy single phylogenies. This argument is partially correct, in the sense that where the lines are drawn between subtypes will be to some extent arbitrary. Subtypes are "roughly" phylogenetically equidistant as opposed to exactly equidistant. Subtypes B and D are about as closely related as the sub-subtypes A1 and A2 (Thomson et al., 2002) but have remained labelled as discrete subtypes for backwards compatibility of the literature. There are also categorisation problems stemming from acyclic graphs (trees) having difficulty handling things like introgressions, recombination, and horizontal gene transfer. However, despite these categorisation issues, the concept of subtypes is still widely used as they are broadly discernible from one another both phylogenetically ("Los Alamos National Laboratory HIV Sequence Database," 2019), at SSPs (Tenzer et al., 2014), and phenotypically (Ariën et al., 2005; Baeten et al., 2007; Fischetti et al.,

2004; Hudgens et al., 2002; Kaleebu et al., 2002; Kiwanuka et al., 2009; Njai et al., 2006; Sarr et al., 2005). As Abecasis and Vandamme (2015) point out later, stratifying HIV strains into subtypes and CRFs is useful for studying the impact HIV-1 genetic diversity on biological markers such as pathogenesis, and also provides a means of monitoring the spread of the different epidemics around the world.

In another often cited review by Pybus and Rambaut (2009), the dominant paradigm is presented again without alteration from previous iterations, but they do offer an alternative explanation for why viral adaptation (i.e. natural selection) does not shape the geographic distribution of subtypes. They posit the lack of protective immunity against HIV as the cause of this but at the same time acknowledge the virus acquired specific mutations to enable efficient human transmission after zoonosis and that it is currently adapting to class I HLAs. Both of these examples are indicative of viral adaptation in the absence of protective immunity which could affect the geographic distribution of subtypes. For example, there is no protective immunity to other types of HIV, such as HIV-1 Group O or HIV-2, but both of these variants account for only a small percentage of global HIV infections and are largely confined to west Africa. Group O appears to have an increased susceptibility to tetherin, and this lack of adaption to human antiviral mechanisms has been proposed as a reason for its lesser geographic spread (Sauter et al., 2009). Similarly, there is no protective immunity to HIV-2 (though many people are naturally long term non-progressors), but it is much less infectious than HIV-1, with five to ten fold lower rates of transmission for

heterosexual sex and 20 to 30 fold lower for vertical transmission (Ingole et al., 2013). HIV-2 is in decline in west Africa relative to HIV-1 despite having likely arrived earlier than HIV-1 in some areas (Omobolaji T. Campbell-Yesufu and Gandhi, 2011; De Cock, 1993).

Subsequent years up until the present saw numerous instances the dominant paradigm being repeated with no change or any additional justifications, e.g. (Abecasis and Vandamme, 2015; Bulla et al., 2010; Cordingley, 2017; Désiré et al., 2018; Faria et al., 2019, 2014; Grant et al., 2020; Hemelaar, 2012; Müller et al., 2011; Papkou et al., 2016; Peeters et al., 2013; Tebit and Arts, 2011; Vermund and Leigh-Brown, 2012; Woo et al., 2010). Ferdinandy et al. (2015) simulated HIV epidemics with an older established strain and a second, recently introduced, invading strain. They found that the invading strain needed >25% better transmission efficiency than the established strain in order to become dominant on a time scale of decades. With less than this, the new strain usually went extinct and otherwise took hundreds of years to replace the established strain, a time scale that is not relevant for the history of the HIV pandemic. In light of this, Ferdinandy et al. (2015) conclude that the dominant paradigm is correct and that the current subtypes dominant in any given area are probably suboptimal strains that expanded by chance and have retained their prevalence simply due to first comer advantage. However, the benefits of first comer advantage in their simulations only exist if infection with one of the strains provides robust protection against superinfection. This does not seem to be the case in real epidemics (Redd et al., 2013).

Furthermore, as covered earlier, newly emergent CRFs have overtaken or are in the process of overtaking previously established strains in a variety of locations including Thailand, Malaysia, Brazil, and west Africa (Lau and Wong, 2013a; Wirachsilp et al., 2007).

It should be noted that the argument here is quite narrow. It has been well established for quite some time that the human immune system, in particular HLA type affects HIV e.g. (Moore et al., 2002) and that HIV subtype affects prognoses e.g. (Kaleebu et al., 2002). Neither of these ideas are very disputed, even by people promoting the dominant paradigm e.g. (Pybus and Rambaut, 2009). Rather, the debate is over whether these effects have an impact on subtype diversity on a global scale. The dominant paradigm is that the observed distribution of subtypes is due entirely to founder effects caused by chance exportations, and that viral adaptation plays no role. We argue that while founder effects are important, they are not the entire explanation. Viruses that are by chance better adapted at evading the HLA profile of a population will be more likely to take root in that population and spread efficiently. This will lead to an observed founder effect, but the founder strain is not selected at random from all possible strains. There will be instances where it really is just a single strain that has been exported by chance and without competition to a new region. However, in all cases any founder strain will acquire mutations that make it better able to spread in the new population, leading to geographically distinct clades. If the affected population in a geographic area shifts, e.g. if a regional epidemic shifts prevalence to a new ethnic group, then selected

for mutations in dominant strain should start to reflect this.

## *HIV, HLA, & Ethnolinguistic Diversity*

The Allele Frequency Database (AFD) is an online repository of immune gene frequencies for varying populations around the world (González-Galarza et al., 2015). Unfortunately, the data for many African countries are either extremely limited or entirely non-existent. However, decades of genetic research both around the world and in Africa specifically has shown a strong link between ethnolinguistic diversity and genetic diversity, and the latter can be used as a viable proxy for the former (Excoffier et al., 1991a, 1987a; Pagani et al., 2012; Patin et al., 2017; Sanchez-Mazas et al., 2011; Tishkoff et al., 2009; Zamani et al., 2013). The HLA loci of geographically dispersed populations show higher levels of local adaptation than loci in the rest of the genome, further supporting this as a proxy (Cao et al., 2004; Meyer et al., 2018).

Using ethnolinguistic information is not without its own issues though. The linguist Max Weinreich famously quipped "A language is a dialect with an army and navy" ("Of dialects, armies and navies," 2010), underscoring that more often it is social and political issues that divide "languages" from "dialects". Although attempts have been made to rigorously define these terms, no single definition handles all cases and they run into the same issues as the term "species" in biology, *e.g.* languages at each end of a dialect continuum being unintelligible but one can trace a stepwise path of mutually

intelligible intermediaries between the two. Additionally, given language's important role in group identity, the topic is often a political minefield. Fortunately, while the minutia can be difficult to sort out there is some consensus at the higher level. Broadly speaking there are more than 2000 different ethnolinguistic groups in Africa which can, excluding a few isolates, be divided into various language families (Center for Geographic Analysis, 2013; Congo Basin Art History Research Center et al., 2001; Tishkoff et al., 2009). If it is 1900 or 2100 different groups is not hugely important for the work here.

My work here demonstrates a strong correlation between ethnolinguistic diversity (as a proxy for genetic diversity) and the HIV-1 sequence and subtype diversity in a country, which casts doubts on the idea that HIV-1 subtype diversity is solely from founder effects.

# Methods

## *Shannon Entropy*

Briefly, Shannon entropy, also known as Shannon Diversity, is a way to quantify uncertainty and information content. With a string of text, including a sequence alignment, there are different letter options and those options have different proportional likelihoods of appearing. Increasing either the amount of options or making the options more proportionally equal makes it more difficult to accurately predict the next letter in the string. The quantification of this unpredictability or uncertainty is the entropy. It is calculated by the following equation (sometimes with different log bases such as ln):

$$H = -\sum_{i=1}^{n} p_i \log p_i \qquad \text{Equation 3.1}$$

In the case of an AA alignment $n$ is a 22 letter alphabet of the 20 standard amino-acids plus a gap character "-" and mask character "X", and $p_i$ is the proportion of amino acid $i$ within the alignment. Each proportion is multiplied by its logarithm and the results multiplied by -1 after being summed together to give the Shannon Entropy $H$.

## *African HIV-1 Diversity*

To calculate the Shannon entropy of HIV-1 group M subtype diversity for each country I searched the LANL HIV database for all HIV-1 sequences originating from Africa ("Los Alamos National Laboratory HIV Sequence Database," 2019), yielding 176,071

sequences. The LANL database automatically filters out "problematic sequences" according a variety of conditions (Table 3.1).

*Table 3.1* List of problem codes and their descriptions from ("Los Alamos National Laboratory HIV Sequence Database," 2019)

| Problem Code | Description | Note |
|---|---|---|
| N | High content of non-ACTG character | More than 100 consecutive non-ACTG characters<br>>7% non-ACTG characters for sequences of length <1000<br>>5% non-ACTG characters for sequences of length 1000-2999<br>>3% non-ACTG characters for sequences of length 3000+ |
| C | Contaminant | Likely contaminated with a laboratory strain |
| H | Hypermutant | G → A hypermutation |
| S | Synthetic | A non-naturally-occurring viral sequence (e.g. containing SIV component) |
| D | Deletion | Artifactual deletion of >100 nucleotides (e.g. when an author concatenates 2 sequences from a single sample but leaves out some intervening sequence. |
| T | Tiny | Less than 50 bp |
| R | Reverse Compliment | Sequence deposited as its reverse compliment |

I then removed non-HIV-1M sequences (N,O, P, and CRFs thereof), filtered for sequences originating in Sub-Saharan African (as defined by the United Nations Statistics Division, not LANL which does not include Chad, Mauritania, Mali, Niger, and variably does not include Sudan), and restricted sequences to 1 sequence per patient yielding 36,044 sequences (Figure 3.1). This last step was carried out in order to avoid oversampling individual patients. The number of subtypes per country, and

the number of each subtype per country was then calculated from these sequences.

To calculate p24 Gag sequence diversity, I downloaded full p24 sequences as well as sequences labelled as fragments that were at least 684 NA in length (~99% of the HXB2 reference sequence p24 length, GenBank: K03455) originating in Sub-Saharan Africa (again using the United Nations Statistics Division definition) and discarded N,O, and P sequences, and CRFs labelled as containing portions of them. In order to avoid artificial oversampling of individual patients I filtered for one sequence per patient, using patient IDs supplied by the submitting authors. In order to avoid the possibility

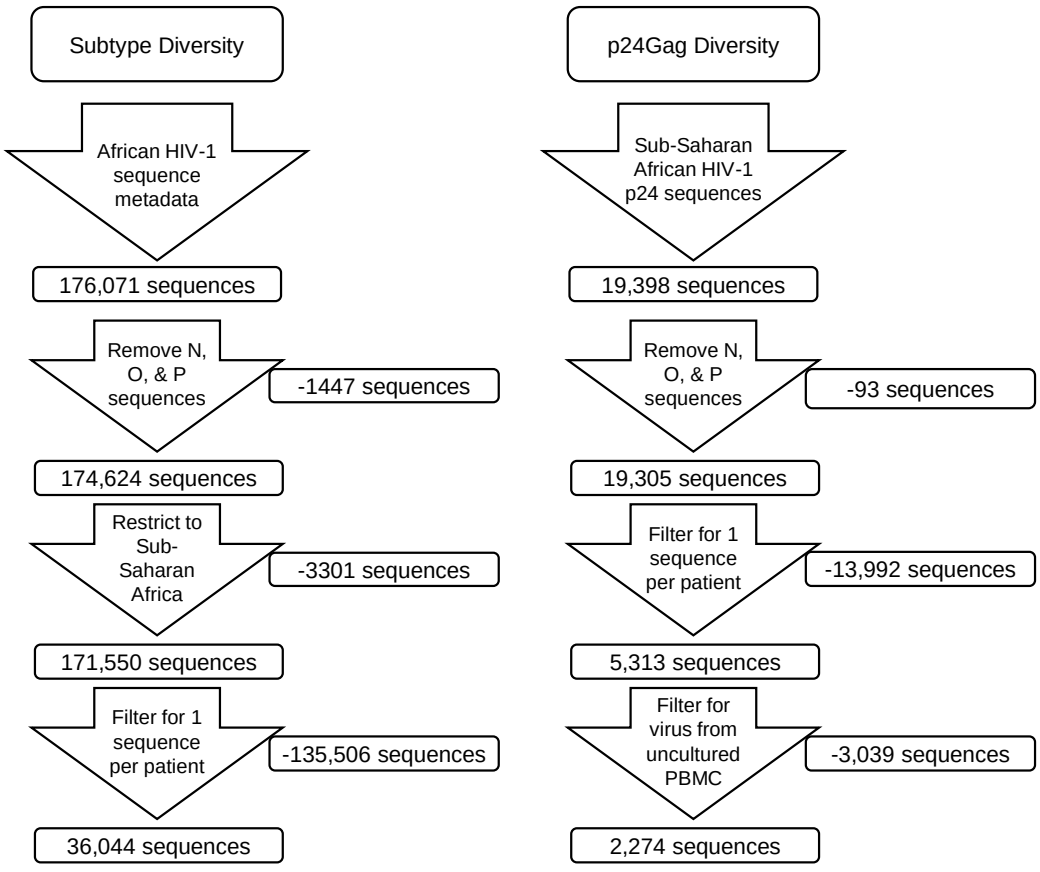



*Figure 3.1* Schematic diagram of filtration steps used in selecting sequences from the LANL HIV Database for analyses. PBMC: peripheral blood mononuclear cells.

of inadvertently obtaining multiple sequences from a single patient I also removed all sequences without an ID. This filtered down from 19,305 sequences to 5,313. I created custom code to extract the accession numbers from the sequence metadata, and modified code provided to me by another student in the lab (Nico Kist) to extract the source papers from the accession numbers. I looked through 264 papers, including 34 unpublished direct submissions, to determine how the sequence was obtained (See appendix 1A for full list). Sequences originating from uncultured peripheral blood mononuclear cells were kept while others were discarded as the rapid mutation rate of HIV means culturing will likely mask the sequence footprints of the original host's immune response. Additionally, culturing could likely select for different sequences and cause an over estimation of sequence diversity. This yielded a final total of 2,274 sequences.

I aligned these sequences on a per codon basis using the LANL HIV Align program ("Los Alamos National Laboratory HIV Sequence Database," 2019). I then stratified sequences by country and made minor manual corrections to each country's alignment as required. LANL HIV Align was chosen over systems such as MUSCLE, MAUVE, or MAFFT as it is purpose built for HIV and handles insertions and deletions (indels) within HIV much better. This was necessary due to the codon based nature of the alignments. I then translated the nucleic acid alignments into amino acid alignments for subsequent analyses.

### *Sub-Saharan African Ethnic Diversity & Ethnolinguistic Groups*

Ethnolinguistic group geospatial data used for constructing the maps in figure 3.2 were obtained from the Center for Geographic Analysis (2013). The two most well accepted analyses for global ethnic diversity (also known as fractionalization in the economics context of the sources) have slightly different methodologies (Alesina et al., 2003; Fearon, 2003). Fearon's work to some extent side steps the intractable problem of how to define an "ethnic group". It does this by looking at what people in each individual country identify as according to their most socially relevant self-defined groupings. As such it will, for example, split Muslim Arabs from Christian Arabs in a country like Lebanon where those two groups are seen as ethnically distinct groups but not in a country like the United States where they are not. As such it is more useful for the social sciences and less useful for the present study as it can artificially distort, primarily by inflation, the ethnic diversity of a country. The other broadly used analysis by Alesina et al. compiles data from a variety of expert sources such as the Encyclopaedia Britannica, the Central Intelligence Agency of the United States, the World Bank, etc. It uses these data to calculate the ethnic diversity of a country. While neither method is perfect and both are best approximations, Alesina's work seems to be closer to the biological diversity that I am interested in and therefore I used it as one of the means of determining ethnic diversity in Sub-Saharan Africa. I manually extracted the ethnic proportions data from Alesina et al. (2003) for Sub-Saharan Africa. The Shannon entropy for each country was then calculated from these data, as per Equation 3.1 where in this instance $n$

indicates the number of ethnicities recorded, and $p_i$ is the proportion of ethnicity $i$ within the country's population.

### *Ethnic Diversity Versus Land Area*

Land area was obtained from the World Bank Data Catalog [sic] of World Development Indicators (World Bank, 2019). I fit an unweighted linear model in R version 3.4.3 (R Core Team, 2018) to assess statistical significance of ethnic diversity versus land area within a country.

### *HIV-1 p24 Gag Diversity Versus Ethnic Diversity*

I calculated mean amino acid Shanon Entropy for each country's p24 Gag sequences using the R version 3.4.3 (R Core Team, 2018) package bio3d version 2.3-3 (Grant et al., 2006). The bio3d package calculates Shannon entropy in the standard fashion described earlier. This generated an entropy for every position in the alignment and I took the mean of these entropies for the sequence entropy. I calculated the mean sequence entropy for each country by taking the mean entropy of all the sequence entropies.

I plotted the mean HIV-1 p24 Gag AA entropy against the ethnic diversity Shannon entropy on a per country basis, and weighted by the number of p24 Gag sequences per country obtained from the LANL database. In order to graphically represent this

weighting I logarithmically scaled each dot. Plots were constructed using ggplot2 version 2.2.1 (Wickham, 2009). I then fit a weighted linear model in R version 3.4.3 (R Core Team, 2018) to assess statistical significance.

### *HIV-1 Subtype Diversity Versus Ethnic Diversity*

When downloading sequences from the LANL database I created a custom sequence naming system that compensated for the disparate naming conventions used by submitting authors and the use of special characters by some groups. This allowed for the efficient extraction of metadata from sequence names. I used this metadata to determine the number of HIV-1M subtypes and CRFs per country in the LANL database and the number of sequences of each subtype per country. I plotted the number of subtypes and CRFs per country against the ethnic diversity for each country weighted by the number of sequences per country using ggplot2 version 2.2.1 (Wickham, 2009). The size of each dot was scaled logarithmically by the number of sequences. In order to determine statistical significance, I fit a generalised linear model in R version 3.4.3 (R Core Team, 2018) using a Poisson error distribution with a log link function:

$$\ln(\mu_y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots \qquad \text{Equation 3.2}$$

linking the natural log of the outcome ($\mu_y$) to a linear combination of the parameterised x variable $\beta_i x_i$. I chose this type of generalised linear model as the number of subtypes and CRFs per country is count data rather than normally distributed continuous data.

In order to create a more intuitive representation of this, I also generated a map with pie charts showing the relative amounts of each subtype in the LANL database for each country as well as the ethnic diversity for each country, using Tableau 10 (Ko and Chang, 2017).

## Graphics

Graphics editing and composition was conducted using Inkscape version 0.92 (Inkscape Project, 2017).

# Results & Discussion

## *Aims & Hypothesis*

**Question:** Is there a link between ethno-linguistic diversity and HIV-1 diversity?

**$H_0$:** Ethno-linguistic diversity and HIV diversity are unrelated

**$H_1$:** Higher HIV-1 diversity will be observed in countries with higher ethno-linguistic diversity.

## *HIV-1 Diversity Versus Ethnic Diversity*

The results of the linear regression between HIV-1 p24-Gag diversity and ethno-linguistic diversity within a country showed a statistically significant relationship

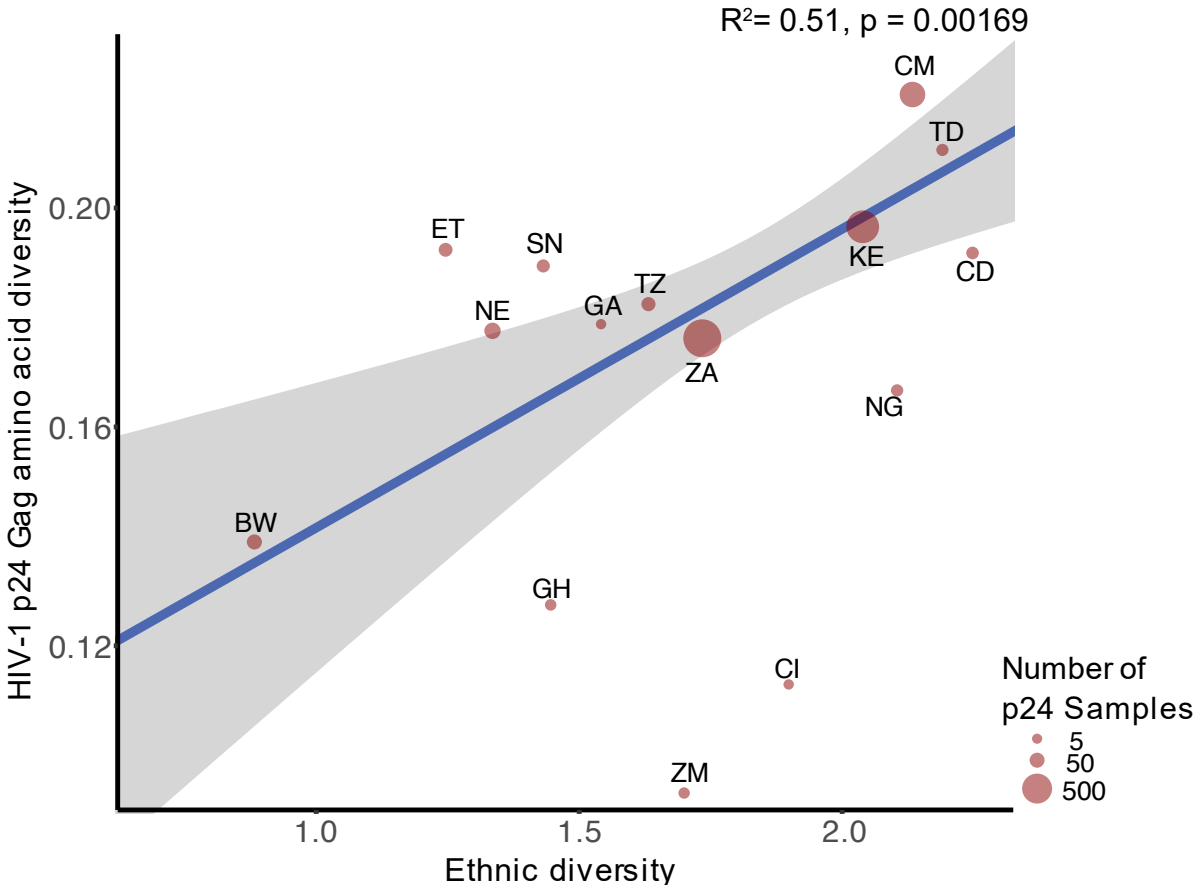



*Figure 3.2* Weighted linear regression of HIV-1 p24 Gag diversity versus ethnic diversity (each determined using Shannon entropy). All Sub-Saharan African countries with ≥ 5 p24 Gag sequences were included; Labels are ISO alpha-2 two-letter country codes and full names are available in appendix 1B. Grey shaded area represents the 95% confidence interval (CI).

($F_{1,13}$ = 15.53, $p$ < .005; Figure 3.2). When comparing subtype diversity and ethno-linguistic diversity within a country, I also found a statistically significant relationship ($p < 2 \times 10^{-16}$; Figure 3.3). Given these results, the null hypothesis, that the two are unrelated, can be rejected.

The DRC and Cameroon are both proposed origin points for HIV-1M, and as such higher diversity would be expected in them, potentially skewing the results. However, the results are robust, remaining statistically significant (HIV-1 p24 Gag diversity versus ethno-linguistic diversity ($p < 0.05$) and subtype diversity versus ethno-linguistic

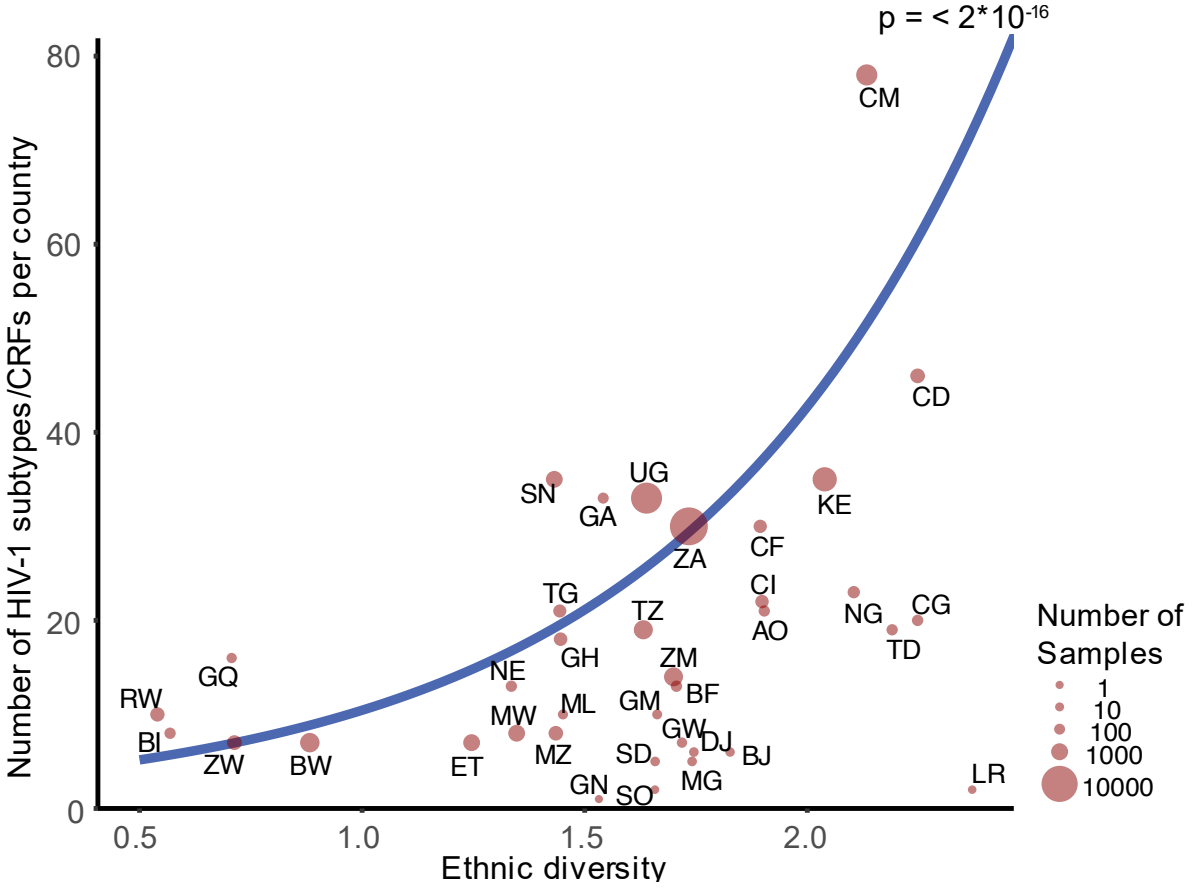



*Figure 3.3* Weighted linear regression of the number of HIV-1 subtypes and circulating recombinant forms (CRFs) versus ethnic diversity within Sub-Saharan Africa (illustrated in d). The linear regression used a generalised linear model fit with a Poisson error distribution with a log link function as the data were counts. Labels are ISO alpha-2 two-letter country codes and full names are available in appendix 1B. The 95% CI is too small to see.

diversity ($p < 0.5 \times 10^{-5}$)), even with both countries removed from the analyses. In light of this, it is difficult to support the idea that it is solely the timing of the epidemic in each African country that determines HIV-1 diversification. These results suggest that human genetic diversity, most likely in the form of HLA diversity, leads to the granular adaptation of HIV-1 to ethnic groups within Sub-Saharan Africa.

Five language families (Afro-Asiatic, Austronesian, Khoe, Niger-Congo, and Nilo-Saharan) comprise most Sub-Saharan African ethno-linguistic groups (Figure 3.4).

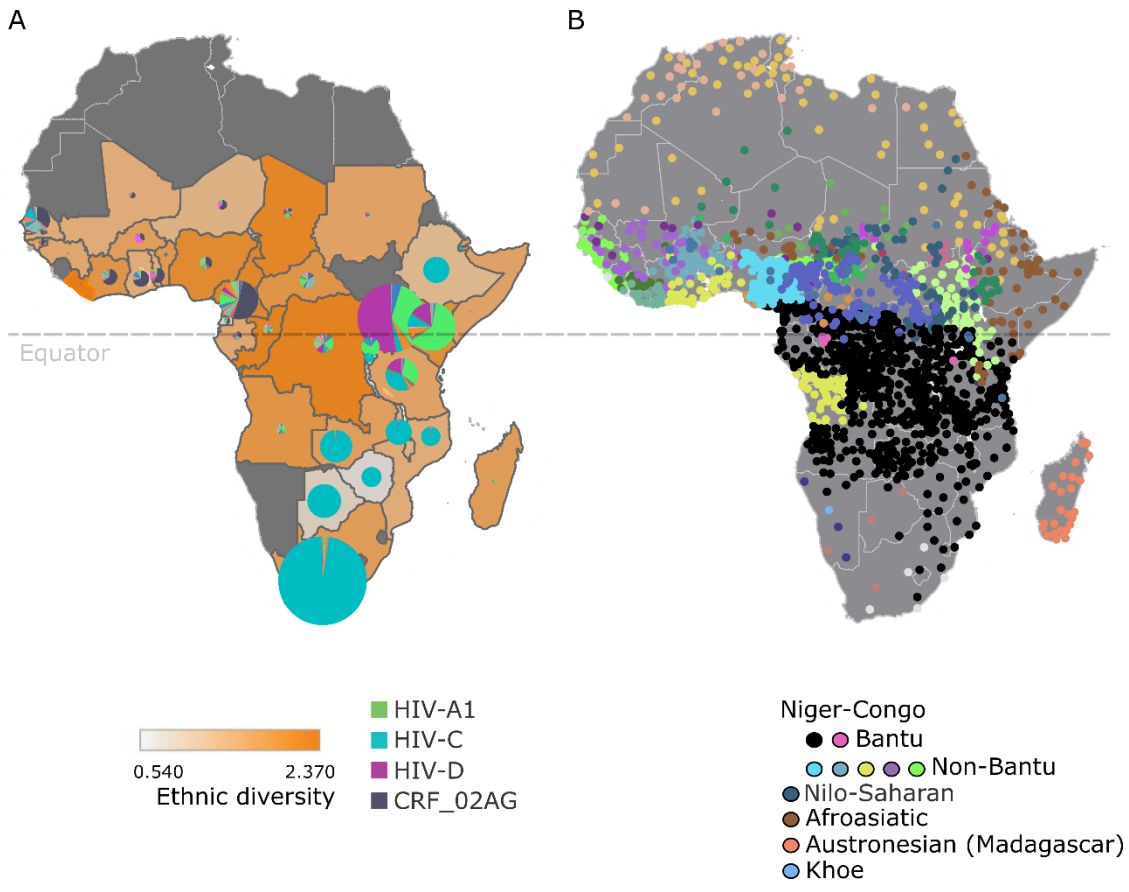



*Figure 3..4* **(A)** Map showing the ethnic diversity within Sub-Saharan Africa (orange shaded background) overlaid with pie charts demonstrating HIV-1 subtype diversity within each country. Missing countries (dark grey) lacked either HIV-1 or ethnic fractionalization data. The pie charts are scaled according to the number of unique patient HIV-1 sequences. **(B)** Map showing the current location of the five major language families within Africa with predominant groups indicated; complete key in appendix 1B.

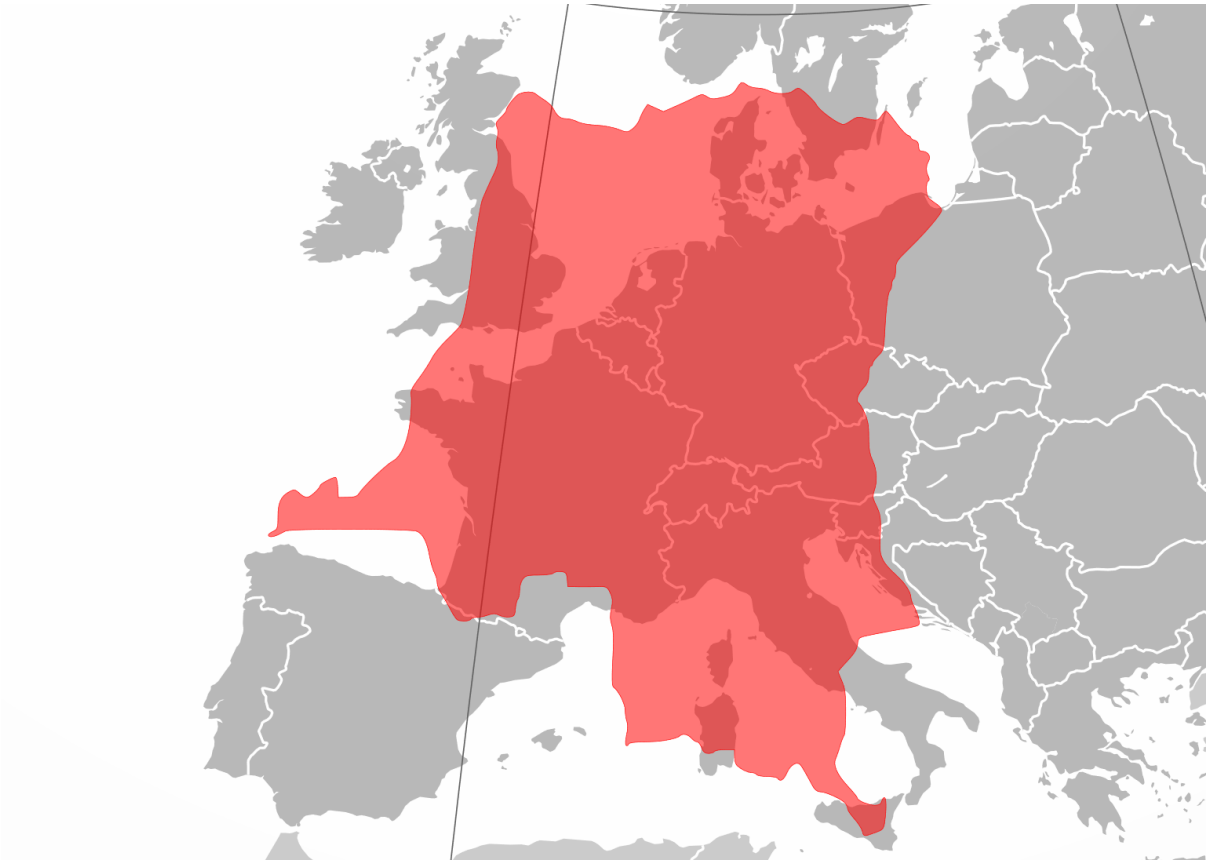


*Figure 3.5* Map of Western Europe with the Democratic Republic of the Congo to scale superimposed.

These can be further subdivided, for example the Niger-Congo family includes both Bantu languages, of which a third of all Sub-Saharan Africans are native speakers, and non-Bantu languages in West Africa. I found the highest concentration of both ethno-linguistic diversity and HIV-1 subtype diversity near the equator (Figure 3.4). A possible criticism of this is that some of the countries near the equator are geographically large, *e.g.* the DRC is roughly the size of Western Europe (Figure 3.5), and therefore are likely to contain a large number of ethno-linguistic groups. However, there does not appear to be any statistically significant link between country size and ethno-linguistic diversity in the Sub-Saharan African countries examined ($F_{1,25}$ = .23, $p$-value = .63).

I examined whether specific HIV-1 subtypes clustered within specific language families

(as ethno-linguistic groups are a viable proxy for human genetic diversity) and observed similar subtype patterns in related ethno-linguistic groups (Figure 3.4). For example, the regions in southern and south-eastern Africa with the largest percentages of Bantu speaking populations are also the areas where HIV subtype C (HIV-C) predominates. More HIV-1 diversity is seen in western and central Bantu speaking countries such as the DRC, which has 55 subtypes listed in LANL. Of these, HIV-A is the most common, but comprises only 27% of the 2703 listed sequences. This is contrasted by the similarly well sampled Zimbabwe where 98% of the 2982 sequences are HIV-C, and there are only seven subtypes present in LANL ("Los Alamos National Laboratory HIV Sequence Database," 2019). The specific history of the Bantu peoples may explain this discrepancy. ~4000-5000 years ago, Bantu speaking peoples expanded out from western central Africa (Patin et al., 2017). A climate crisis that fragmented the central African rainforest (first on the periphery ~4000 years ago and reaching the core ~2500 years ago) likely facilitated this dispersal (Bostoen et al., 2015). As the Bantu populations moved southwards they would have encountered local populations. Genetic analysis has indicated that ~800 years ago there was admixture of rainforest hunter-gatherers into western and central Bantu populations in the region (Patin et al., 2017). Strikingly, the level of rainforest hunter-gatherers ancestry in the HLA region is over six standard deviations higher than the level of ancestry in the rest of the genome in western and central Bantu populations (Patin et al., 2017). This makes the HLA variability of western and central Bantu populations markedly distinct and more diverse than their southern and south eastern counterparts. Non-Bantu languages from

the Niger-Congo family are predominant in West Africa (Figure 3.4). HIV-C is rarer in these regions, and HIV-G and the recombinant CRF-02AG are dominant in the region. HIV-A, HIV-C, and HIV-D are frequently found in East Africa where Nilo-Saharan, Afro-Asiatic, and Bantu languages are spoken.

This is not to say that founder effects can be totally disregarded. HIV-C is the predominant strain in Ethiopia, with 99.5% of the sequences in LANL being HIV-C ("Los Alamos National Laboratory HIV Sequence Database," 2019) and the actual epidemic closely reflecting this distribution (Kalu et al., 2017). The epidemic there is believed to have started in the early 1980s (Pollakis et al., 2003a). However, when examining the genomes of HIV-C in Ethiopia, they show strong geographical clustering, distinct from southern HIV-C strains, and the HIV-C in Ethiopia has diverged extensively from the earliest published near full length Ethiopian strain from 1986 (Abebe et al., 2000; Amogne et al., 2016). Furthermore, the HIV-C in Ethiopia has longer phylogenetic branch lengths and has significantly higher diversity than the HIV-C seen in southern Africa (Amogne et al., 2016). This is in line with my results and would be expected for a country like Ethiopia where a virus more adapted to southern Bantu HLA profiles would be only relatively recently coming into contact with a mixture of Cushitic, Omotic, and Semitic peoples (all from the Afro-Asiatic family) as well as Nilo-Saharan populations.

These results suggest that human HLA diversity helps drive HIV-1 subtype

diversification by exerting varied HLA pressures on the virus as it encounters new host populations. Although HIV-1 mutates and recombines rapidly, viruses still may take time to adapt to the host population. When Cameroon, the Congo, and the DRC, the epicentre of the HIV-1 pandemic achieved independence from France and Belgium in 1960, there were no resulting HIV-1 epidemics in Europe brought back by returning colonialists, nor in the preceding decades of colonial rule. Similarly, in the Americas, it wasn't until around 1966 (1962-1970) that an ancestor strain to HIV-B is believed to have been brought to Haiti from the DRC, and then been transferred to the United States around 1969 (1966–1972) (Gilbert et al., 2007). There was presumably ample exposure to any HIV-1 circulating in Haiti, as the country had 100,000 tourists per year by the late 1970s, 70,000 of them American, and Haiti's rapidly growing sex industry took advantage of this (Girard, 2005 pg. 101). Although in retrospect, there were signs of the virus circulating in the US, primarily in marginalised populations such as homeless intravenous drug users in New York dying from an epidemic of "junkie pneumonia" and "the dwindles", HIV-1 circulated largely cryptically for ~12 years in the US before the epidemic reached a level high enough to be detected (Gilbert et al., 2007).

If African HIV-1 strains were less well adapted to European HLA profiles they would be less readily transmissible to people of European descent, which might at least partially explain the lag. That HLAs can affect acquisition rates on an individual level, has already been established. HLA-A*68:02 and the B*42-C*17 haplotype have been

shown to be associated with faster acquisition of HIV-1 in HIV-1-exposed seronegative partners of seropositive patients, even after controlling for other factors (Song et al., 2011). It is not hard to imagine that similar factors could be playing out at the population level.

These results suggest several possible avenues for further research. One is looking at whether we can see adaptation of a subtype recently introduced to a new a population, or one where the ethnic demographics of a localised epidemic are in the process of changing. Indeed, this is something we saw in HIV-B becoming more HIV-C like in the US as African Americans become an increasingly large percentage of the US epidemic (see published paper by Kist et al. (2020) in appendix 1C). The Ethiopian HIV-C epidemic has distinct clades that are apparently unrelated to geographic location, risk groups (e.g. IV drug users, commercial sex workers), or time of collection (Abebe et al., 2000). It would be interesting to see if the clades are associated with ethno-linguistic groups within Ethiopia. Additionally, these results have implications for HIV-1 vaccine design. Rather than focusing on the HIV-1 subtype, the starting point should be examining the HLA composition of the target population itself. In this context, epitopes, preferably from more constrained proteins such as Gag, could be targeted. Distinct groups of amino acids in Gag with linked mutations and limited escape pathways (if targeted in parallel) have been identified as potential targets of immunological vulnerability (Dahirel et al., 2011a).

Using this information, epitopes that can be presented by the host population's HLAs and that are processed in small amounts can be identified. While these epitopes may not be presented enough naturally to prime the host immune system, they can be presented as part of a vaccine construct in order to artificially prime it. Once primed, the host immune system will be able to recognise and respond to the much more limited presentation occurring naturally. If epitope selection has been carried out effectively, the virus should have difficultly utilizing escape mutations without incurring catastrophic fitness costs. Such epitope focused vaccine design has been demonstrated in proof of principle experiments for respiratory syncytial virus (Correia et al., 2014) and has been suggested as a possible avenue forward for highly antigenically variable pathogens such as HIV-1 and influenza.

# Works Cited


Abebe, A., Pollakis, G., Fontanet, A.L., Fisseha, B., Tegbaru, B., Kliphuis, A., Tesfaye, G., Negassa, H., Cornelissen, M., Goudsmit, J., Rinke de Wit, T.F., 2000. Identification of a genetic subcluster of HIV type 1 subtype C (C') widespread in Ethiopia. AIDS Res. Hum. Retroviruses 16, 1909–1914. https://doi.org/10.1089/08892220050195865

Abecasis, A., Vandamme, A.-M., 2015. Origin and Distribution of HIV-1 Subtypes, in: Hope, T.J., Stevenson, M., Richman, D. (Eds.), Encyclopedia of AIDS. Springer New York, New York, NY, pp. 1–16. https://doi.org/10.1007/978-1-4614-9610-6_130-2

Abecasis, A.B., Lemey, P., Vidal, N., de Oliveira, T., Peeters, M., Camacho, R., Shapiro, B., Rambaut, A., Vandamme, A.-M., 2007. Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form. J. Virol. 81, 8543–8551. https://doi.org/10.1128/JVI.00463-07

Abram, M.E., Ferris, A.L., Shao, W., Alvord, W.G., Hughes, S.H., 2010. Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. J. Virol. 84, 9864–9878. https://doi.org/10.1128/JVI.00915-10

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., Wacziarg, R., 2003. Fractionalization. J. Econ. Growth 8, 155–94.

Amogne, W., Bontell, I., Grossmann, S., Aderaye, G., Lindquist, L., Sönnerborg, A., Neogi, U., 2016. Phylogenetic Analysis of Ethiopian HIV-1 Subtype C Near Full-Length Genomes Reveals High Intrasubtype Diversity and a Strong Geographical Cluster. AIDS Res. Hum. Retroviruses 32, 471–474. https://doi.org/10.1089/aid.2015.0380

Anastassopoulou, C., Kostrikis, L., 2006. Global Genetic Variation of HIV-1 Infection. Curr. HIV Res. 4, 365–373. https://doi.org/10.2174/157016206777709456

Antiretroviral Therapy Cohort Collaboration, 2008. Life expectancy of individuals on combination antiretroviral therapy in high-income countries: a collaborative analysis of 14 cohort studies. Lancet Lond. Engl. 372, 293–299. https://doi.org/10.1016/S0140-6736(08)61113-7

Archer, J., Robertson, D.L., 2007a. Understanding the diversification of HIV-1 groups M and O: AIDS 21, 1693–1700. https://doi.org/10.1097/QAD.0b013e32825eabd0

Archer, J., Robertson, D.L., 2007b. Understanding the diversification of HIV-1 groups M and O: AIDS 21, 1693–1700. https://doi.org/10.1097/QAD.0b013e32825eabd0

Ariën, K.K., Abraha, A., Quiñones-Mateu, M.E., Kestens, L., Vanham, G., Arts, E.J., 2005. The replicative fitness of primary human immunodeficiency virus type 1 (HIV-1) group M, HIV-1 group O, and HIV-2 isolates. J. Virol. 79, 8979–8990. https://doi.org/10.1128/JVI.79.14.8979-8990.2005

Arrildt, K.T., Joseph, S.B., Swanstrom, R., 2012. The HIV-1 env protein: a coat of many colors. Curr. HIV/AIDS Rep. 9, 52–63. https://doi.org/10.1007/s11904-011-0107-3

Baeten, J.M., Chohan, B., Lavreys, L., Chohan, V., McClelland, R.S., Certain, L., Mandaliya, K., Jaoko, W., Overbaugh, J., 2007. HIV-1 subtype D infection is associated with faster disease progression than subtype A in spite of similar

plasma HIV-1 loads. J. Infect. Dis. 195, 1177–1180. https://doi.org/10.1086/512682
Ball, S.C., Abraha, A., Collins, K.R., Marozsan, A.J., Baird, H., Quiñones-Mateu, M.E., Penn-Nicholson, A., Murray, M., Richard, N., Lobritz, M., Zimmerman, P.A., Kawamura, T., Blauvelt, A., Arts, E.J., 2003. Comparing the ex vivo fitness of CCR5-tropic human immunodeficiency virus type 1 isolates of subtypes B and C. J. Virol. 77, 1021–1038. https://doi.org/10.1128/jvi.77.2.1021-1038.2003
Barré-Sinoussi, F., Chermann, J.C., Rey, F., Nugeyre, M.T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vézinet-Brun, F., Rouzioux, C., Rozenbaum, W., Montagnier, L., 1983. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science 220, 868–871. https://doi.org/10.1126/science.6189183
Basu, V.P., Song, M., Gao, L., Rigby, S.T., Hanson, M.N., Bambara, R.A., 2008. Strand transfer events during HIV-1 reverse transcription. Virus Res. 134, 19–38. https://doi.org/10.1016/j.virusres.2007.12.017
Bostoen, K., Clist, B., Doumenge, C., Grollemund, R., Hombert, J.-M., Muluwa, J.K., Maley, J., 2015. Middle to Late Holocene Paleoclimatic Change and the Early Bantu Expansion in the Rain Forests of Western Central Africa. Curr. Anthropol. 56, 354–384. https://doi.org/10.1086/681436
Boutwell, C.L., Rowley, C.F., Essex, M., 2009. Reduced viral replication capacity of human immunodeficiency virus type 1 subtype C caused by cytotoxic-T-lymphocyte escape mutations in HLA-B57 epitopes of capsid protein. J. Virol. 83, 2460–2468. https://doi.org/10.1128/JVI.01970-08
Bulla, I., Schultz, A.-K., Schreiber, F., Zhang, M., Leitner, T., Korber, B., Morgenstern, B., Stanke, M., 2010. HIV classification using the coalescent theory. Bioinformatics 26, 1409–1415. https://doi.org/10.1093/bioinformatics/btq159
Bunnell, R.E., Yanpaisarn, S., Kilmarx, P.H., Rhodes, P.H., Limpakarnjanarat, K., Srismith, R., Mastro, T.D., St Louis, M.E., 1999. HIV-1 seroprevalence among childbearing women in northern Thailand: monitoring a rapidly evolving epidemic. AIDS Lond. Engl. 13, 509–515. https://doi.org/10.1097/00002030-199903110-00010
Buonaguro, L., Tornesello, M.L., Buonaguro, F.M., 2007. Human Immunodeficiency Virus Type 1 Subtype Distribution in the Worldwide Epidemic: Pathogenetic and Therapeutic Implications. J. Virol. 81, 10209–10219. https://doi.org/10.1128/JVI.00872-07
Burchell, A.N., Raboud, J., Donelle, J., Loutfy, M.R., Rourke, S.B., Rogers, T., Rosenes, R., Liddy, C., Kendall, C.E., 2019. Cause-specific mortality among HIV-infected people in Ontario, 1995-2014: a population-based retrospective cohort study. CMAJ Open 7, E1–E7. https://doi.org/10.9778/cmajo.20180159
Burke, D., 1997. Recombination in HIV: An Important Viral Evolutionary Strategy. Emerg. Infect. Dis. 3, 253–259. https://doi.org/10.3201/eid0303.970301
Burnie, J., Guzzo, C., 2019. The Incorporation of Host Proteins into the External HIV-1 Envelope. Viruses 11. https://doi.org/10.3390/v11010085
Campbell-Yesufu, Omobolaji T., Gandhi, R.T., 2011. Update on human immunodeficiency virus (HIV)-2 infection. Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am. 52, 780–787. https://doi.org/10.1093/cid/ciq248
Campbell-Yesufu, O. T., Gandhi, R.T., 2011. Update on Human Immunodeficiency Virus (HIV)-2 Infection. Clin. Infect. Dis. 52, 780–787. https://doi.org/10.1093/cid/ciq248

Cao, K., Moormann, A.M., Lyke, K.E., Masaberg, C., Sumba, O.P., Doumbo, O.K., Koech, D., Lancaster, A., Nelson, M., Meyer, D., Single, R., Hartzman, R.J., Plowe, C.V., Kazura, J., Mann, D.L., Sztein, M.B., Thomson, G., Fernández-Viña, M.A., 2004. Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci. Tissue Antigens 63, 293–325.

Case Western Reserve University Institutional Research, 2020. Diversity | Institutional Research | Case Western Reserve University [WWW Document]. URL https://case.edu/ir/cwru-facts/diversity (accessed 12.18.20).

Center for Geographic Analysis, 2013. Ethnicity Felix 2001 [WWW Document]. URL https://worldmap.harvard.edu/africamap/

Chen, R.X.F., 2018. A Brief Introduction to Shannon's Information Theory. ArXiv161209316v2 Cs Math.

Chen, Z., DeSalle, R., Schiffman, M., Herrero, R., Wood, C.E., Ruiz, J.C., Clifford, G.M., Chan, P.K.S., Burk, R.D., 2018. Niche adaptation and viral transmission of human papillomaviruses from archaic hominins to modern humans. PLOS Pathog. 14, e1007352. https://doi.org/10.1371/journal.ppat.1007352

Congo Basin Art History Research Center, Meur, C., Felix, M.L., Tribal Arts S.P.R.L, 2001. Peoples of Africa: ethno-linguistic map.

Cordingley, M.G., 2017. Viruses: Agents of Evolutionary Invention. Harvard University Press.

Correia, B.E., Bates, J.T., Loomis, R.J., Baneyx, G., Carrico, C., Jardine, J.G., Rupert, P., Correnti, C., Kalyuzhniy, O., Vittal, V., Connell, M.J., Stevens, E., Schroeter, A., Chen, M., Macpherson, S., Serra, A.M., Adachi, Y., Holmes, M.A., Li, Y., Klevit, R.E., Graham, B.S., Wyatt, R.T., Baker, D., Strong, R.K., Crowe, J.E., Johnson, P.R., Schief, W.R., 2014. Proof of principle for epitope-focused vaccine design. Nature 507, 201–206. https://doi.org/10.1038/nature12966

Crawford, H., Matthews, P.C., Schaefer, M., Carlson, J.M., Leslie, A., Kilembe, W., Allen, S., Ndung'u, T., Heckerman, D., Hunter, E., Goulder, P.J.R., 2011. The hypervariable HIV-1 capsid protein residues comprise HLA-driven CD8+ T-cell escape mutations and covarying HLA-independent polymorphisms. J. Virol. 85, 1384–1390. https://doi.org/10.1128/JVI.01879-10

Cromer, D., Schlub, T.E., Smyth, R.P., Grimm, A.J., Chopra, A., Mallal, S., Davenport, M.P., Mak, J., 2016. HIV-1 Mutation and Recombination Rates Are Different in Macrophages and T-cells. Viruses 8, 118. https://doi.org/10.3390/v8040118

Dahirel, V., Shekhar, K., Pereyra, F., Miura, T., Artyomov, M., Talsania, S., Allen, T.M., Altfeld, M., Carrington, M., Irvine, D.J., Walker, B.D., Chakraborty, A.K., 2011a. Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. Proc. Natl. Acad. Sci. U. S. A. 108, 11530–11535. https://doi.org/10.1073/pnas.1105315108

Dahirel, V., Shekhar, K., Pereyra, F., Miura, T., Artyomov, M., Talsania, S., Allen, T.M., Altfeld, M., Carrington, M., Irvine, D.J., Walker, B.D., Chakraborty, A.K., 2011b. Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. Proc. Natl. Acad. Sci. U. S. A. 108, 11530–11535. https://doi.org/10.1073/pnas.1105315108

De Cock, K.M., 1993. Epidemiology and Transmission of HIV-2: Why There Is No HIV-2 Pandemic. JAMA 270, 2083. https://doi.org/10.1001/jama.1993.03510170073033

de Silva, T.I., Cotten, M., Rowland-Jones, S.L., 2008. HIV-2: the forgotten AIDS virus. Trends Microbiol. 16, 588–595. https://doi.org/10.1016/j.tim.2008.09.003

Désiré, N., Cerutti, L., Le Hingrat, Q., Perrier, M., Emler, S., Calvez, V., Descamps, D., Marcelin, A.-G., Hué, S., Visseaux, B., 2018. Characterization update of HIV-1 M subtypes diversity and proposal for subtypes A and D sub-subtypes reclassification. Retrovirology 15. https://doi.org/10.1186/s12977-018-0461-y

Duke-Sylvester, S.M., Biek, R., Real, L.A., 2013. Molecular evolutionary signatures reveal the role of host ecological dynamics in viral disease emergence and spread. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 368, 20120194. https://doi.org/10.1098/rstb.2012.0194

Esbjörnsson, J., Månsson, F., Kvist, A., Isberg, P.-E., Nowroozalizadeh, S., Biague, A.J., da Silva, Z.J., Jansson, M., Fenyö, E.M., Norrgren, H., Medstrand, P., 2012. Inhibition of HIV-1 disease progression by contemporaneous HIV-2 infection. N. Engl. J. Med. 367, 224–232. https://doi.org/10.1056/NEJMoa1113244

Esparza, J., 2013. A brief history of the global effort to develop a preventive HIV vaccine. Vaccine 31, 3502–3518. https://doi.org/10.1016/j.vaccine.2013.05.018

Excoffier, L., Harding, R.M., Sokal, R.R., Pellegrini, B., Sanchez-Mazas, A., 1991a. Spatial differentiation of RH and GM haplotype frequencies in Sub-Saharan Africa and its relation to linguistic affinities. Hum. Biol. 63, 273–307.

Excoffier, L., Harding, R.M., Sokal, R.R., Pellegrini, B., Sanchez-Mazas, A., 1991b. Spatial differentiation of RH and GM haplotype frequencies in Sub-Saharan Africa and its relation to linguistic affinities. Hum. Biol. 63, 273–307.

Excoffier, L., Pellegrini, B., Sanchez-Mazas, A., Simon, C., Langaney, A., 1987a. Genetics and history of sub-Saharan Africa. Am. J. Phys. Anthropol. 30, 151–194. https://doi.org/10.1002/ajpa.1330300510

Excoffier, L., Pellegrini, B., Sanchez-Mazas, A., Simon, C., Langaney, A., 1987b. Genetics and history of sub-Saharan Africa. Am. J. Phys. Anthropol. 30, 151–194. https://doi.org/10.1002/ajpa.1330300510

Faria, N.R., Rambaut, A., Suchard, M.A., Baele, G., Bedford, T., Ward, M.J., Tatem, A.J., Sousa, J.D., Arinaminpathy, N., Pépin, J., Posada, D., Peeters, M., Pybus, O.G., Lemey, P., 2014. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. Science 346, 56–61. https://doi.org/10.1126/science.1256739

Faria, N.R., Vidal, N., Lourenco, J., Raghwani, J., Sigaloff, K.C.E., Tatem, A.J., van de Vijver, D.A.M., Pineda-Peña, A.-C., Rose, R., Wallis, C.L., Ahuka-Mundeke, S., Muyembe-Tamfum, J.-J., Muwonga, J., Suchard, M.A., Rinke de Wit, T.F., Hamers, R.L., Ndembi, N., Baele, G., Peeters, M., Pybus, O.G., Lemey, P., Dellicour, S., 2019. Distinct rates and patterns of spread of the major HIV-1 subtypes in Central and East Africa. PLOS Pathog. 15, e1007976. https://doi.org/10.1371/journal.ppat.1007976

Fauci, A.S., Folkers, G.K., Marston, H.D., 2014a. Ending the global HIV/AIDS pandemic: the critical role of an HIV vaccine. Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am. 59 Suppl 2, S80-84. https://doi.org/10.1093/cid/ciu420

Fauci, A.S., Marston, H.D., Folkers, G.K., 2014b. An HIV cure: feasibility, discovery, and implementation. JAMA 312, 335–336. https://doi.org/10.1001/jama.2014.4754

Fearon, J.D., 2003. Ethnic and Cultural Diversity by Country. J. Econ. Growth 8, 195–222.

Felsenstein, J., Yokoyama, S., 1976. The evolutionary advantage of recombination. II. Individual selection for recombination. Genetics 83, 845–859.
Ferdinandy, B., Mones, E., Vicsek, T., Müller, V., 2015. HIV Competition Dynamics over Sexual Networks: First Comer Advantage Conserves Founder Effects. PLOS Comput. Biol. 11, e1004093. https://doi.org/10.1371/journal.pcbi.1004093
Fischetti, L., Opare-Sem, O., Candotti, D., Lee, H., Allain, J.-P., 2004. Higher viral load may explain the dominance of CRF02_AG in the molecular epidemiology of HIV in Ghana. AIDS Lond. Engl. 18, 1208–1210. https://doi.org/10.1097/00002030-200405210-00017
Foy, H.M., Kunanusont, C., Kreiss, J.K., Phanuphak, P., Raktham, S., Pau, C.-P., Young, N.L., Rerks-Ngarm, S., 1995. HIV-1 subtypes and male-to-female transmission in Thailand. The Lancet 345, 1078–1083. https://doi.org/10.1016/S0140-6736(95)90818-8
Galetto, R., Negroni, M., 2005. Mechanistic features of recombination in HIV. AIDS Rev. 7, 92–102.
Gallo, R.C., Montagnier, L., 2003. The Discovery of HIV as the Cause of AIDS. N. Engl. J. Med. 349, 2283–2285. https://doi.org/10.1056/NEJMp038194
Gilbert, M.T.P., Rambaut, A., Wlasiuk, G., Spira, T.J., Pitchenik, A.E., Worobey, M., 2007. The emergence of HIV/AIDS in the Americas and beyond. Proc. Natl. Acad. Sci. 104, 18566–18570. https://doi.org/10.1073/pnas.0705329104
Girard, P.R., 2005. Paradise Lost. Palgrave Macmillan US, New York. https://doi.org/10.1057/9781403980311
González-Galarza, F.F., Takeshita, L.Y.C., Santos, E.J.M., Kempson, F., Maia, M.H.T., da Silva, A.L.S., Teles e Silva, A.L., Ghattaoraya, G.S., Alfirevic, A., Jones, A.R., Middleton, D., 2015. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. Nucleic Acids Res. 43, D784-788. https://doi.org/10.1093/nar/gku1166
Grant, B.J., Rodrigues, A.P.C., ElSawy, K.M., McCammon, J.A., Caves, L.S.D., 2006. Bio3d: an R package for the comparative analysis of protein structures. Bioinforma. Oxf. Engl. 22, 2695–2696. https://doi.org/10.1093/bioinformatics/btl461
Grant, H.E., Hodcroft, E.B., Ssemwanga, D., Kitayimbwa, J.M., Yebra, G., Esquivel Gomez, L.R., Frampton, D., Gall, A., Kellam, P., de Oliveira, T., Bbosa, N., Nsubuga, R.N., Kibengo, F., Kwan, T.H., Lycett, S., Kao, R., Robertson, D.L., Ratmann, O., Fraser, C., Pillay, D., Kaleebu, P., Leigh Brown, A.J., 2020. Pervasive and non-random recombination in near full-length HIV genomes from Uganda. Virus Evol. 6, veaa004. https://doi.org/10.1093/ve/veaa004
Halper-Stromberg, A., Nussenzweig, M.C., 2016. Towards HIV-1 remission: potential roles for broadly neutralizing antibodies. J. Clin. Invest. 126, 415–423. https://doi.org/10.1172/JCI80561
Headrick, D.R., 2014. Sleeping sickness epidemics and colonial responses in East and Central Africa, 1900-1940. PLoS Negl. Trop. Dis. 8, e2772. https://doi.org/10.1371/journal.pntd.0002772
Hemelaar, J., 2012. The origin and diversity of the HIV-1 pandemic. Trends Mol. Med. 18, 182–192. https://doi.org/10.1016/j.molmed.2011.12.001
Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Fleminger, I., Kirtley, S., Williams, B., Gouws-Williams, E., Ghys, P.D., WHO–UNAIDS Network for HIV Isolation Characterisation, 2019. Global and regional molecular epidemiology of HIV-1, 1990-2015: a systematic review, global survey, and

trend analysis. Lancet Infect. Dis. 19, 143–155. https://doi.org/10.1016/S1473-3099(18)30647-9
Hemelaar, J., Gouws, E., Ghys, P.D., Osmanov, S., 2006. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. AIDS Lond. Engl. 20, W13-23. https://doi.org/10.1097/01.aids.0000247564.73009.bc
Ho, L., Chan, S.Y., Burk, R.D., Das, B.C., Fujinaga, K., Icenogle, J.P., Kahn, T., Kiviat, N., Lancaster, W., Mavromara-Nazos, P., 1993. The genetic drift of human papillomavirus type 16 is a means of reconstructing prehistoric viral spread and the movement of ancient human populations. J. Virol. 67, 6413–6423. https://doi.org/10.1128/JVI.67.11.6413-6423.1993
Holmes, E.C., 2009. RNA virus genomics: a world of possibilities. J. Clin. Invest. 119, 2488–2495. https://doi.org/10.1172/JCI38050
Holmes, E.C., 2004. The phylogeography of human viruses. Mol. Ecol. 13, 745–756. https://doi.org/10.1046/j.1365-294X.2003.02051.x
Holmes, E.C., Nee, S., Rambaut, A., Garnett, G.P., Harvey, P.H., 1995. Revealing the history of infectious disease epidemics through phylogenetic trees. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 349, 33–40. https://doi.org/10.1098/rstb.1995.0088
Honeyborne, I., Prendergast, A., Pereyra, F., Leslie, A., Crawford, H., Payne, R., Reddy, S., Bishop, K., Moodley, E., Nair, K., van der Stok, M., McCarthy, N., Rousseau, C.M., Addo, M., Mullins, J.I., Brander, C., Kiepiela, P., Walker, B.D., Goulder, P.J.R., 2007. Control of human immunodeficiency virus type 1 is associated with HLA-B*13 and targeting of multiple gag-specific CD8+ T-cell epitopes. J. Virol. 81, 3667–3672. https://doi.org/10.1128/JVI.02689-06
Hu, D.J., Buvé, A., Baggs, J., van der Groen, G., Dondero, T.J., 1999. What role does HIV-1 subtype play in transmission and pathogenesis? An epidemiological perspective: AIDS 13, 873–881. https://doi.org/10.1097/00002030-199905280-00002
Hudgens, M.G., Longini, I.M., Vanichseni, S., Hu, D.J., Kitayaporn, D., Mock, P.A., Halloran, M.E., Satten, G.A., Choopanya, K., Mastro, T.D., 2002. Subtype-specific transmission probabilities for human immunodeficiency virus type 1 among injecting drug users in Bangkok, Thailand. Am. J. Epidemiol. 155, 159–168. https://doi.org/10.1093/aje/155.2.159
Iacob, S.A., Iacob, D.G., Jugulete, G., 2017. Improving the Adherence to Antiretroviral Therapy, a Difficult but Essential Task for a Successful HIV Treatment-Clinical Points of View and Practical Considerations. Front. Pharmacol. 8, 831. https://doi.org/10.3389/fphar.2017.00831
Ingole, N.A., Sarkate, P.P., Paranjpe, S.M., Shinde, S.D., Lall, S.S., Mehta, P.R., 2013. HIV-2 Infection: Where Are We Today? J. Glob. Infect. Dis. 5, 110–113. https://doi.org/10.4103/0974-777X.116872
Inkscape Project, 2017. Inkscape.
Junqueira, D.M., Almeida, S.E. de M., 2016. HIV-1 subtype B: Traces of a pandemic. Virology 495, 173–184. https://doi.org/10.1016/j.virol.2016.05.003
Kaleebu, P., French, N., Mahe, C., Yirrell, D., Watera, C., Lyagoba, F., Nakiyingi, J., Rutebemberwa, A., Morgan, D., Weber, J., Gilks, C., Whitworth, J., 2002. Effect of human immunodeficiency virus (HIV) type 1 envelope subtypes A and D on disease progression in a large cohort of HIV-1-positive persons in Uganda. J. Infect. Dis. 185, 1244–1250. https://doi.org/10.1086/340130
Kalish, M.L., Luo, C.C., Weniger, B.G., Limpakarnjanarat, K., Young, N., Ou, C.Y., Schochetman, G., 1994. Early HIV type 1 strains in Thailand were not

responsible for the current epidemic. AIDS Res. Hum. Retroviruses 10, 1573–1575. https://doi.org/10.1089/aid.1994.10.1573

Kalish, M.L., Robbins, K.E., Pieniazek, D., Schaefer, A., Nzilambi, N., Quinn, T.C., St Louis, M.E., Youngpairoj, A.S., Phillips, J., Jaffe, H.W., Folks, T.M., 2004. Recombinant viruses and early global HIV-1 epidemic. Emerg. Infect. Dis. 10, 1227–1234. https://doi.org/10.3201/eid1007.030904

Kalish, M.L., Wolfe, N.D., Ndongmo, C.B., McNicholl, J., Robbins, K.E., Aidoo, M., Fonjungo, P.N., Alemnji, G., Zeh, C., Djoko, C.F., Mpoudi-Ngole, E., Burke, D.S., Folks, T.M., 2005. Central African hunters exposed to simian immunodeficiency virus. Emerg. Infect. Dis. 11, 1928–1930. https://doi.org/10.3201/eid1112.050394

Kalu, A.W., Telele, N.F., Gebreselasie, S., Fekade, D., Abdurahman, S., Marrone, G., Sönnerborg, A., 2017. Monophylogenetic HIV-1C epidemic in Ethiopia is dominated by CCR5-tropic viruses-an analysis of a prospective country-wide cohort. BMC Infect. Dis. 17, 37. https://doi.org/10.1186/s12879-016-2163-1

Keele, B.F., Van Heuverswyn, F., Li, Y., Bailes, E., Takehisa, J., Santiago, M.L., Bibollet-Ruche, F., Chen, Y., Wain, L.V., Liegeois, F., Loul, S., Ngole, E.M., Bienvenue, Y., Delaporte, E., Brookfield, J.F.Y., Sharp, P.M., Shaw, G.M., Peeters, M., Hahn, B.H., 2006. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. Science 313, 523–526. https://doi.org/10.1126/science.1126531

Kiepiela, P., Ngumbela, K., Thobakgale, C., Ramduth, D., Honeyborne, I., Moodley, E., Reddy, S., de Pierres, C., Mncube, Z., Mkhwanazi, N., Bishop, K., van der Stok, M., Nair, K., Khan, N., Crawford, H., Payne, R., Leslie, A., Prado, J., Prendergast, A., Frater, J., McCarthy, N., Brander, C., Learn, G.H., Nickle, D., Rousseau, C., Coovadia, H., Mullins, J.I., Heckerman, D., Walker, B.D., Goulder, P., 2007. CD8+ T-cell responses to different HIV proteins have discordant associations with viral load. Nat. Med. 13, 46–53. https://doi.org/10.1038/nm1520

Kist, N.C., Lambert, B., Campbell, S., Katzourakis, A., Lunn, D., Lemey, P., Iversen, A.K.N., 2020. HIV-1 p24Gag adaptation to modern and archaic HLA-allele frequency differences in ethnic groups contributes to viral subtype diversification. Virus Evol. https://doi.org/10.1093/ve/veaa085

Kiwanuka, N., Laeyendecker, O., Quinn, T.C., Wawer, M.J., Shepherd, J., Robb, M., Kigozi, G., Kagaayi, J., Serwadda, D., Makumbi, F.E., Reynolds, S.J., Gray, R.H., 2009. HIV-1 subtypes and differences in heterosexual HIV transmission among HIV-discordant couples in Rakai, Uganda. AIDS Lond. Engl. 23, 2479–2484. https://doi.org/10.1097/QAD.0b013e328330cc08

Klein, F., Halper-Stromberg, A., Horwitz, J.A., Gruell, H., Scheid, J.F., Bournazos, S., Mouquet, H., Spatz, L.A., Diskin, R., Abadir, A., Zang, T., Dorner, M., Billerbeck, E., Labitt, R.N., Gaebler, C., Marcovecchio, P., Incesu, R.-B., Eisenreich, T.R., Bieniasz, P.D., Seaman, M.S., Bjorkman, P.J., Ravetch, J.V., Ploss, A., Nussenzweig, M.C., 2012. HIV therapy by a combination of broadly neutralizing antibodies in humanized mice. Nature 492, 118–122. https://doi.org/10.1038/nature11604

Ko, I., Chang, H., 2017. Interactive Visualization of Healthcare Data Using Tableau. Healthc. Inform. Res. 23, 349–354. https://doi.org/10.4258/hir.2017.23.4.349

Korber, B., 2000. Timing the Ancestor of the HIV-1 Pandemic Strains. Science 288, 1789–1796. https://doi.org/10.1126/science.288.5472.1789

Korber, B., Gaschen, B., Yusim, K., Thakallapally, R., Kesmir, C., Detours, V., 2001. Evolutionary and immunological implications of contemporary HIV-1 variation. Br. Med. Bull. 58, 19–42. https://doi.org/10.1093/bmb/58.1.19

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B.H., Wolinsky, S., Bhattacharya, T., 2000. Timing the ancestor of the HIV-1 pandemic strains. Science 288, 1789–1796. https://doi.org/10.1126/science.288.5472.1789

Kuiken, C., 2000. Genetic Analysis Reveals Epidemiologic Patterns in the Spread of Human Immunodeficiency Virus. Am. J. Epidemiol. 152, 814–822. https://doi.org/10.1093/aje/152.9.814

Kumar, R., Qureshi, H., Deshpande, S., Bhattacharya, J., 2018. Broadly neutralizing antibodies in HIV-1 treatment and prevention. Ther. Adv. Vaccines Immunother. 6, 61–68. https://doi.org/10.1177/2515135518800689

Lau, K.A., Wang, B., Miranda-Saksena, M., Boadle, R., Kamarulzaman, A., Ng, K.-P., Saksena, N.K., 2010. Evidence for possible biological advantages of the newly emerging HIV-1 circulating recombinant form from Malaysia - CRF33_01B in comparison to its progenitors - CRF01_AE and subtype B. Curr. HIV Res. 8, 259–271. https://doi.org/10.2174/157016210791111151

Lau, K.A., Wong, J.J.L., 2013a. Current Trends of HIV Recombination Worldwide. Infect. Dis. Rep. 5, e4. https://doi.org/10.4081/idr.2013.s1.e4

Lau, K.A., Wong, J.J.L., 2013b. Current Trends of HIV Recombination Worldwide. Infect. Dis. Rep. 5, e4. https://doi.org/10.4081/idr.2013.s1.e4

Leonard, C.K., Spellman, M.W., Riddle, L., Harris, R.J., Thomas, J.N., Gregory, T.J., 1990. Assignment of intrachain disulfide bonds and characterization of potential glycosylation sites of the type 1 recombinant human immunodeficiency virus envelope glycoprotein (gp120) expressed in Chinese hamster ovary cells. J. Biol. Chem. 265, 10373–10382.

Liégeois, F., Boué, V., Butel, C., Mouinga-Ondémé, A., Sica, J., Zamba, C., Peeters, M., Delaporte, E., Rouet, F., 2013. HIV type-1 group O infection in Gabon: low prevalence rate but circulation of genetically diverse and drug-resistant HIV type-1 group O strains. AIDS Res. Hum. Retroviruses 29, 1085–1090. https://doi.org/10.1089/aid.2012.0375

Linde, M.E., Colquhoun, D.R., Ubaida Mohien, C., Kole, T., Aquino, V., Cotter, R., Edwards, N., Hildreth, J.E.K., Graham, D.R., 2013. The conserved set of host proteins incorporated into HIV-1 virions suggests a common egress pathway in multiple cell types. J. Proteome Res. 12, 2045–2054. https://doi.org/10.1021/pr300918r

Los Alamos National Laboratory HIV Sequence Database [WWW Document], 2019. URL http://www.hiv.lanl.gov/

Magiorkinis, G., Sypsa, V., Magiorkinis, E., Paraskevis, D., Katsoulidou, A., Belshaw, R., Fraser, C., Pybus, O.G., Hatzakis, A., 2013. Integrating phylodynamics and epidemiology to estimate transmission diversity in viral epidemics. PLoS Comput. Biol. 9, e1002876. https://doi.org/10.1371/journal.pcbi.1002876

Makin, J.K., Francis, K.L., Polonsky, M.J., Renzaho, A.M.N., 2019. Interventions to Increase Blood Donation among Ethnic/Racial Minorities: A Systematic Review. J. Environ. Public Health 2019, 6810959. https://doi.org/10.1155/2019/6810959

Marx, P.A., Li, Y., Lerche, N.W., Sutjipto, S., Gettie, A., Yee, J.A., Brotman, B.H., Prince, A.M., Hanson, A., Webster, R.G., 1991. Isolation of a simian

immunodeficiency virus related to human immunodeficiency virus type 2 from a west African pet sooty mangabey. J. Virol. 65, 4480–4485.
McMichael, A., Klenerman, P., 2002. HIV/AIDS. HLA leaves its footprints on HIV. Science 296, 1410–1411. https://doi.org/10.1126/science.1072492
McVean, G., Awadalla, P., Fearnhead, P., 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160, 1231–1241.
Meyer, D., C Aguiar, V.R., Bitarello, B.D., C Brandt, D.Y., Nunes, K., 2018. A genomic perspective on HLA evolution. Immunogenetics 70, 5–27. https://doi.org/10.1007/s00251-017-1017-3
Milner, D.A., 2020. Diagnostic Pathology: infectious diseases.
Moore, C.B., John, M., James, I.R., Christiansen, F.T., Witt, C.S., Mallal, S.A., 2002. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. Science 296, 1439–1443. https://doi.org/10.1126/science.1069660
Morse, S.S., 1995. Factors in the emergence of infectious diseases. Emerg. Infect. Dis. 1, 7–15. https://doi.org/10.3201/eid0101.950102
Muller, H.J., 1964. THE RELATION OF RECOMBINATION TO MUTATIONAL ADVANCE. Mutat. Res. 106, 2–9. https://doi.org/10.1016/0027-5107(64)90047-8
Müller, V., Fraser, C., Herbeck, J.T., 2011. A Strong Case for Viral Genetic Factors in HIV Virulence. Viruses 3, 204–216. https://doi.org/10.3390/v3030204
Myers, G., 1994. Tenth anniversary perspectives on AIDS. HIV: between past and future. AIDS Res. Hum. Retroviruses 10, 1317–1324. https://doi.org/10.1089/aid.1994.10.1317
Myers, G., Korber, B., Berzofsky, J.A., Pavlakis, G., Smith, R.F. (Eds.), 1992. Human retroviruses and aids, 1992. A compilation and analysis of nucleic acid and amino acid sequences.
Myers, G., Korber, B., Wain-Hobson, S., Jeang, K.-T., Henderson, L.E., Pavlakis, G.N., 1995. Human retroviruses and AIDS 1994 (No. LA-UR--94-4171, 10116202). https://doi.org/10.2172/10116202
Njai, H.F., Gali, Y., Vanham, G., Clybergh, C., Jennes, W., Vidal, N., Butel, C., Mpoudi-Ngolle, E., Peeters, M., Ariën, K.K., 2006. The predominance of Human Immunodeficiency Virus type 1 (HIV-1) circulating recombinant form 02 (CRF02_AG) in West Central Africa may be related to its replicative fitness. Retrovirology 3, 40. https://doi.org/10.1186/1742-4690-3-40
Of dialects, armies and navies, 2010. . The Economist.
Ong, C.K., Chan, S.Y., Campo, M.S., Fujinaga, K., Mavromara-Nazos, P., Labropoulou, V., Pfister, H., Tay, S.K., ter Meulen, J., Villa, L.L., 1993. Evolution of human papillomavirus type 18: an ancient phylogenetic root in Africa and intratype diversity reflect coevolution with human ethnic groups. J. Virol. 67, 6424–6431. https://doi.org/10.1128/JVI.67.11.6424-6431.1993
Pagani, L., Kivisild, T., Tarekegn, A., Ekong, R., Plaster, C., Gallego Romero, I., Ayub, Q., Mehdi, S.Q., Thomas, M.G., Luiselli, D., Bekele, E., Bradman, N., Balding, D.J., Tyler-Smith, C., 2012. Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. Am. J. Hum. Genet. 91, 83–96. https://doi.org/10.1016/j.ajhg.2012.05.015
Papkou, A., Gokhale, C.S., Traulsen, A., Schulenburg, H., 2016. Host–parasite coevolution: why changing population size matters. Zoology 119, 330–338. https://doi.org/10.1016/j.zool.2016.02.001

Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G.H., Barreiro, L.B., Froment, A., Heyer, E., Massougbodji, A., Fortes-Lima, C., Migot-Nabias, F., Bellis, G., Dugoujon, J.-M., Pereira, J.B., Fernandes, V., Pereira, L., Van der Veen, L., Mouguiama-Daouda, P., Bustamante, C.D., Hombert, J.-M., Quintana-Murci, L., 2017. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. Science 356, 543–546. https://doi.org/10.1126/science.aal1988

Peeters, M., Jung, M., Ayouba, A., 2013. The origin and molecular epidemiology of HIV. Expert Rev. Anti Infect. Ther. 11, 885–896. https://doi.org/10.1586/14787210.2013.825443

Peeters, M., Sharp, P.M., 2000. Genetic diversity of HIV-1: the moving target. AIDS Lond. Engl. 14 Suppl 3, S129-140.

Pepin, J., 2011. The Origins of AIDS. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139005234

Plantier, J.-C., Leoz, M., Dickerson, J.E., De Oliveira, F., Cordonnier, F., Lemée, V., Damond, F., Robertson, D.L., Simon, F., 2009. A new human immunodeficiency virus derived from gorillas. Nat. Med. 15, 871–872. https://doi.org/10.1038/nm.2016

Pollakis, G., Abebe, A., Kliphuis, A., De Wit, T.F.R., Fisseha, B., Tegbaru, B., Tesfaye, G., Negassa, H., Mengistu, Y., Fontanet, A.L., Cornelissen, M., Goudsmit, J., 2003a. Recombination of HIV type 1C (C'/C") in Ethiopia: possible link of EthHIV-1C' to subtype C sequences from the high-prevalence epidemics in India and Southern Africa. AIDS Res. Hum. Retroviruses 19, 999–1008. https://doi.org/10.1089/088922203322588350

Pollakis, G., Abebe, A., Kliphuis, A., De Wit, T.F.R., Fisseha, B., Tegbaru, B., Tesfaye, G., Negassa, H., Mengistu, Y., Fontanet, A.L., Cornelissen, M., Goudsmit, J., 2003b. Recombination of HIV type 1C (C'/C") in Ethiopia: possible link of EthHIV-1C' to subtype C sequences from the high-prevalence epidemics in India and Southern Africa. AIDS Res. Hum. Retroviruses 19, 999–1008. https://doi.org/10.1089/088922203322588350

Pritchard, L.K., Vasiljevic, S., Ozorowski, G., Seabright, G.E., Cupo, A., Ringe, R., Kim, H.J., Sanders, R.W., Doores, K.J., Burton, D.R., Wilson, I.A., Ward, A.B., Moore, J.P., Crispin, M., 2015. Structural Constraints Determine the Glycosylation of HIV-1 Envelope Trimers. Cell Rep. 11, 1604–1613. https://doi.org/10.1016/j.celrep.2015.05.017

Pybus, O.G., Rambaut, A., 2009. Evolutionary analysis of the dynamics of viral infectious disease. Nat. Rev. Genet. 10, 540–550. https://doi.org/10.1038/nrg2583

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rambaut, A., Posada, D., Crandall, K.A., Holmes, E.C., 2004. The causes and consequences of HIV evolution. Nat. Rev. Genet. 5, 52–61. https://doi.org/10.1038/nrg1246

Rambaut, A., Robertson, D.L., Pybus, O.G., Peeters, M., Holmes, E.C., 2001. Human immunodeficiency virus. Phylogeny and the origin of HIV-1. Nature 410, 1047–1048. https://doi.org/10.1038/35074179

Redd, A.D., Quinn, T.C., Tobian, A.A.R., 2013. Frequency and implications of HIV superinfection. Lancet Infect. Dis. 13, 622–628. https://doi.org/10.1016/S1473-3099(13)70066-5

René, B., Mauffret, O., Fossé, P., 2018. Retroviral nucleocapsid proteins and DNA strand transfers. Biochim. Open 7, 10–25. https://doi.org/10.1016/j.biopen.2018.07.001

Robertson, D.L., Anderson, J.P., Bradac, J.A., Carr, J.K., Foley, B., Funkhouser, R.K., Gao, F., Hahn, B.H., Kalish, M.L., Kuiken, C., Learn, G.H., Leitner, T., McCutchan, F., Osmanov, S., Peeters, M., Pieniazek, D., Salminen, M., Sharp, P.M., Wolinsky, S., Korber, B., 2000. HIV-1 nomenclature proposal. Science 288, 55–56. https://doi.org/10.1126/science.288.5463.55d

Rolland, M., Carlson, J.M., Manocheewa, S., Swain, J.V., Lanxon-Cookson, E., Deng, W., Rousseau, C.M., Raugi, D.N., Learn, G.H., Maust, B.S., Coovadia, H., Ndung'u, T., Goulder, P.J.R., Walker, B.D., Brander, C., Heckerman, D.E., Mullins, J.I., 2010. Amino-acid co-variation in HIV-1 Gag subtype C: HLA-mediated selection pressure and compensatory dynamics. PloS One 5. https://doi.org/10.1371/journal.pone.0012463

Roques, P., Robertson, D.L., Souquière, S., Damond, F., Ayouba, A., Farfara, I., Depienne, C., Nerrienet, E., Dormont, D., Brun-Vézinet, F., Simon, F., Mauclère, P., 2002. Phylogenetic Analysis of 49 Newly Derived HIV-1 Group O Strains: High Viral Diversity but No Group M-like Subtype Structure. Virology 302, 259–273. https://doi.org/10.1006/viro.2002.1430

Salemi, M., Strimmer, K., Hall, W.W., Duffy, M., Delaporte, E., Mboup, S., Peeters, M., Vandamme, A.M., 2001. Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution. FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. 15, 276–278. https://doi.org/10.1096/fj.00-0449fje

Sanchez-Mazas, A., Fernandez-Viña, M., Middleton, D., Hollenbach, J.A., Buhler, S., Di, D., Rajalingam, R., Dugoujon, J.-M., Mack, S.J., Thorsby, E., 2011. Immunogenetics as a tool in anthropological studies. Immunology 133, 143–164. https://doi.org/10.1111/j.1365-2567.2011.03438.x

Sanjuan, R., Nebot, M.R., Chirico, N., Mansky, L.M., Belshaw, R., 2010. Viral Mutation Rates. J. Virol. 84, 9733–9748. https://doi.org/10.1128/JVI.00694-10

Sarr, A.D., Eisen, G., Guèye-Ndiaye, A., Mullins, C., Traoré, I., Dia, M.C., Sankalé, J.-L., Faye, D., Mboup, S., Kanki, P., 2005. Viral dynamics of primary HIV-1 infection in Senegal, West Africa. J. Infect. Dis. 191, 1460–1467. https://doi.org/10.1086/429409

Sauter, D., Schindler, M., Specht, A., Landford, W.N., Münch, J., Kim, K.-A., Votteler, J., Schubert, U., Bibollet-Ruche, F., Keele, B.F., Takehisa, J., Ogando, Y., Ochsenbauer, C., Kappes, J.C., Ayouba, A., Peeters, M., Learn, G.H., Shaw, G., Sharp, P.M., Bieniasz, P., Hahn, B.H., Hatziioannou, T., Kirchhoff, F., 2009. Tetherin-driven adaptation of Vpu and Nef function and the evolution of pandemic and nonpandemic HIV-1 strains. Cell Host Microbe 6, 409–421. https://doi.org/10.1016/j.chom.2009.10.004

Schlub, T.E., Grimm, A.J., Smyth, R.P., Cromer, D., Chopra, A., Mallal, S., Venturi, V., Waugh, C., Mak, J., Davenport, M.P., 2014. Fifteen to twenty percent of HIV substitution mutations are associated with recombination. J. Virol. 88, 3837–3849. https://doi.org/10.1128/JVI.03136-13

Schneidewind, A., Brockman, M.A., Sidney, J., Wang, Y.E., Chen, H., Suscovich, T.J., Li, B., Adam, R.I., Allgaier, R.L., Mothé, B.R., Kuntzen, T., Oniangue-Ndza, C., Trocha, A., Yu, X.G., Brander, C., Sette, A., Walker, B.D., Allen, T.M., 2008. Structural and functional constraints limit options for cytotoxic T-lymphocyte escape in the immunodominant HLA-B27-restricted epitope in

human immunodeficiency virus type 1 capsid. J. Virol. 82, 5594–5605. https://doi.org/10.1128/JVI.02356-07
Shannon, C.E., 1948. A mathematical theory of communication. ACM SIGMOBILE Mob. Comput. Commun. Rev. 5, 3. https://doi.org/10.1145/584091.584093
Sharp, P.M., Hahn, B.H., 2008. AIDS: prehistory of HIV-1. Nature 455, 605–606. https://doi.org/10.1038/455605a
Shaw, G.M., Hunter, E., 2012. HIV Transmission. Cold Spring Harb. Perspect. Med. 2, a006965–a006965. https://doi.org/10.1101/cshperspect.a006965
Simon-Loriere, E., Martin, D.P., Weeks, K.M., Negroni, M., 2010. RNA Structures Facilitate Recombination-Mediated Gene Swapping in HIV-1. J. Virol. 84. https://doi.org/10.1128/JVI.01302-10
Smith, D.G., 1990. Thailand: AIDS crisis looms. Lancet Lond. Engl. 335, 781–782. https://doi.org/10.1016/0140-6736(90)90883-7
Smyth, R.P., Schlub, T.E., Grimm, A.J., Waugh, C., Ellenberg, P., Chopra, A., Mallal, S., Cromer, D., Mak, J., Davenport, M.P., Hahn, B.H., 2014. Identifying Recombination Hot Spots in the HIV-1 Genome. J. Virol. 88, 2891–2902. https://doi.org/10.1128/JVI.03014-13
Song, W., He, D., Brill, I., Malhotra, R., Mulenga, J., Allen, S., Hunter, E., Tang, J., Kaslow, R.A., 2011. Disparate associations of HLA class I markers with HIV-1 acquisition and control of viremia in an African population. PloS One 6, e23469. https://doi.org/10.1371/journal.pone.0023469
Tarosso, L.F., Vieira, V.A., Sauer, M.M., Tomiyama, H.I., Kalil, J., Kallas, E.G., 2017. Conserved HIV-1 Gag p24 Epitopes Elicit Cellular Immune Responses That Impact Disease Outcome. AIDS Res. Hum. Retroviruses 33, 832–842. https://doi.org/10.1089/AID.2016.0168
Taylor, B.S., Sobieszczyk, M.E., McCutchan, F.E., Hammer, S.M., 2008. The challenge of HIV-1 subtype diversity. N. Engl. J. Med. 358, 1590–1602. https://doi.org/10.1056/NEJMra0706737
Tebit, D.M., Arts, E.J., 2011. Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. Lancet Infect. Dis. 11, 45–56. https://doi.org/10.1016/S1473-3099(10)70186-9
Tebit, D.M., Lobritz, M., Lalonde, M., Immonen, T., Singh, K., Sarafianos, S., Herchenröder, O., Kräusslich, H.-G., Arts, E.J., 2010. Divergent Evolution in Reverse Transcriptase (RT) of HIV-1 Group O and M Lineages: Impact on Structure, Fitness, and Sensitivity to Nonnucleoside RT Inhibitors. J. Virol. 84, 9817–9830. https://doi.org/10.1128/JVI.00991-10
Tenzer, S., Crawford, H., Pymm, P., Gifford, R., Sreenu, V.B., Weimershaus, M., de Oliveira, T., Burgevin, A., Gerstoft, J., Akkad, N., Lunn, D., Fugger, L., Bell, J., Schild, H., van Endert, P., Iversen, A.K.N., 2014. HIV-1 adaptation to antigen processing results in population-level immune evasion and affects subtype diversification. Cell Rep. 7, 448–463. https://doi.org/10.1016/j.celrep.2014.03.031
Thomas Splettstoesser, 2014. . SCIstyle. URL https://www.scistyle.com/ (accessed 7.30.19).
Thomson, M.M., Pérez-Alvarez, L., Nájera, R., 2002. Molecular epidemiology of HIV-1 genetic forms and its significance for vaccine development and therapy. Lancet Infect. Dis. 2, 461–471. https://doi.org/10.1016/s1473-3099(02)00343-2
Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.-M., Doumbo, O., Ibrahim, M., Juma, A.T., Kotze, M.J., Lema, G., Moore, J.H., Mortensen, H., Nyambo, T.B., Omar,

S.A., Powell, K., Pretorius, G.S., Smith, M.W., Thera, M.A., Wambebe, C., Weber, J.L., Williams, S.M., 2009. The genetic structure and history of Africans and African Americans. Science 324, 1035–1044. https://doi.org/10.1126/science.1172257

Triques, K., Bourgeois, A., Vidal, N., Mpoudi-Ngole, E., Mulanga-Kabeya, C., Nzilambi, N., Torimiro, N., Saman, E., Delaporte, E., Peeters, M., 2000. Near-full-length genome sequencing of divergent African HIV type 1 subtype F viruses leads to the identification of a new HIV type 1 subtype designated K. AIDS Res. Hum. Retroviruses 16, 139–151. https://doi.org/10.1089/088922200309485

UNAIDS, 2019. Global HIV & AIDS statistics — 2019 fact sheet.

Vallari, A., Bodelle, P., Ngansop, C., Makamche, F., Ndembi, N., Mbanya, D., Kaptué, L., Gürtler, L.G., McArthur, C.P., Devare, S.G., Brennan, C.A., 2010. Four new HIV-1 group N isolates from Cameroon: Prevalence continues to be low. AIDS Res. Hum. Retroviruses 26, 109–115. https://doi.org/10.1089/aid.2009.0178

Vallari, A., Holzmayer, V., Harris, B., Yamaguchi, J., Ngansop, C., Makamche, F., Mbanya, D., Kaptué, L., Ndembi, N., Gürtler, L., Devare, S., Brennan, C.A., 2011. Confirmation of putative HIV-1 group P in Cameroon. J. Virol. 85, 1403–1407. https://doi.org/10.1128/JVI.02005-10

Van Doorslaer, K., 2013. Evolution of the Papillomaviridae. Virology 445, 11–20. https://doi.org/10.1016/j.virol.2013.05.012

Van Heuverswyn, F., Li, Y., Bailes, E., Neel, C., Lafay, B., Keele, B.F., Shaw, K.S., Takehisa, J., Kraus, M.H., Loul, S., Butel, C., Liegeois, F., Yangda, B., Sharp, P.M., Mpoudi-Ngole, E., Delaporte, E., Hahn, B.H., Peeters, M., 2007. Genetic diversity and phylogeographic clustering of SIVcpzPtt in wild chimpanzees in Cameroon. Virology 368, 155–171. https://doi.org/10.1016/j.virol.2007.06.018

Veillette, M., Coutu, M., Richard, J., Batraville, L.-A., Dagher, O., Bernard, N., Tremblay, C., Kaufmann, D.E., Roger, M., Finzi, A., 2015. The HIV-1 gp120 CD4-bound conformation is preferentially targeted by antibody-dependent cellular cytotoxicity-mediating antibodies in sera from HIV-1-infected individuals. J. Virol. 89, 545–551. https://doi.org/10.1128/JVI.02868-14

Vermund, S.H., Leigh-Brown, A.J., 2012. The HIV Epidemic: High-Income Countries. Cold Spring Harb. Perspect. Med. 2, a007195–a007195. https://doi.org/10.1101/cshperspect.a007195

Vidal, N., Peeters, M., Mulanga-Kabeya, C., Nzilambi, N., Robertson, D., Ilunga, W., Sema, H., Tshimanga, K., Bongo, B., Delaporte, E., 2000. Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. J. Virol. 74, 10498–10507. https://doi.org/10.1128/jvi.74.22.10498-10507.2000

Volberding, P.A. (Ed.), 2008. Global HIV/AIDS medicine, 1st publ. ed. Saunders Elsevier, Philadlphia, PA.

Wangroongsarb, Y., Weniger, B.G., Wasi, C., Traisupa, A., Kunasol, P., Rojanapithayakorn, W., Fucharoen, S., 1985. Prevalence of HTLV-III/LAV antibody in selected populations in Thailand. Southeast Asian J. Trop. Med. Public Health 16, 517–520.

WHO network for HIV isolation and characterization, 1994. HIV type 1 variation in World Health Organization-sponsored vaccine evaluation sites: genetic screening, sequence analysis, and preliminary biological characterization of selected viral strains. WHO Network for HIV Isolation and Characterization.

AIDS Res. Hum. Retroviruses 10, 1327–1343. https://doi.org/10.1089/aid.1994.10.1327

Wickham, H., 2009. Ggplot2: elegant graphics for data analysis, Use R! Springer, New York.

Wirachsilp, P., Kantakamalakul, W., Foongladda, S., Chuenchitra, T., Kohriangudom, S., Athipanyasilp, N., Tanprasertsuk, S., Gasitrong, M., Sutthent, R., 2007. Surveillance of subtype and genetic variation of the circulating strains of HIV-1 in Thailand. Southeast Asian J. Trop. Med. Public Health 38, 814–827.

Wolfe, N.D., Switzer, W.M., Carr, J.K., Bhullar, V.B., Shanmugam, V., Tamoufe, U., Prosser, A.T., Torimiro, J.N., Wright, A., Mpoudi-Ngole, E., McCutchan, F.E., Birx, D.L., Folks, T.M., Burke, D.S., Heneine, W., 2004. Naturally acquired simian retrovirus infections in central African hunters. Lancet Lond. Engl. 363, 932–937. https://doi.org/10.1016/S0140-6736(04)15787-5

Woo, J., Robertson, D.L., Lovell, S.C., 2010. Constraints on HIV-1 diversity from protein structure. J. Virol. 84, 12995–13003. https://doi.org/10.1128/JVI.00702-10

Worboys, M., 1994. The comparative history of sleeping sickness in east and central Africa. Hist. Sci. Annu. Rev. Lit. Res. Teach. 32, 89–102. https://doi.org/10.1177/007327539403200103

World Bank, 2019. World Development Indicators [WWW Document]. URL https://databank.worldbank.org

World Health Organization, 2016. Global health sector strategy on Sexually Transmitted Infections, 2016-2021. 2016. WHO.

Worobey, M., Gemmel, M., Teuwen, D.E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J.-J., Kabongo, J.-M.M., Kalengayi, R.M., Van Marck, E., Gilbert, M.T.P., Wolinsky, S.M., 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature 455, 661–664. https://doi.org/10.1038/nature07390

Yamaguchi, J., Bodelle, P., Kaptué, L., Zekeng, L., Gürtler, L.G., Devare, S.G., Brennan, C.A., 2003. Near Full-Length Genomes of 15 HIV Type 1 Group O Isolates. AIDS Res. Hum. Retroviruses 19, 979–988. https://doi.org/10.1089/088922203322588332

Yuan, T., Li, J., Zhang, M.-Y., 2013. HIV-1 envelope glycoprotein variable loops are indispensable for envelope structural integrity and virus entry. PloS One 8, e69789. https://doi.org/10.1371/journal.pone.0069789

Zamani, C., Elzey, J.D., Hildreth, J.E., 2013. Greater ethnic diversity correlates with lower HIV prevalence in Africa: justification for an alloimmunity vaccine. HIVAIDS Auckl. NZ 5, 75–80. https://doi.org/10.2147/HIV.S38922

Zuñiga, R., Lucchetti, A., Galvan, P., Sanchez, S., Sanchez, C., Hernandez, A., Sanchez, H., Frahm, N., Linde, C.H., Hewitt, H.S., Hildebrand, W., Altfeld, M., Allen, T.M., Walker, B.D., Korber, B.T., Leitner, T., Sanchez, J., Brander, C., 2006. Relative dominance of Gag p24-specific cytotoxic T lymphocytes is associated with human immunodeficiency virus control. J. Virol. 80, 3122–3125. https://doi.org/10.1128/JVI.80.6.3122-3125.2006

# Appendix 1A

List of papers checked to determine how HIV sequence submitted LANL was obtained:


Abecasis, A.B., Lemey, P., Vidal, N., de Oliveira, T., Peeters, M., Camacho, R., Shapiro, B., Rambaut, A., Vandamme, A.-M., 2007. Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form. J. Virol. 81, 8543–8551. https://doi.org/10.1128/JVI.00463-07

Alizon, M., Wain-Hobson, S., Montagnier, L., Sonigo, P., 1986. Genetic variability of the AIDS virus: nucleotide sequence analysis of two isolates from African patients. Cell 46, 63–74. https://doi.org/10.1016/0092-8674(86)90860-3

Aman, S., Borggren, M., Fomsgaard, A., Kebe, K., Jaye, A., 2015. Optimization of full-length sequencing of HIV type 1 from Senegal. Unpublished direct submission to Genbank.

Amogne, W., Bontell, I., Grossmann, S., Aderaye, G., Lindquist, L., Sönnerborg, A., Neogi, U., 2016. Phylogenetic Analysis of Ethiopian HIV-1 Subtype C Near Full-Length Genomes Reveals High Intrasubtype Diversity and a Strong Geographical Cluster. AIDS Res. Hum. Retroviruses 32, 471–474. https://doi.org/10.1089/aid.2015.0380

Anderson, J.P., Rodrigo, A.G., Learn, G.H., Madan, A., Delahunty, C., Coon, M., Girard, M., Osmanov, S., Hood, L., Mullins, J.I., 2000. Testing the hypothesis of a recombinant origin of human immunodeficiency virus type 1 subtype E. J. Virol. 74, 10752–10765. https://doi.org/10.1128/jvi.74.22.10752-10765.2000

Arroyo, M.A., Hoelscher, M., Sanders-Buell, E., Herbinger, K.-H., Samky, E., Maboko, L., Hoffmann, O., Robb, M.R., Birx, D.L., McCutchan, F.E., 2004. HIV type 1 subtypes among blood donors in the Mbeya region of southwest Tanzania. AIDS Res. Hum. Retroviruses 20, 895–901. https://doi.org/10.1089/0889222041725235

Arroyo, M.A., Hoelscher, M., Sateren, W., Samky, E., Maboko, L., Hoffmann, O., Kijak, G., Robb, M., Birx, D.L., McCutchan, F.E., 2005. HIV-1 diversity and prevalence differ between urban and rural areas in the Mbeya region of Tanzania. AIDS 19, 1517–1524. https://doi.org/10.1097/01.aids.0000183515.14642.76

Baalwa, J., Wang, S., Parrish, N.F., Decker, J.M., Keele, B.F., Learn, G.H., Yue, L., Ruzagira, E., Ssemwanga, D., Kamali, A., Amornkul, P.N., Price, M.A., Kappes, J.C., Karita, E., Kaleebu, P., Sanders, E., Gilmour, J., Allen, S., Hunter, E., Montefiori, D.C., Haynes, B.F., Cormier, E., Hahn, B.H., Shaw, G.M., 2013. Molecular identification, cloning and characterization of transmitted/founder HIV-1 subtype A, D and A/D infectious molecular clones. Virology 436, 33–48. https://doi.org/10.1016/j.virol.2012.10.009

Baesi, K., Moallemi, S., Farrokhi, M., Alinaghi, S.A.S., Truong, H.-H.M., 2014. Subtype classification of Iranian HIV-1 sequences registered in the HIV databases, 2006-2013. PLoS ONE 9, e105098. https://doi.org/10.1371/journal.pone.0105098

Baker, C.A.R., Bousheri, S., Ssewanyana, I., Jones, N.G., K'Aluoch, O., Baliruno, D., Ssali, F., Cao, H., 2007. HIV Subtypes Distribution and Implication for Antiretroviral Treatment in a Ugandan Population. J Int Assoc Physicians AIDS Care (Chic) 6, 260–263. https://doi.org/10.1177/1545109707303938

Baldrich-Rubio, E., Anagonou, S., Stirrups, K., Lafia, E., Candotti, D., Lee, H., Allain, J.P., 2001. A complex human immunodeficiency virus type 1 A/G/J recombinant virus isolated from a seronegative patient with AIDS from Benin, West Africa. J. Gen. Virol. 82, 1095–1106. https://doi.org/10.1099/0022-1317-82-5-1095

Bártolo, I., Calado, R., Borrego, P., Leitner, T., Taveira, N., 2016. Rare HIV-1 Subtype J Genomes and a New H/U/CRF02_AG Recombinant Genome Suggests an Ancient Origin of HIV-1 in Angola.

AIDS Res. Hum. Retroviruses 32, 822–828. https://doi.org/10.1089/AID.2016.0084
Barugahare, B., Baker, C., K'Aluoch, O., Donovan, R., Elrefaei, M., Eggena, M., Jones, N., Mutalya, S., Kityo, C., Mugyenyi, P., Cao, H., 2005. Human immunodeficiency virus-specific responses in adult Ugandans: patterns of cross-clade recognition. J. Virol. 79, 4132–4139. https://doi.org/10.1128/JVI.79.7.4132-4139.2005
Berg, M.G., Yamaguchi, J., Alessandri-Gradt, E., Tell, R.W., Plantier, J.-C., Brennan, C.A., 2016. A Pan-HIV Strategy for Complete Genome Sequencing. J. Clin. Microbiol. 54, 868–882. https://doi.org/10.1128/JCM.02479-15
Billings, E., Sanders-Buell, E., Bose, M., Bradfield, A., Lei, E., Kijak, G.H., Arroyo, M.A., Kibaya, R.M., Scott, P.T., Wasunna, M.K., Sawe, F.K., Shaffer, D.N., Birx, D.L., McCutchan, F.E., Michael, N.L., Robb, M.L., Kim, J.H., Tovanabutra, S., 2015. The Number and Complexity of Pure and Recombinant HIV-1 Strains Observed within Incident Infections during the HIV and Malaria Cohort Study Conducted in Kericho, Kenya, from 2003 to 2006. PLoS ONE 10, e0135124. https://doi.org/10.1371/journal.pone.0135124
Billings, E., Sanders-Buell, E., Bose, M., Kijak, G.H., Bradfield, A., Crossler, J., Arroyo, M.A., Maboko, L., Hoffmann, O., Geis, S., Birx, D.L., Kim, J.H., Michael, N.L., Robb, M.L., Hoelscher, M., Tovanabutra, S., 2017. HIV-1 Genetic Diversity Among Incident Infections in Mbeya, Tanzania. AIDS Res. Hum. Retroviruses 33, 373–381. https://doi.org/10.1089/AID.2016.0111
Birditt, B.A., Rousseau, C.M., Korber, B.T., Goulder, P., Brander, C., Kiepiela, P., Walker, B.D., Mullins, J.I., 2005. HIV HLA epitope mapping from Durban, South Africa. Unpublished direct submission to Genbank.
Bodelle, P., Vallari, A., Coffey, R., McArthur, C.P., Beyeme, M., Devare, S.G., Schochetman, G., Brennan, C.A., 2004. Identification and genomic sequence of an HIV type 1 group N isolate from Cameroon. AIDS Res. Hum. Retroviruses 20, 902–908. https://doi.org/10.1089/0889222041725262
Brave, A., Ljungberg, K., Wahren, B., 2002. HIV-1 92UG031 p17p24gag. Unpublished direct submission to Genbank.
Bredell, H., Martin, D.P., Van Harmelen, J., Varsani, A., Sheppard, H.W., Donovan, R., Gray, C.M., HIVNET028 Study Team, Williamson, C., 2007. HIV type 1 subtype C gag and nef diversity in Southern Africa. AIDS Res. Hum. Retroviruses 23, 477–481. https://doi.org/10.1089/aid.2006.0232
Brennan, C.A., Bodelle, P., Coffey, R., Devare, S.G., Golden, A., Hackett, J., Harris, B., Holzmayer, V., Luk, K.-C., Schochetman, G., Swanson, P., Yamaguchi, J., Vallari, A., Ndembi, N., Ngansop, C., Makamche, F., Mbanya, D., Gürtler, L.G., Zekeng, L., Kaptué, L., 2008. The prevalence of diverse HIV-1 strains was stable in Cameroonian blood donors from 1996 to 2004. J. Acquir. Immune Defic. Syndr. 49, 432–439. https://doi.org/10.1097/QAI.0b013e31818a6561
Brown, B.K., Darden, J.M., Tovanabutra, S., Oblander, T., Frost, J., Sanders-Buell, E., de Souza, M.S., Birx, D.L., McCutchan, F.E., Polonis, V.R., 2005. Biologic and genetic characterization of a panel of 60 human immunodeficiency virus type 1 isolates, representing clades A, B, C, D, CRF01_AE, and CRF02_AG, for the development and assessment of candidate vaccines. J. Virol. 79, 6089–6101. https://doi.org/10.1128/JVI.79.10.6089-6101.2005
Burns, C.C., Gleason, L.M., Mozaffarian, A., Giachetti, C., Carr, J.K., Overbaugh, J., 2002. Sequence variability of the integrase protein from a diverse collection of HIV type 1 isolates representing several subtypes. AIDS Res. Hum. Retroviruses 18, 1031–1041. https://doi.org/10.1089/08892220260235399
Campbell, M.S., Mullins, J.I., Hughes, J.P., Celum, C., Wong, K.G., Raugi, D.N., Sorensen, S., Stoddard, J.N., Zhao, H., Deng, W., Kahle, E., Panteleeff, D., Baeten, J.M., McCutchan, F.E., Albert, J., Leitner, T., Wald, A., Corey, L., Lingappa, J.R., Partners in Prevention HSV/HIV Transmission Study Team, 2011. Viral linkage in HIV-1 seroconverters and their partners in an HIV-1 prevention clinical trial. PLoS ONE 6, e16986. https://doi.org/10.1371/journal.pone.0016986
Candotti, D., Adu-Sarkodie, Y., Davies, F., Baldrich-Rubio, E., Stirrups, K., Lee, H., Allain, J.P., 2000. AIDS in an HIV-seronegative Ghanaian woman with intersubtype A/G recombinant HIV-1 infection. J. Med. Virol. 62, 1–8. https://doi.org/10.1002/1096-9071(200009)62:1<1::aid-jmv1>3.0.co;2-3

Cappy, P., De Oliveira, F., Gueudin, M., Alessandri-Gradt, E., Plantier, J.-C., 2016. A Multiplex PCR Approach for Detecting Dual Infections and Recombinants Involving Major HIV Variants. J. Clin. Microbiol. 54, 1282–1288. https://doi.org/10.1128/JCM.03222-15

Carlson, J.M., Du, V.Y., Pfeifer, N., Bansal, A., Tan, V.Y.F., Power, K., Brumme, C.J., Kreimer, A., DeZiel, C.E., Fusi, N., Schaefer, M., Brockman, M.A., Gilmour, J., Price, M.A., Kilembe, W., Haubrich, R., John, M., Mallal, S., Shapiro, R., Frater, J., Harrigan, P.R., Ndung'u, T., Allen, S., Heckerman, D., Sidney, J., Allen, T.M., Goulder, P.J.R., Brumme, Z.L., Hunter, E., Goepfert, P.A., 2016. Impact of pre-adapted HIV transmission. Nat. Med. 22, 606–613. https://doi.org/10.1038/nm.4100

Carlson, J.M., Schaefer, M., Monaco, D.C., Batorsky, R., Claiborne, D.T., Prince, J., Deymier, M.J., Ende, Z.S., Klatt, N.R., DeZiel, C.E., Lin, T.-H., Peng, J., Seese, A.M., Shapiro, R., Frater, J., Ndung'u, T., Tang, J., Goepfert, P., Gilmour, J., Price, M.A., Kilembe, W., Heckerman, D., Goulder, P.J.R., Allen, T.M., Allen, S., Hunter, E., 2014. HIV transmission. Selection bias at the heterosexual HIV-1 transmission bottleneck. Science 345, 1254031. https://doi.org/10.1126/science.1254031

Carr, J.K., Kim, B., Sanders-Buell, E., Salminen, M.O., Alaeus, A., Albert, J.A., Birx, D.L., McCutchan, F.E., 1998. HIV-1 isolate VI354 from Gabon, complete genome. Unpublished direct submission to Genbank.

Carr, J.K., Lar, P., Bemis, K., Jelpe, J., Ayuba, L., Zella, D., Blattner, W., Kanki, P., Abimiku, A.G., 2005. HIV Subtype and Drug Resistance Patterns among Drug Naive Persons in Jos, Nigeria. Unpublished direct submission to Genbank.

Carr, J.K., Salminen, M.O., Albert, J., Sanders-Buell, E., Gotte, D., Birx, D.L., McCutchan, F.E., 1998. Full genome sequences of human immunodeficiency virus type 1 subtypes G and A/G intersubtype recombinants. Virology 247, 22–31. https://doi.org/10.1006/viro.1998.9211

Carr, J.K., Torimiro, J.N., Wolfe, N.D., Eitel, M.N., Kim, B., Sanders-Buell, E., Jagodzinski, L.L., Gotte, D., Burke, D.S., Birx, D.L., McCutchan, F.E., 2001. The AG recombinant IbNG and novel strains of group M HIV-1 are common in Cameroon. Virology 286, 168–181. https://doi.org/10.1006/viro.2001.0976

Carr, J.K., Torimiro, J.N., Wolfe, N.D., Tamoufe, U., Mpoudi-Ngole, E., Kim, B., Sanders-Buell, E., Eyzaguirre, L.M., Jagodzinski, L.L., Gotte, D., Birx, D.L., McCutchan, F.E., Burke, D.S., 2006. High Genetic Complexity of HIV-1 in low-prevalence, rural regions of Cameroon. Unpublished direct submission to Genbank.

Carr, J.K., Wolfe, N.D., Torimiro, J.N., Tamoufe, U., Mpoudi-Ngole, E., Eyzaguirre, L., Birx, D.L., McCutchan, F.E., Burke, D.S., 2010. HIV-1 recombinants with multiple parental strains in low-prevalence, remote regions of Cameroon: evolutionary relics? Retrovirology 7, 39. https://doi.org/10.1186/1742-4690-7-39

Carr, J.K., Zhao, J., Tang, S., Viswanath, R., Awazi, B., Hewlett,I., 2009. Identification of a novel CRF22_01A1 recombinant form in Cameroon. Unpublished direct submission to Genbank.

Casado, G., Thomson, M.M., Sierra, M., Nájera, R., 2005. Identification of a novel HIV-1 circulating ADG intersubtype recombinant form (CRF19_cpx) in Cuba. J. Acquir. Immune Defic. Syndr. 40, 532–537. https://doi.org/10.1097/01.qai.0000186363.27587.c0

Cham, F., Heyndrickx, L., Janssens, W., Van der Auwera, G., Vereecken, K., De Houwer, K., Coppens, S., Whittle, H., van der Groen, G., 2000. Study of HIV type 1 gag/env variability in The Gambia, using a multiplex DNA polymerase chain reaction. AIDS Res. Hum. Retroviruses 16, 1915–1919. https://doi.org/10.1089/08892220050195874

Chang, C.-H., Kist, N.C., Stuart Chester, T.L., Sreenu, V.B., Herman, M., Luo, M., Lunn, D., Bell, J., Plummer, F.A., Ball, T.B., Katzourakis, A., Iversen, A.K.N., 2015. HIV-infected sex workers with beneficial HLA-variants are potential hubs for selection of HIV-1 recombinants that may affect disease progression. Sci Rep 5, 11253. https://doi.org/10.1038/srep11253

Charurat, M., Nasidi, A., Delaney, K., Saidu, A., Croxton, T., Mondal, P., Aliyu, G.G., Constantine, N., Abimiku, A., Carr, J.K., Vertefeuille, J., Blattner, W., 2012. Characterization of acute HIV-1 infection in high-risk Nigerian populations. J. Infect. Dis. 205, 1239–1247. https://doi.org/10.1093/infdis/jis103

Choge, I., Cilliers, T., Walker, P., Taylor, N., Phoswa, M., Meyers, T., Viljoen, J., Violari, A., Gray, G., Moore, P.L., Papathanosopoulos, M., Morris, L., 2006. Genotypic and phenotypic characterization of viral isolates from HIV-1 subtype C-infected children with slow and rapid disease progression. AIDS Res. Hum. Retroviruses 22, 458–465. https://doi.org/10.1089/aid.2006.22.458

Choi, D.J., Dube, S., Spicer, T.P., Slade, H.B., Jensen, F.C., Poiesz, B.J., 1997. HIV type 1 isolate Z321, the strain used to make a therapeutic HIV type 1 immunogen, is intersubtype recombinant. AIDS Res. Hum. Retroviruses 13, 357–361. https://doi.org/10.1089/aid.1997.13.357

Chopera, D.R., Mann, J.K., Mwimanzi, P., Omarjee, S., Kuang, X.T., Ndabambi, N., Goodier, S., Martin, E., Naranbhai, V., Karim, S.A., Karim, Q.A., Brumme, Z.L., Ndung'u, T., Williamson, C., Brockman, M.A., CAPRISA 004 TRAPS Team, 2013. No evidence for selection of HIV-1 with enhanced gag-protease or Nef function among breakthrough infections in the CAPRISA 004 tenofovir microbicide trial. PLoS ONE 8, e71758. https://doi.org/10.1371/journal.pone.0071758

Chopera, D.R., Mlotshwa, M., Woodman, Z., Mlisana, K., de Assis Rosa, D., Martin, D.P., Abdool Karim, S., Gray, C.M., Williamson, C., CAPRISA 002 Study Team, 2011. Virological and immunological factors associated with HIV-1 differential disease progression in HLA-B 58:01-positive individuals. J. Virol. 85, 7070–7080. https://doi.org/10.1128/JVI.02543-10

Chopera, D.R., Woodman, Z., Mlisana, K., Mlotshwa, M., Martin, D.P., Seoighe, C., Treurnicht, F., de Rosa, D.A., Hide, W., Karim, S.A., Gray, C.M., Williamson, C., CAPRISA 002 Study Team, 2008. Transmission of HIV-1 CTL escape variants provides HLA-mismatched recipients with a survival advantage. PLoS Pathog. 4, e1000033. https://doi.org/10.1371/journal.ppat.1000033

Cilliers, T., Nhlapo, J., Coetzer, M., Orlovic, D., Ketas, T., Olson, W.C., Moore, J.P., Trkola, A., Morris, L., 2003. The CCR5 and CXCR4 coreceptors are both used by human immunodeficiency virus type 1 primary isolates from subtype C. J. Virol. 77, 4449–4456. https://doi.org/10.1128/jvi.77.7.4449-4456.2003

Cilliers, T., Patience, T., Pillay, C., Papathanasopoulos, M., Morris, L., 2004. Sensitivity of HIV type 1 subtype C isolates to the entry inhibitor T-20. AIDS Res. Hum. Retroviruses 20, 477–482. https://doi.org/10.1089/088922204323087714

Claiborne, D.T., Prince, J.L., Scully, E., Macharia, G., Micci, L., Lawson, B., Kopycinski, J., Deymier, M.J., Vanderford, T.H., Nganou-Makamdop, K., Ende, Z., Brooks, K., Tang, J., Yu, T., Lakhi, S., Kilembe, W., Silvestri, G., Douek, D., Goepfert, P.A., Price, M.A., Allen, S.A., Paiardini, M., Altfeld, M., Gilmour, J., Hunter, E., 2015. Replicative fitness of transmitted HIV-1 drives acute immune activation, proviral load in memory CD4+ T cells, and disease progression. Proc. Natl. Acad. Sci. U.S.A. 112, E1480-1489. https://doi.org/10.1073/pnas.1421607112

Coetzer, M., Cilliers, T., Papathanasopoulos, M., Ramjee, G., Karim, S.A., Williamson, C., Morris, L., 2007. Longitudinal analysis of HIV type 1 subtype C envelope sequences from South Africa. AIDS Res. Hum. Retroviruses 23, 316–321. https://doi.org/10.1089/aid.2006.0207

Cornelissen, M., van Den Burg, R., Zorgdrager, F., Goudsmit, J., 2000. Spread of distinct human immunodeficiency virus type 1 AG recombinant lineages in Africa. J. Gen. Virol. 81, 515–523. https://doi.org/10.1099/0022-1317-81-2-515

Currier, J.R., Harris, M.E., Cox, J.H., McCutchan, F.E., Birx, D.L., Maayan, S., Ferrari, G., 2005. Immunodominance and cross-reactivity of B5703-restricted CD8 T lymphocytes from HIV type 1 subtype C-infected Ethiopians. AIDS Res. Hum. Retroviruses 21, 239–245. https://doi.org/10.1089/aid.2005.21.239

Currier, J.R., Visawapoka, U., Tovanabutra, S., Mason, C.J., Birx, D.L., McCutchan, F.E., Cox, J.H., 2006. CTL epitope distribution patterns in the Gag and Nef proteins of HIV-1 from subtype A infected subjects in Kenya: use of multiple peptide sets increases the detectable breadth of the CTL response. BMC Immunol. 7, 8. https://doi.org/10.1186/1471-2172-7-8

D'Arrigo, R., Ciccozzi, M., Gori, C., Montieri, S., Aquaro, S., Bellagamba, R., Boumis, E., Di Perri, G., Pizzi, D., Antinori, A., Rezza, G., Perno, C.F., 2007. gp41 sequence variability in HIV type 1 non-B subtypes infected patients undergoing enfuvirtide pressure. AIDS Res. Hum. Retroviruses 23, 1296–1302. https://doi.org/10.1089/aid.2007.0095

De Mareuil, J., Salaun, D., Chermann, J.C., Hirsch, I., 1995. Fusogenic determinants of highly cytopathic subtype D Zairian isolate HIV-1 NDK. Virology 209, 649–653. https://doi.org/10.1006/viro.1995.1298

De Oliveira, F., Mourez, T., Vessiere, A., Ngoupo, P.-A., Alessandri-Gradt, E., Simon, F., Rousset, D., Plantier, J.-C., 2017. Multiple HIV-1/M + HIV-1/O dual infections and new HIV-1/MO inter-group recombinant forms detected in Cameroon. Retrovirology 14, 1. https://doi.org/10.1186/s12977-016-0324-3

De Oliveira,F., Cappy,P., Lemee,V., 2016. HIV-1/MO recombinants detection in France: impact on diagnosis and therapeutic management. Unpublished direct submission to Genbank.

de Silva, T.I., Turner, R., Hué, S., Trikha, R., van Tienen, C., Onyango, C., Jaye, A., Foley, B., Whittle, H., Rowland-Jones, S.L., Cotten, M., 2010. HIV-1 subtype distribution in the Gambia and the significant presence of CRF49_cpx, a novel circulating recombinant form. Retrovirology 7, 82. https://doi.org/10.1186/1742-4690-7-82

Deymier, M.J., Ende, Z., Fenton-May, A.E., Dilernia, D.A., Kilembe, W., Allen, S.A., Borrow, P., Hunter, E., 2015. Heterosexual Transmission of Subtype C HIV-1 Selects Consensus-Like Variants without Increased Replicative Capacity or Interferon-α Resistance. PLoS Pathog. 11, e1005154. https://doi.org/10.1371/journal.ppat.1005154

Diop-Ndiaye, H., Toure-Kane, C., Leye, N., Ngom-Gueye, N.F., Montavon, C., Peeters, M., Mboup, S., 2010. Antiretroviral drug resistance mutations in antiretroviral-naive patients from Senegal. AIDS Res. Hum. Retroviruses 26, 1133–1138. https://doi.org/10.1089/aid.2009.0295

Dowling, W.E., Kim, B., Mason, C.J., Wasunna, K.M., Alam, U., Elson, L., Birx, D.L., Robb, M.L., McCutchan, F.E., Carr, J.K., 2002. Forty-one near full-length HIV-1 sequences from Kenya reveal an epidemic of subtype A and A-containing recombinants. AIDS 16, 1809–1820. https://doi.org/10.1097/00002030-200209060-00015

Engelbrecht, S., de Villiers, T., Sampson, C.C., zur Megede, J., Barnett, S.W., van Rensburg, E.J., 2001. Genetic analysis of the complete gag and env genes of HIV type 1 subtype C primary isolates from South Africa. AIDS Res. Hum. Retroviruses 17, 1533–1547. https://doi.org/10.1089/08892220152644241

Fang, G., Kuiken, C., Weiser, B., Rowland-Jones, S., Plummer, F., Chen, C.-H., Kaul, R., Anzala, A.O., Bwayo, J., Kimani, J., Philpott, S.M., Kitchen, C., Sinsheimer, J.S., Gaschen, B., Lang, D., Shi, B., Kemal, K.S., Rostron, T., Brunner, C., Beddows, S., Sattenau, Q., Paxinos, E., Oyugi, J., Burger, H., 2004a. Long-term survivors in Nairobi: complete HIV-1 RNA sequences and immunogenetic associations. J. Infect. Dis. 190, 697–701. https://doi.org/10.1086/421504

Fang, G., Weiser, B., Kuiken, C., Philpott, S.M., Rowland-Jones, S., Plummer, F., Kimani, J., Shi, B., Kaul, R., Bwayo, J., Anzala, O., Burger, H., 2004b. Recombination following superinfection by HIV-1. AIDS 18, 153–159. https://doi.org/10.1097/00002030-200401230-00003

Frahm, N., Nickle, D.C., Linde, C.H., Cohen, D.E., Zuñiga, R., Lucchetti, A., Roach, T., Walker, B.D., Allen, T.M., Korber, B.T., Mullins, J.I., Brander, C., 2008. Increased detection of HIV-specific T cell responses by combination of central sequences with comparable immunogenicity. AIDS 22, 447–456. https://doi.org/10.1097/QAD.0b013e3282f42412

Gao, F., 2002. Complete genomes of HIV-1 subtype C viruses in South Africa. Unpublished direct submission to Genbank.

Gao, F., Robertson, D.L., Carruthers, C.D., Li, Y., Bailes, E., Kostrikis, L.G., Salminen, M.O., Bibollet-Ruche, F., Peeters, M., Ho, D.D., Shaw, G.M., Sharp, P.M., Hahn, B.H., 1998a. An isolate of human immunodeficiency virus type 1 originally classified as subtype I represents a complex mosaic comprising three different group M subtypes (A, G, and I). J. Virol. 72, 10234–10241.

Gao, F., Robertson, D.L., Carruthers, C.D., Morrison, S.G., Jian, B., Chen, Y., Barré-Sinoussi, F., Girard, M., Srinivasan, A., Abimiku, A.G., Shaw, G.M., Sharp, P.M., Hahn, B.H., 1998b. A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. J. Virol. 72, 5680–5698.

Gao, F., Robertson, D.L., Morrison, S.G., Hui, H., Craig, S., Decker, J., Fultz, P.N., Girard, M., Shaw, G.M.,

Hahn, B.H., Sharp, P.M., 1996. The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. J. Virol. 70, 7013–7029.

Gao, F., Trask, S.A., Hui, H., Mamaeva, O., Chen, Y., Theodore, T.S., Foley, B.T., Korber, B.T., Shaw, G.M., Hahn, B.H., 2001a. Molecular characterization of a highly divergent HIV type 1 isolate obtained early in the AIDS epidemic from the Democratic Republic of Congo. AIDS Res. Hum. Retroviruses 17, 1217–1222. https://doi.org/10.1089/088922201316912826

Gao, F., Vidal, N., Li, Y., Trask, S.A., Chen, Y., Kostrikis, L.G., Ho, D.D., Kim, J., Oh, M.D., Choe, K., Salminen, M., Robertson, D.L., Shaw, G.M., Hahn, B.H., Peeters, M., 2001b. Evidence of two distinct subsubtypes within the HIV-1 subtype A radiation. AIDS Res. Hum. Retroviruses 17, 675–688. https://doi.org/10.1089/088922201750236951

Gao, F., Yue, L., Craig, S., Thornton, C.L., Robertson, D.L., McCutchan, F.E., Bradac, J.A., Sharp, P.M., Hahn, B.H., 1994. Genetic variation of HIV type 1 in four World Health Organization-sponsored vaccine evaluation sites: generation of functional envelope (glycoprotein 160) clones representative of sequence subtypes A, B, C, and E. WHO Network for HIV Isolation and Characterization. AIDS Res. Hum. Retroviruses 10, 1359–1368. https://doi.org/10.1089/aid.1994.10.1359

Garcia-Knight, M.A., Slyker, J., Payne, B.L., Pond, S.L.K., de Silva, T.I., Chohan, B., Khasimwa, B., Mbori-Ngacha, D., John-Stewart, G., Rowland-Jones, S.L., Esbjörnsson, J., 2016. Viral Evolution and Cytotoxic T Cell Restricted Selection in Acute Infant HIV-1 Infection. Sci Rep 6, 29536. https://doi.org/10.1038/srep29536

Geldmacher, C., Metzler, I.S., Tovanabutra, S., Asher, T.E., Gostick, E., Ambrozak, D.R., Petrovas, C., Schuetz, A., Ngwenyama, N., Kijak, G., Maboko, L., Hoelscher, M., McCutchan, F., Price, D.A., Douek, D.C., Koup, R.A., 2009. Minor viral and host genetic polymorphisms can dramatically impact the biologic outcome of an epitope-specific CD8 T-cell response. Blood 114, 1553–1562. https://doi.org/10.1182/blood-2009-02-206193

Gerhardt, M., Mloka, D., Tovanabutra, S., Sanders-Buell, E., Hoffmann, O., Maboko, L., Mmbando, D., Birx, D.L., McCutchan, F.E., Hoelscher, M., 2005. In-depth, longitudinal analysis of viral quasispecies from an individual triply infected with late-stage human immunodeficiency virus type 1, using a multiple PCR primer approach. J. Virol. 79, 8249–8261. https://doi.org/10.1128/JVI.79.13.8249-8261.2005

Giandhari, J., Basson, A.E., Coovadia, A., Kuhn, L., Abrams, E.J., Strehlau, R., Morris, L., Hunt, G.M., 2015. Genetic Changes in HIV-1 Gag-Protease Associated with Protease Inhibitor-Based Therapy Failure in Pediatric Patients. AIDS Res. Hum. Retroviruses 31, 776–782. https://doi.org/10.1089/AID.2014.0349

Gounder, K., Oyaro, M., Padayachi, N., Zulu, T.M., de Oliveira, T., Wylie, J., Ndung'u, T., 2017. Complex Subtype Diversity of HIV-1 Among Drug Users in Major Kenyan Cities. AIDS Res. Hum. Retroviruses 33, 500–510. https://doi.org/10.1089/AID.2016.0321

Gounder, K., Padayachi, N., Mann, J.K., Radebe, M., Mokgoro, M., van der Stok, M., Mkhize, L., Mncube, Z., Jaggernath, M., Reddy, T., Walker, B.D., Ndung'u, T., 2015. High frequency of transmitted HIV-1 Gag HLA class I-driven immune escape variants but minimal immune selection over the first year of clade C infection. PLoS ONE 10, e0119886. https://doi.org/10.1371/journal.pone.0119886

Grisson, R.D., Chenine, A.-L., Yeh, L.-Y., He, J., Wood, C., Bhat, G.J., Xu, W., Kankasa, C., Ruprecht, R.M., 2004. Infectious molecular clone of a recently transmitted pediatric human immunodeficiency virus clade C isolate from Africa: evidence of intraclade recombination. J. Virol. 78, 14066–14069. https://doi.org/10.1128/JVI.78.24.14066-14069.2004

Guimarães, M.L., Vicente, A.C.P., Otsuki, K., da Silva, R.F.F.C., Francisco, M., da Silva, F.G., Serrano, D., Morgado, M.G., Bello, G., 2009. Close phylogenetic relationship between Angolan and Romanian HIV-1 subtype F1 isolates. Retrovirology 6, 39. https://doi.org/10.1186/1742-4690-6-39

Gupta, R.K., Kohli, A., McCormick, A.L., Towers, G.J., Pillay, D., Parry, C.M., 2010. Full-length HIV-1 Gag determines protease inhibitor susceptibility within in vitro assays. AIDS 24, 1651–1655.

https://doi.org/10.1097/qad.0b013e3283398216
Gürtler, L.G., Hauser, P.H., Eberle, J., von Brunn, A., Knapp, S., Zekeng, L., Tsague, J.M., Kaptue, L., 1994. A new subtype of human immunodeficiency virus type 1 (MVP-5180) from Cameroon. J. Virol. 68, 1581–1585.
Hackett, J., Zekeng, L., Brennan, C.A., Lund, J.K., Vallari, A.S., Hickman, R.K., Gürtler, L., Kaptué, L., Devare, S.G., 1997. Genetic analysis of HIV type 1 group O p24gag sequences from Cameroon and Equatorial Guinea. AIDS Res. Hum. Retroviruses 13, 1155–1158. https://doi.org/10.1089/aid.1997.13.1155
Harris, M.E., Serwadda, D., Sewankambo, N., Kim, B., Kigozi, G., Kiwanuka, N., Phillips, J.B., Wabwire, F., Meehen, M., Lutalo, T., Lane, J.R., Merling, R., Gray, R., Wawer, M., Birx, D.L., Robb, M.L., McCutchan, F.E., 2002. Among 46 near full length HIV type 1 genome sequences from Rakai District, Uganda, subtype D and AD recombinants predominate. AIDS Res. Hum. Retroviruses 18, 1281–1290. https://doi.org/10.1089/088922202320886325
Heipertz, R.A., Ayemoba, O., Sanders-Buell, E., Poltavee, K., Pham, P., Kijak, G.H., Lei, E., Bose, M., Howell, S., O'Sullivan, A.M., Bates, A., Cervenka, T., Kuroiwa, J., Akintunde, A., Ibezim, O., Alabi, A., Okoye, O., Manak, M., Malia, J., Peel, S., Maisaka, M., Singer, D., O'Connell, R.J., Robb, M.L., Kim, J.H., Michael, N.L., Njoku, O., Tovanabutra, S., 2016. Significant contribution of subtype G to HIV-1 genetic complexity in Nigeria identified by a newly developed subtyping assay specific for subtype G and CRF02_AG. Medicine (Baltimore) 95, e4346. https://doi.org/10.1097/MD.0000000000004346
Hertz, T., Logan, M.G., Rolland, M., Magaret, C.A., Rademeyer, C., Fiore-Gartland, A., Edlefsen, P.T., DeCamp, A., Ahmed, H., Ngandu, N., Larsen, B.B., Frahm, N., Marais, J., Thebus, R., Geraghty, D., Hural, J., Corey, L., Kublin, J., Gray, G., McElrath, M.J., Mullins, J.I., Gilbert, P.B., Williamson, C., 2016. A study of vaccine-induced immune pressure on breakthrough infections in the Phambili phase 2b HIV-1 vaccine efficacy trial. Vaccine 34, 5792–5801. https://doi.org/10.1016/j.vaccine.2016.09.054
HIV type 1 variation in World Health Organization-sponsored vaccine evaluation sites: genetic screening, sequence analysis, and preliminary biological characterization of selected viral strains. WHO Network for HIV Isolation and Characterization, 1994. . AIDS Res. Hum. Retroviruses 10, 1327–1343. https://doi.org/10.1089/aid.1994.10.1327
Hoelscher, M., Kim, B., Maboko, L., Mhalu, F., von Sonnenburg, F., Birx, D.L., McCutchan, F.E., UNAIDS Network for HIV Isolation and Characterization, 2001. High proportion of unrelated HIV-1 intersubtype recombinants in the Mbeya region of southwest Tanzania. AIDS 15, 1461–1470. https://doi.org/10.1097/00002030-200108170-00002
Holzmayer, V., Zekeng, L., Kaptué, L., Gürtler, L., Devare, S.G., Hackett, J., 2005. Near-full-length genomic sequence of a human immunodeficiency type 1 subtype G strain from Cameroon. AIDS Res. Hum. Retroviruses 21, 414–419. https://doi.org/10.1089/aid.2005.21.414
Hora, B., Chen,Y., Denny,T.N., Hewlett,I., Ragupathy,V., Gao,F., 2013. Characterization of near full-length genome sequences for standard panels of HIV-1 isolates established at the External Quality Assurance Program Oversight Laboratory (EQAPOL). Unpublished direct submission to Genbank.
Hora, B., Keating, S.M., Chen, Y., Sanchez, A.M., Sabino, E., Hunt, G., Ledwaba, J., Hackett, J., Swanson, P., Hewlett, I., Ragupathy, V., Vikram Vemula, S., Zeng, P., Tee, K.-K., Chow, W.Z., Ji, H., Sandstrom, P., Denny, T.N., Busch, M.P., Gao, F., REDS-III and EQAPOL programs, 2016. Genetic Characterization of a Panel of Diverse HIV-1 Isolates at Seven International Sites. PLoS ONE 11, e0157340. https://doi.org/10.1371/journal.pone.0157340
Howard, T.M., Olaleye, D.O., Rasheed, S., 1995. Molecular characterization of the gp120 encoding region of a new subtype A strain of HIV-1 (HIV-1IbNg) from Nigeria. Unpublished direct submission to Genbank.
Howard, T.M., Olaylele, D.O., Rasheed, S., 1994. Sequence analysis of the glycoprotein 120 coding region of a new HIV type 1 subtype A strain (HIV-1IbNg) from Nigeria. AIDS Res. Hum. Retroviruses 10,

1755–1757. https://doi.org/10.1089/aid.1994.10.1755
Howard, T.M., Rasheed, S., 1996. Genomic structure and nucleotide sequence analysis of a new HIV type 1 subtype A strain from Nigeria. AIDS Res. Hum. Retroviruses 12, 1413–1425. https://doi.org/10.1089/aid.1996.12.1413
Huang, K.-H.G., Goedhals, D., Carlson, J.M., Brockman, M.A., Mishra, S., Brumme, Z.L., Hickling, S., Tang, C.S.W., Miura, T., Seebregts, C., Heckerman, D., Ndung'u, T., Walker, B., Klenerman, P., Steyn, D., Goulder, P., Phillips, R., Bloemfontein-Oxford Collaborative Group, van Vuuren, C., Frater, J., 2011. Progression to AIDS in South Africa is associated with both reverting and compensatory viral mutations. PLoS ONE 6, e19018. https://doi.org/10.1371/journal.pone.0019018
Huang, K.-H.G., Goedhals, D., Fryer, H., van Vuuren, C., Katzourakis, A., De Oliveira, T., Brown, H., Cassol, S., Seebregts, C., McLean, A., Klenerman, P., Phillips, R., Frater, J., Bloemfontein-Oxford Collaborative Group, 2009. Prevalence of HIV type-1 drug-associated mutations in pre-therapy patients in the Free State, South Africa. Antivir. Ther. (Lond.) 14, 975–984. https://doi.org/10.3851/IMP1416
Huet, T., Dazza, M.C., Brun-Vézinet, F., Roelants, G.E., Wain-Hobson, S., 1989. A highly defective HIV-1 strain isolated from a healthy Gabonese individual presenting an atypical western blot. AIDS 3, 707–715. https://doi.org/10.1097/00002030-198911000-00004
Hunt, G.M., Papathanasopoulos, M.A., Gray, G.E., Tiemessen, C.T., 2003. Characterisation of near-full length genome sequences of three South African human immunodeficiency virus type 1 subtype C isolates. Virus Genes 26, 49–56.
Iwabu, Y., Kinomoto, M., Tatsumi, M., Fujita, H., Shimura, M., Tanaka, Y., Ishizaka, Y., Nolan, D., Mallal, S., Sata, T., Tokunaga, K., 2010. Differential anti-APOBEC3G activity of HIV-1 Vif proteins derived from different subtypes. J. Biol. Chem. 285, 35350–35358. https://doi.org/10.1074/jbc.M110.173286
Iweriebor, B.C., Bessong, P.O., Mavhandu, L.G., Masebe, T.M., Nwobegahay, J., Moyo, S.R., Mphahlele, J.M., 2011. Genetic analysis of the near full-length genome of an HIV type 1 A1/C unique recombinant form from northern South Africa. AIDS Res. Hum. Retroviruses 27, 911–915. https://doi.org/10.1089/aid.2010.0286
Iweriebor, B.C., Masebe, T., Nwobegahay, J., Mphahlele, J.M., Bessong, P.O., 2010. Genetic Characterization of HIV-1 Unique C/CRF11_cpx Recombinant Form from Limpopo Province, South Africa. Unpublished direct submission to Genbank.
Iyer, S.S., Bibollet-Ruche, F., Sherrill-Mix, S., Learn, G.H., Plenderleith, L., Smith, A.G., Barbian, H.J., Russell, R.M., Gondim, M.V.P., Bahari, C.Y., Shaw, C.M., Li, Y., Decker, T., Haynes, B.F., Shaw, G.M., Sharp, P.M., Borrow, P., Hahn, B.H., 2017. Resistance to type 1 interferons is a major determinant of HIV-1 transmission fitness. Proc. Natl. Acad. Sci. U.S.A. 114, E590–E599. https://doi.org/10.1073/pnas.1620144114
Jacobs, G.B., Bock, S., Schuch, A., Moschall, R., Schrom, E.-M., Zahn, J., Reuter, C., Preiser, W., Rethwilm, A., Engelbrecht, S., Kerkau, T., Bodem, J., 2012. Construction of a high titer infectious HIV-1 subtype C proviral clone from South Africa. Viruses 4, 1830–1843. https://doi.org/10.3390/v4091830
Jacobs, G.B., Loxton, A.G., Laten, A., Engelbrecht, S., 2007. Complete genome sequencing of a non-syncytium-inducing HIV type 1 subtype D strain from Cape Town, South Africa. AIDS Res. Hum. Retroviruses 23, 1575–1578. https://doi.org/10.1089/aid.2007.0167
Janssens, W., Heyndrickx, L., Fransen, K., Motte, J., Peeters, M., Nkengasong, J.N., Ndumbe, P.M., Delaporte, E., Perret, J.L., Atende, C., 1994. Genetic and phylogenetic analysis of env subtypes G and H in central Africa. AIDS Res. Hum. Retroviruses 10, 877–879. https://doi.org/10.1089/aid.1994.10.877
Janssens, W., Laukkanen, T., Salminen, M.O., Carr, J.K., Van der Auwera, G., Heyndrickx, L., van der Groen, G., McCutchan, F.E., 2000. HIV-1 subtype H near-full length genome reference strains and analysis of subtype-H-containing inter-subtype recombinants. AIDS 14, 1533–1543. https://doi.org/10.1097/00002030-200007280-00009

Kahle, E., Campbell, M., Lingappa, J., Donnell, D., Celum, C., Ondondo, R., Mujugira, A., Fife, K., Mugo, N., Kapiga, S., Mullins, J.I., Baeten, J.M., Partners in Prevention HSVHIV Transmission Study Team, 2014. HIV-1 subtype C is not associated with higher risk of heterosexual HIV-1 transmission: a multinational study among HIV-1 serodiscordant couples. AIDS 28, 235–243. https://doi.org/10.1097/QAD.0000000000000024

K"Aluoch, O., Donovan, R., Barugahare, B., Mugyenyi, P., Sheppard, H., Cao, H., 2005. High Frequency of Recombinant A and D HIV-1 subtypes in a Population from Uganda. Unpublished direct submission to Genbank.

Kandathil, A.J., Kannangai, R., Abraham, O.C., Rupali, P., Pulimood, S.A., Verghese, V.P., Grant, P., Pillay, D., Sridharan, G., 2009. The frequency of HIV-I drug resistance mutations among treatment-naive individuals at a tertiary care centre in south India. Int J STD AIDS 20, 522–526. https://doi.org/10.1258/ijsa.2008.008403

Kane, C.T., Montavon, C., Toure, M.A., Faye, M.A., Ndiaye, A.G., Diallo, A.G., Ndoye, I., Liegeois, F., Delaporte, E., Mboup, S., Peeters, M., 2001. Full-length genome sequencing of HIV type 1 group O viruses isolated from a heterosexual transmission cluster in Senegal. AIDS Res. Hum. Retroviruses 17, 1211–1216. https://doi.org/10.1089/088922201316912817

Kantor, R., Katzenstein, D.A., Efron, B., Carvalho, A.P., Wynhoven, B., Cane, P., Clarke, J., Sirivichayakul, S., Soares, M.A., Snoeck, J., Pillay, C., Rudich, H., Rodrigues, R., Holguin, A., Ariyoshi, K., Bouzas, M.B., Cahn, P., Sugiura, W., Soriano, V., Brigido, L.F., Grossman, Z., Morris, L., Vandamme, A.-M., Tanuri, A., Phanuphak, P., Weber, J.N., Pillay, D., Harrigan, P.R., Camacho, R., Schapiro, J.M., Shafer, R.W., 2005. Impact of HIV-1 subtype and antiretroviral therapy on protease and reverse transcriptase genotype: results of a global collaboration. PLoS Med. 2, e112. https://doi.org/10.1371/journal.pmed.0020112

Kawashima, Y., Pfafferott, K., Frater, J., Matthews, P., Payne, R., Addo, M., Gatanaga, H., Fujiwara, M., Hachiya, A., Koizumi, H., Kuse, N., Oka, S., Duda, A., Prendergast, A., Crawford, H., Leslie, A., Brumme, Z., Brumme, C., Allen, T., Brander, C., Kaslow, R., Tang, J., Hunter, E., Allen, S., Mulenga, J., Branch, S., Roach, T., John, M., Mallal, S., Ogwu, A., Shapiro, R., Prado, J.G., Fidler, S., Weber, J., Pybus, O.G., Klenerman, P., Ndung'u, T., Phillips, R., Heckerman, D., Harrigan, P.R., Walker, B.D., Takiguchi, M., Goulder, P., 2009. Adaptation of HIV-1 to human leukocyte antigen class I. Nature 458, 641–645. https://doi.org/10.1038/nature07746

Kemal, K.S., Anastos, K., Weiser, B., Ramirez, C.M., Shi, Q., Burger, H., 2013. Molecular epidemiology of HIV type 1 subtypes in Rwanda. AIDS Res. Hum. Retroviruses 29, 957–962. https://doi.org/10.1089/AID.2012.0095

Kiepiela, P., Leslie, A.J., Honeyborne, I., Ramduth, D., Thobakgale, C., Chetty, S., Rathnavalu, P., Moore, C., Pfafferott, K.J., Hilton, L., Zimbwa, P., Moore, S., Allen, T., Brander, C., Addo, M.M., Altfeld, M., James, I., Mallal, S., Bunce, M., Barber, L.D., Szinger, J., Day, C., Klenerman, P., Mullins, J., Korber, B., Coovadia, H.M., Walker, B.D., Goulder, P.J.R., 2004. Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. Nature 432, 769–775. https://doi.org/10.1038/nature03113

Kiguoya, M.W., Mann, J.K., Chopera, D., Gounder, K., Lee, G.Q., Hunt, P.W., Martin, J.N., Ball, T.B., Kimani, J., Brumme, Z.L., Brockman, M.A., Ndung'u, T., 2017. Subtype-Specific Differences in Gag-Protease-Driven Replication Capacity Are Consistent with Intersubtype Differences in HIV-1 Disease Progression. J. Virol. 91. https://doi.org/10.1128/JVI.00253-17

Kijak, G.H., Sanders-Buell, E., Chenine, A.-L., Eller, M.A., Goonetilleke, N., Thomas, R., Leviyang, S., Harbolick, E.A., Bose, M., Pham, P., Oropeza, C., Poltavee, K., O'Sullivan, A.M., Billings, E., Merbah, M., Costanzo, M.C., Warren, J.A., Slike, B., Li, H., Peachman, K.K., Fischer, W., Gao, F., Cicala, C., Arthos, J., Eller, L.A., O'Connell, R.J., Sinei, S., Maganga, L., Kibuuka, H., Nitayaphan, S., Rao, M., Marovich, M.A., Krebs, S.J., Rolland, M., Korber, B.T., Shaw, G.M., Michael, N.L., Robb, M.L., Tovanabutra, S., Kim, J.H., 2017. Rare HIV-1 transmitted/founder lineages identified by deep viral sequencing contribute to rapid shifts in dominant quasispecies during acute and early infection. PLoS Pathog. 13, e1006510. https://doi.org/10.1371/journal.ppat.1006510

Kijak, G.H., Sanders-Buell, E., Wolfe, N.D., Mpoudi-Ngole, E., Kim, B., Brown, B., Robb, M.L., Birx, D.L., Burke, D.S., Carr, J.K., McCutchan, F.E., 2004. Development and application of a high-throughput HIV type 1 genotyping assay to identify CRF02_AG in West/West Central Africa. AIDS Res. Hum. Retroviruses 20, 521–530. https://doi.org/10.1089/088922204323087778

Klatt, N.R., Hawkins, P., Lumm, W., Farmer, P., Bansal, A., Mulenga, J., Yue, L., Allen, S., Derdeyn, C.A., Goepfert, P.A., Hunter, E., 2011. Transmission of Homogenous Gag after HIV Infection is not dependent on pressure from CTL escape. Unpublished direct submission to Genbank.

Koofhethile, C.K., Ndhlovu, Z.M., Thobakgale-Tshabalala, C., Prado, J.G., Ismail, N., Mncube, Z., Mkhize, L., van der Stok, M., Yende, N., Walker, B.D., Goulder, P.J.R., Ndung'u, T., 2016. CD8+ T Cell Breadth and Ex Vivo Virus Inhibition Capacity Distinguish between Viremic Controllers with and without Protective HLA Class I Alleles. J. Virol. 90, 6818–6831. https://doi.org/10.1128/JVI.00276-16

Koulinska, I.N., Msamanga, G., Mwakagile, D., Essex, M., Renjifo, B., 2002. Common genetic arrangements among human immunodeficiency virus type 1 subtype A and D recombinant genomes vertically transmitted in Tanzania. AIDS Res. Hum. Retroviruses 18, 947–956. https://doi.org/10.1089/088922202760265614

Koulinska, I.N., Ndung'u, T., Mwakagile, D., Msamanga, G., Kagoma, C., Fawzi, W., Essex, M., Renjifo, B., 2001. A new human immunodeficiency virus type 1 circulating recombinant form from Tanzania. AIDS Res. Hum. Retroviruses 17, 423–431. https://doi.org/10.1089/088922201750102508

Kusagawa, S., Takebe, Y., Yang, R., Motomura, K., Ampofo, W., Brandful, J., Koyanagi, Y., Yamamoto, N., Sata, T., Ishikawa, K., Hachiya, A., Oka, S., Nagai, Y., Tatsumi, M., 2000. Isolation and characterization of full-length molecular DNA clones of Ghanaian HIV-1 intersubtype A/G recombinant (CRF02_AG) which is replication-competent in restricted host-range. Unpublished direct submission to Genbank.

Kusagawa, S., Takebe, Y., Yang, R., Motomura, K., Ampofo, W., Brandful, J., Koyanagi, Y., Yamamoto, N., Sata, T., Ishikawa, K., Nagai, Y., Tatsumi, M., 2001. Isolation and characterization of a full-length molecular DNA clone of Ghanaian HIV type 1 intersubtype A/G recombinant CRF02_AG, which is replication competent in a restricted host range. AIDS Res. Hum. Retroviruses 17, 649–655. https://doi.org/10.1089/088922201300119761

Land, A.M., Ball, T.B., Luo, M., Rutherford, J., Sarna, C., Wachihi, C., Kimani, J., Plummer, F.A., 2008. Full-length HIV type 1 proviral sequencing of 10 highly exposed women from Nairobi, Kenya reveals a high proportion of intersubtype recombinants. AIDS Res. Hum. Retroviruses 24, 865–872. https://doi.org/10.1089/aid.2007.0200

Leitman, E.M., Willberg, C.B., Tsai, M.-H., Chen, H., Buus, S., Chen, F., Riddell, L., Haas, D., Fellay, J., Goedert, J.J., Piechocka-Trocha, A., Walker, B.D., Martin, J., Deeks, S., Wolinsky, S.M., Martinson, J., Martin, M., Qi, Y., Sáez-Cirión, A., Yang, O.O., Matthews, P.C., Carrington, M., Goulder, P.J.R., 2017. HLA-B*14:02-Restricted Env-Specific CD8+ T-Cell Activity Has Highly Potent Antiviral Efficacy Associated with Immune Control of HIV Infection. J. Virol. 91. https://doi.org/10.1128/JVI.00544-17

Liang, B., Luo, M., Scott-Herridge, J., Semeniuk, C., Mendoza, M., Capina, R., Sheardown, B., Ji, H., Kimani, J., Ball, B.T., Van Domselaar, G., Graham, M., Tyler, S., Jones, S.J.M., Plummer, F.A., 2011. A comparison of parallel pyrosequencing and sanger clone-based sequencing and its impact on the characterization of the genetic diversity of HIV-1. PLoS ONE 6, e26745. https://doi.org/10.1371/journal.pone.0026745

Liégeois, F., Boué, V., Butel, C., Mouinga-Ondémé, A., Sica, J., Zamba, C., Peeters, M., Delaporte, E., Rouet, F., 2013. HIV type-1 group O infection in Gabon: low prevalence rate but circulation of genetically diverse and drug-resistant HIV type-1 group O strains. AIDS Res. Hum. Retroviruses 29, 1085–1090. https://doi.org/10.1089/aid.2012.0375

Liu, M.K.P., Hawkins, N., Ritchie, A.J., Ganusov, V.V., Whale, V., Brackenridge, S., Li, H., Pavlicek, J.W., Cai, F., Rose-Abrahams, M., Treurnicht, F., Hraber, P., Riou, C., Gray, C., Ferrari, G., Tanner, R., Ping, L.-H., Anderson, J.A., Swanstrom, R., CHAVI Core B, Cohen, M., Karim, S.S.A., Haynes, B.,

Borrow, P., Perelson, A.S., Shaw, G.M., Hahn, B.H., Williamson, C., Korber, B.T., Gao, F., Self, S., McMichael, A., Goonetilleke, N., 2013. Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. J. Clin. Invest. 123, 380–393. https://doi.org/10.1172/JCI65330

Louwagie, J., Delwart, E.L., Mullins, J.I., McCutchan, F.E., Eddy, G., Burke, D.S., 1994. Genetic analysis of HIV-1 isolates from Brazil reveals presence of two distinct genetic subtypes. AIDS Res. Hum. Retroviruses 10, 561–567. https://doi.org/10.1089/aid.1994.10.561

Louwagie, J., McCutchan, F.E., Peeters, M., Brennan, T.P., Sanders-Buell, E., Eddy, G.A., van der Groen, G., Fransen, K., Gershy-Damet, G.M., Deleys, R., 1993. Phylogenetic analysis of gag genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. AIDS 7, 769–780. https://doi.org/10.1097/00002030-199306000-00003

Loxton, A.G., Treurnicht, F., Laten, A., van Rensburg, E.J., Engelbrecht, S., 2005. Sequence analysis of near full-length HIV type 1 subtype D primary strains isolated in Cape Town, South Africa, from 1984 to 1986. AIDS Res. Hum. Retroviruses 21, 410–413. https://doi.org/10.1089/aid.2005.21.410

Luk, K.-C., Berg, M.G., Naccache, S.N., Kabre, B., Federman, S., Mbanya, D., Kaptué, L., Chiu, C.Y., Brennan, C.A., Hackett, J., 2015. Utility of Metagenomic Next-Generation Sequencing for Characterization of HIV and Human Pegivirus Diversity. PLoS ONE 10, e0141723. https://doi.org/10.1371/journal.pone.0141723

Luk, K.-C., Holzmayer, V., Ndembi, N., Swanson, P., Brennan, C.A., Ngansop, C., Mbanya, D., Kaptué, L., Gürtler, L., Devare, S.G., Hackett, J., 2008. Near full-length genome characterization of an HIV type 1 CRF25_cpx strain from Cameroon. AIDS Res. Hum. Retroviruses 24, 1309–1314. https://doi.org/10.1089/aid.2008.0124

Luk, K.-C., Holzmayer, V., Yamaguchi, J., Swanson, P., Brennan, C.A., Ngansop, C., Mbanya, D., Gayum, H., Djuidje, M.N., Ndembi, N., Kamdem, D., Kaptué, L., Gürtler, L., Devare, S.G., Hackett, J., 2007. Near full-length genome characterization of three additional HIV type 1 CRF13_cpx strains from Cameroon. AIDS Res. Hum. Retroviruses 23, 297–302. https://doi.org/10.1089/aid.2006.0177

Mahasha, P.W., Cassol, E., Danaviah, S., Bond, R., Seebregts, C., van der Merwe, S.W., Rossouw, T., de Oliveira, T., 2013. Differential Distribution and Evolution of HIV-1 RNA Variants in the Gastrointestinal Tract of Antiretroviral Naive African AIDS Patients with Diarrhea and (or) Weight Loss. Unpublished direct submission to Genbank.

Mamadou, S., Montavon, C., Ben, A., Djibo, A., Rabiou, S., Mboup, S., Delaporte, E., Peeters, M., 2002. Predominance of CRF02-AG and CRF06-cpx in Niger, West Africa. AIDS Res. Hum. Retroviruses 18, 723–726. https://doi.org/10.1089/088922202760072357

Mamadou, S., Vidal, N., Montavon, C., Ben, A., Djibo, A., Rabiou, S., Soga, G., Delaporte, E., Mboup, S., Peeters, M., 2003. Emergence of complex and diverse CRF02-AG/CRF06-cpx recombinant HIV type 1 strains in Niger, West Africa. AIDS Res. Hum. Retroviruses 19, 77–82. https://doi.org/10.1089/08892220360474005

Manak, M., Sina, S., Anekella, B., Hewlett, I., Sanders-Buell, E., Ragupathy, V., Kim, J., Vermeulen, M., Stramer, S.L., Sabino, E., Grabarczyk, P., Michael, N., Peel, S., Garrett, P., Tovanabutra, S., Busch, M.P., Schito, M., 2012. Pilot studies for development of an HIV subtype panel for surveillance of global diversity. AIDS Res. Hum. Retroviruses 28, 594–606. https://doi.org/10.1089/AID.2011.0271

Matthews, P.C., Prendergast, A., Leslie, A., Crawford, H., Payne, R., Rousseau, C., Rolland, M., Honeyborne, I., Carlson, J., Kadie, C., Brander, C., Bishop, K., Mlotshwa, N., Mullins, J.I., Coovadia, H., Ndung'u, T., Walker, B.D., Heckerman, D., Goulder, P.J.R., 2008. Central role of reverting mutations in HLA associations with human immunodeficiency virus set point. J. Virol. 82, 8548–8559. https://doi.org/10.1128/JVI.00580-08

McCutchan, F.E., Hoelscher, M., Tovanabutra, S., Piyasirisilp, S., Sanders-Buell, E., Ramos, G., Jagodzinski, L., Polonis, V., Maboko, L., Mmbando, D., Hoffmann, O., Riedner, G., von Sonnenburg, F., Robb, M., Birx, D.L., 2005. In-depth analysis of a heterosexually acquired human immunodeficiency virus type 1 superinfection: evolution, temporal fluctuation, and intercompartment dynamics from the seronegative window period through 30 months postinfection. J. Virol. 79, 11693–11704.

https://doi.org/10.1128/JVI.79.18.11693-11704.2005
McCutchan, F.E., Sankale, J.-L., M'Boup, S., Kim, B., Tovanabutra, S., Hamel, D.J., Brodine, S.K., Kanki, P.J., Birx, D.L., 2004. HIV type 1 circulating recombinant form CRF09_cpx from west Africa combines subtypes A, F, G, and may share ancestors with CRF02_AG and Z321. AIDS Res. Hum. Retroviruses 20, 819–826. https://doi.org/10.1089/0889222041725163
McCutchan, F.E., Ungar, B.L., Hegerich, P., Roberts, C.R., Fowler, A.K., Hira, S.K., Perine, P.L., Burke, D.S., 1992. Genetic analysis of HIV-1 isolates from Zambia and an expanded phylogenetic tree for HIV-1. J. Acquir. Immune Defic. Syndr. 5, 441–449.
Meloni, S.T., Kim, B., Sankalé, J.-L., Hamel, D.J., Tovanabutra, S., Mboup, S., McCutchan, F.E., Kanki, P.J., 2004. Distinct human immunodeficiency virus type 1 subtype A virus circulating in West Africa: sub-subtype A3. J. Virol. 78, 12438–12445. https://doi.org/10.1128/JVI.78.22.12438-12445.2004
Mokili, J.L.K., Rogers, M., Carr, J.K., Simmonds, P., Bopopi, J.M., Foley, B.T., Korber, B.T., Birx, D.L., McCutchan, F.E., 2002. Identification of a novel clade of human immunodeficiency virus type 1 in Democratic Republic of Congo. AIDS Res. Hum. Retroviruses 18, 817–823. https://doi.org/10.1089/08892220260139567
Mónaco, D.C., Dilernia, D.A., Fiore-Gartland, A., Yu, T., Prince, J.L., Dennis, K.K., Qin, K., Schaefer, M., Claiborne, D.T., Kilembe, W., Tang, J., Price, M.A., Farmer, P., Gilmour, J., Bansal, A., Allen, S., Goepfert, P., Hunter, E., 2016. Balance between transmitted HLA preadapted and nonassociated polymorphisms is a major determinant of HIV-1 disease progression. J. Exp. Med. 213, 2049–2063. https://doi.org/10.1084/jem.20151984
Montavon, C., Bibollet-Ruche, F., Robertson, D., Koumare, B., Mulanga, C., Esu-Williams, E., Toure, C., Mboup, S., Saman, E., Delaporte, E., Peeters, M., 1999. The identification of a complex A/G/I/J recombinant HIV type 1 virus in various West African countries. AIDS Res. Hum. Retroviruses 15, 1707–1712. https://doi.org/10.1089/088922299309757
Montavon, C., Delaporte, E., Peeters, M., 2000. Two new complete genomes of HIV-1 recombinant AGJ BFP90-like circulating in Mali and Senegal. Unpublished direct submission to Genbank.
Montavon, C., Toure-Kane, C., Liegeois, F., Mpoudi, E., Bourgeois, A., Vergne, L., Perret, J.L., Boumah, A., Saman, E., Mboup, S., Delaporte, E., Peeters, M., 2000. Most env and gag subtype A HIV-1 viruses circulating in West and West Central Africa are similar to the prototype AG recombinant virus IBNG. J. Acquir. Immune Defic. Syndr. 23, 363–374. https://doi.org/10.1097/00126334-200004150-00001
Montavon, C., Toure-Kane, C., Nkengasong, J.N., Vergne, L., Hertogs, K., Mboup, S., Delaporte, E., Peeters, M., 2002a. CRF06-cpx: a new circulating recombinant form of HIV-1 in West Africa involving subtypes A, G, K, and J. J. Acquir. Immune Defic. Syndr. 29, 522–530. https://doi.org/10.1097/00126334-200204150-00014
Montavon, C., Vergne, L., Bourgeois, A., Mpoudi-Ngole, E., Malonga-Mouellet, G., Butel, C., Toure-Kane, C., Delaporte, E., Peeters, M., 2002b. Identification of a new circulating recombinant form of HIV type 1, CRF11-cpx, involving subtypes A, G, J, and CRF01-AE, in Central Africa. AIDS Res. Hum. Retroviruses 18, 231–236. https://doi.org/10.1089/08892220252781301
Muenchhoff, M., Madurai, S., Hempenstall, A.J., Adland, E., Carlqvist, A., Moonsamy, A., Jaggernath, M., Mlotshwa, B., Siboto, E., Ndung'u, T., Goulder, P.J.R., 2014. Evaluation of the NucliSens EasyQ v2.0 assay in comparison with the Roche Amplicor v1.5 and the Roche CAP/CTM HIV-1 Test v2.0 in quantification of C-clade HIV-1 in plasma. PLoS ONE 9, e103983. https://doi.org/10.1371/journal.pone.0103983
Murphy, E., Korber, B., Georges-Courbot, M.C., You, B., Pinter, A., Cook, D., Kieny, M.P., Georges, A., Mathiot, C., Barré-Sinoussi, F., 1993. Diversity of V3 region sequences of human immunodeficiency viruses type 1 from the central African Republic. AIDS Res. Hum. Retroviruses 9, 997–1006. https://doi.org/10.1089/aid.1993.9.997
Ndiaye, H.D., Tchiakpe, E., Vidal, N., Ndiaye, O., Diop, A.K., Peeters, M., Mboup, S., Toure-Kane, C., 2013. HIV type 1 subtype C remains the predominant subtype in men having sex with men in Senegal. AIDS Res. Hum. Retroviruses 29, 1265–1272. https://doi.org/10.1089/aid.2013.0140

Ndung'u, T., Renjifo, B., Novitsky, V.A., McLane, M.F., Gaolekwe, S., Essex, M., 2000. Molecular cloning and biological characterization of full-length HIV-1 subtype C from Botswana. Virology 278, 390–399. https://doi.org/10.1006/viro.2000.0583

Neilson, J.R., John, G.C., Carr, J.K., Lewis, P., Kreiss, J.K., Jackson, S., Nduati, R.W., Mbori-Ngacha, D., Panteleeff, D.D., Bodrug, S., Giachetti, C., Bott, M.A., Richardson, B.A., Bwayo, J., Ndinya-Achola, J., Overbaugh, J., 1999. Subtypes of human immunodeficiency virus type 1 and disease stage among women in Nairobi, Kenya. J. Virol. 73, 4393–4403.

Niama, F.R., Toure-Kane, C., Vidal, N., Obengui, P., Bikandou, B., Ndoundou Nkodia, M.Y., Montavon, C., Diop-Ndiaye, H., Mombouli, J.V., Mokondzimobe, E., Diallo, A.G., Delaporte, E., Parra, H.-J., Peeters, M., Mboup, S., 2006. HIV-1 subtypes and recombinants in the Republic of Congo. Infect. Genet. Evol. 6, 337–343. https://doi.org/10.1016/j.meegid.2005.12.001

Niama, F.R., Vidal, N., Bazepeo, S.E., Mpoudi, E., Toure-Kane, C., Parra, H.J., Delaporte, E., Peeters, M., 2009. CRF45_AKU, a circulating recombinant from Central Africa, is probably the common ancestor of HIV type 1 MAL and HIV type 1 NOGIL. AIDS Res. Hum. Retroviruses 25, 1345–1353. https://doi.org/10.1089/aid.2009.0169

Novelli, P., Vella, C., Oxford, J., Daniels, R.S., 2002. Construction and characterization of a full-length HIV-1(92UG001) subtype D infectious molecular clone. AIDS Res. Hum. Retroviruses 18, 85–88. https://doi.org/10.1089/088922202753394754

Novitsky, V., Wang, R., Margolin, L., Baca, J., Kebaabetswe, L., Rossenkhan, R., Bonney, C., Herzig, M., Nkwe, D., Moyo, S., Musonda, R., Woldegabriel, E., van Widenfelt, E., Makhema, J., Lagakos, S., Essex, M., 2009. Timing constraints of in vivo gag mutations during primary HIV-1 subtype C infection. PLoS ONE 4, e7727. https://doi.org/10.1371/journal.pone.0007727

Novitsky, V., Wang, R., Rossenkhan, R., Moyo, S., Essex, M., 2013. Intra-host evolutionary rates in HIV-1C env and gag during primary infection. Infect. Genet. Evol. 19, 361–368. https://doi.org/10.1016/j.meegid.2013.02.023

Novitsky, V.A., Gaolekwe, S., McLane, M.F., Ndung'u, T.P., Foley, B.T., Vannberg, F., Marlink, R., Essex, M., 2000. HIV type 1 A/J recombinant with a pronounced pol gene mosaicism. AIDS Res. Hum. Retroviruses 16, 1015–1020. https://doi.org/10.1089/08892220050058434

Novitsky, V.A., Montano, M.A., McLane, M.F., Renjifo, B., Vannberg, F., Foley, B.T., Ndung'u, T.P., Rahman, M., Makhema, M.J., Marlink, R., Essex, M., 1999. Molecular cloning and phylogenetic analysis of human immunodeficiency virus type 1 subtype C: a set of 23 full-length clones from Botswana. J. Virol. 73, 4427–4432.

Novitsky,V., 2002. HIV-1 subtype C/Denv/C recombinant from Botswana. Unpublished direct submission to Genbank.

Ntale, R.S., Chopera, D.R., Ngandu, N.K., Assis de Rosa, D., Zembe, L., Gamieldien, H., Mlotshwa, M., Werner, L., Woodman, Z., Mlisana, K., Abdool Karim, S., Gray, C.M., Williamson, C., CAPRISA 002 Study Team, 2012. Temporal association of HLA-B*81:01- and HLA-B*39:10-mediated HIV-1 p24 sequence evolution with disease progression. J. Virol. 86, 12013–12024. https://doi.org/10.1128/JVI.00539-12

Obuku, A.E., Bugembe, D.L., Musinguzi, K., Watera, C., Serwanga, J., Ndembi, N., Levin, J., Kaleebu, P., Pala, P., 2016. Macrophage Inflammatory Protein-1 Beta and Interferon Gamma Responses in Ugandans with HIV-1 Acute/Early Infections. AIDS Res. Hum. Retroviruses 32, 237–246. https://doi.org/10.1089/AID.2015.0157

Olaleye, D.O., Sheng, Z., Howard, T.M., Rasheed, S., 1996. Isolation and characterization of a new subtype A variant of human immunodeficiency virus type I from Nigeria. Trop. Med. Int. Health 1, 97–106. https://doi.org/10.1046/j.1365-3156.1996.d01-4.x

Op de Coul, E., van der Schoot, A., Goudsmit, J., van den Burg, R., Janssens, W., Heyndrickx, L., van der Groen, G., Cornelissen, M., 2000. Independent introduction of transmissible F/D recombinant HIV-1 from Africa into Belgium and The Netherlands. Virology 270, 267–277. https://doi.org/10.1006/viro.2000.0247

Oram, J.D., Downing, R.G., Roff, M., Clegg, J.C., Serwadda, D., Carswell, J.W., 1990. Nucleotide sequence

of a Ugandan HIV-1 provirus reveals genetic diversity from other HIV-1 isolates. AIDS Res. Hum. Retroviruses 6, 1073–1078. https://doi.org/10.1089/aid.1990.6.1073

Palm, A.A., Esbjörnsson, J., Månsson, F., Biague, A., da Silva, Z.J., Norrgren, H., Jansson, M., Medstrand, P., 2015. Cocirculation of several similar but unique HIV-1 recombinant forms in Guinea-Bissau revealed by near full-length genomic sequencing. AIDS Res. Hum. Retroviruses 31, 938–945. https://doi.org/10.1089/AID.2015.0073

Pandrea, I., Robertson, D.L., Onanga, R., Gao, F., Makuwa, M., Ngari, P., Bedjabaga, I., Roques, P., Simon, F., Apetrei, C., 2002. Analysis of partial pol and env sequences indicates a high prevalence of HIV type 1 recombinant strains circulating in Gabon. AIDS Res. Hum. Retroviruses 18, 1103–1116. https://doi.org/10.1089/088922202320567842

Papathanasopoulos, M.A., Cilliers, T., Morris, L., Mokili, J.L., Dowling, W., Birx, D.L., McCutchan, F.E., 2002. Full-length genome analysis of HIV-1 subtype C utilizing CXCR4 and intersubtype recombinants isolated in South Africa. AIDS Res. Hum. Retroviruses 18, 879–886. https://doi.org/10.1089/08892220260190362

Papathanasopoulos, M.A., Patience, T., Meyers, T.M., McCutchan, F.E., Morris, L., 2003. Full-length genome characterization of HIV type 1 subtype C isolates from two slow-progressing perinatally infected siblings in South Africa. AIDS Res. Hum. Retroviruses 19, 1033–1037. https://doi.org/10.1089/088922203322588396

Parrish, N.F., Gao, F., Li, H., Giorgi, E.E., Barbian, H.J., Parrish, E.H., Zajic, L., Iyer, S.S., Decker, J.M., Kumar, A., Hora, B., Berg, A., Cai, F., Hopper, J., Denny, T.N., Ding, H., Ochsenbauer, C., Kappes, J.C., Galimidi, R.P., West, A.P., Bjorkman, P.J., Wilen, C.B., Doms, R.W., O'Brien, M., Bhardwaj, N., Borrow, P., Haynes, B.F., Muldoon, M., Theiler, J.P., Korber, B., Shaw, G.M., Hahn, B.H., 2013. Phenotypic properties of transmitted founder HIV-1. Proc. Natl. Acad. Sci. U.S.A. 110, 6626–6633. https://doi.org/10.1073/pnas.1304288110

Parrish, N.F., Wilen, C.B., Banks, L.B., Iyer, S.S., Pfaff, J.M., Salazar-Gonzalez, J.F., Salazar, M.G., Decker, J.M., Parrish, E.H., Berg, A., Hopper, J., Hora, B., Kumar, A., Mahlokozera, T., Yuan, S., Coleman, C., Vermeulen, M., Ding, H., Ochsenbauer, C., Tilton, J.C., Permar, S.R., Kappes, J.C., Betts, M.R., Busch, M.P., Gao, F., Montefiori, D., Haynes, B.F., Shaw, G.M., Hahn, B.H., Doms, R.W., 2012. Transmitted/founder and chronic subtype C HIV-1 use CD4 and CCR5 receptors with equal efficiency and are not inhibited by blocking the integrin α4β7. PLoS Pathog. 8, e1002686. https://doi.org/10.1371/journal.ppat.1002686

Payne, R., Muenchhoff, M., Mann, J., Roberts, H.E., Matthews, P., Adland, E., Hempenstall, A., Huang, K.-H., Brockman, M., Brumme, Z., Sinclair, M., Miura, T., Frater, J., Essex, M., Shapiro, R., Walker, B.D., Ndung'u, T., McLean, A.R., Carlson, J.M., Goulder, P.J.R., 2014. Impact of HLA-driven HIV adaptation on virulence in populations of high HIV seroprevalence. Proc. Natl. Acad. Sci. U.S.A. 111, E5393-5400. https://doi.org/10.1073/pnas.1413339111

Peeters, M., Liegeois, F., Torimiro, N., Bourgeois, A., Mpoudi, E., Vergne, L., Saman, E., Delaporte, E., Saragosti, S., 1999. Characterization of a highly replicative intergroup M/O human immunodeficiency virus type 1 recombinant isolated from a Cameroonian patient. J. Virol. 73, 7368–7375.

Permar, S.R., Salazar, M.G., Gao, F., Cai, F., Learn, G.H., Kalilani, L., Hahn, B.H., Shaw, G.M., Salazar-Gonzalez, J.F., 2013. Clonal amplification and maternal-infant transmission of nevirapine-resistant HIV-1 variants in breast milk following single-dose nevirapine prophylaxis. Retrovirology 10, 88. https://doi.org/10.1186/1742-4690-10-88

Pessôa, R., Watanabe, J.T., Calabria, P., Alencar, C.S., Loureiro, P., Lopes, M.E., Proetti, A.B., Félix, A.C., Sabino, E.C., Busch, M.P., Sanabani, S.S., International Component of the NHLBI Recipient Epidemiology and Donor Evaluation Study-III (REDS-III), 2015. Enhanced detection of viral diversity using partial and near full-length genomes of human immunodeficiency virus Type 1 provirus deep sequencing data from recently infected donors at four blood centers in Brazil. Transfusion 55, 980–990. https://doi.org/10.1111/trf.12936

Peters, H.O., Mendoza, M.G., Capina, R.E., Luo, M., Mao, X., Gubbins, M., Nagelkerke, N.J.D., Macarthur,

I., Sheardown, B.B., Kimani, J., Wachihi, C., Thavaneswaran, S., Plummer, F.A., 2008. An integrative bioinformatic approach for studying escape mutations in human immunodeficiency virus type 1 gag in the Pumwani Sex Worker Cohort. J. Virol. 82, 1980–1992. https://doi.org/10.1128/JVI.02742-06

Pollakis, G., Abebe, A., Kliphuis, A., De Wit, T.F.R., Fisseha, B., Tegbaru, B., Tesfaye, G., Negassa, H., Mengistu, Y., Fontanet, A.L., Cornelissen, M., Goudsmit, J., 2003. Recombination of HIV type 1C (C'/C") in Ethiopia: possible link of EthHIV-1C' to subtype C sequences from the high-prevalence epidemics in India and Southern Africa. AIDS Res. Hum. Retroviruses 19, 999–1008. https://doi.org/10.1089/088922203322588350

Poss, M., Gosink, J., Thomas, E., Kreiss, J.K., Ndinya-Achola, J., Mandaliya, K., Bwayo, J., Overbaugh, J., 1997. Phylogenetic evaluation of Kenyan HIV type 1 isolates. AIDS Res. Hum. Retroviruses 13, 493–499. https://doi.org/10.1089/aid.1997.13.493

Poss, M., Overbaugh, J., 1999. Variants from the diverse virus population identified at seroconversion of a clade A human immunodeficiency virus type 1-infected woman have distinct biological properties. J. Virol. 73, 5255–5264.

Poss, M., Rodrigo, A.G., Gosink, J.J., Learn, G.H., de Vange Panteleeff, D., Martin, H.L., Bwayo, J., Kreiss, J.K., Overbaugh, J., 1998. Evolution of envelope sequences from the genital tract and peripheral blood of women infected with clade A human immunodeficiency virus type 1. J. Virol. 72, 8240–8251.

Powell, R.L.R., Zhao, J., Konings, F.A.J., Tang, S., Ewane, L., Burda, S., Urbanski, M.M., Saa, D.R., Hewlett, I., Nyambi, P.N., 2007a. Circulating recombinant form (CRF) 37_cpx: an old strain in Cameroon composed of diverse, genetically distant lineages of subtypes A and G. AIDS Res. Hum. Retroviruses 23, 923–933. https://doi.org/10.1089/aid.2007.0040

Powell, R.L.R., Zhao, J., Konings, F.A.J., Tang, S., Nanfack, A., Burda, S., Urbanski, M.M., Saa, D.R., Hewlett, I., Nyambi, P.N., 2007b. Identification of a novel circulating recombinant form (CRF) 36_cpx in Cameroon that combines two CRFs (01_AE and 02_AG) with ancestral lineages of subtypes A and G. AIDS Res. Hum. Retroviruses 23, 1008–1019. https://doi.org/10.1089/aid.2006.0289

Price, M.A., Wallis, C.L., Lakhi, S., Karita, E., Kamali, A., Anzala, O., Sanders, E.J., Bekker, L.-G., Twesigye, R., Hunter, E., Kaleebu, P., Kayitenkore, K., Allen, S., Ruzagira, E., Mwangome, M., Mutua, G., Amornkul, P.N., Stevens, G., Pond, S.L.K., Schaefer, M., Papathanasopoulos, M.A., Stevens, W., Gilmour, J., IAVI Early Infection Cohort Study Group, 2011. Transmitted HIV type 1 drug resistance among individuals with recent HIV infection in East and Southern Africa. AIDS Res. Hum. Retroviruses 27, 5–12. https://doi.org/10.1089/aid.2010.0030

Ragupathy,V., Voskanian-Kordi, A., Karagiannis, K., Wang, A., Zhao, J., Mbondji, C., Gao, F., Hora, B., Sanchez, A., Denny,T., Schito, M., Busch, M., Simonyan, V., Hewlett,I., 2016. PCR amplicon free whole genome charecterization of Human immunodeficiency virus type -1 genotypes. Unpublished direct submission to Genbank.

Rodenburg, C.M., Li, Y., Trask, S.A., Chen, Y., Decker, J., Robertson, D.L., Kalish, M.L., Shaw, G.M., Allen, S., Hahn, B.H., Gao, F., UNAIDS and NIAID Networks for HIV Isolation and Characterization, 2001. Near full-length clones and reference sequences for subtype C isolates of HIV type 1 from three different continents. AIDS Res. Hum. Retroviruses 17, 161–168. https://doi.org/10.1089/08892220150217247

Rodgers, M. A., Vallari, A.S., Harris, B., Yamaguchi, J., Holzmayer, V., Forberg, K., Berg, M.G., Kenmenge, J., Ngansop, C., Awazi, B., Mbanya, D., Kaptue, L., Brennan, C., Cloherty, G., Ndembi, N., 2017. Identification of rare HIV-1 Group N, HBV AE, and HTLV-3 strains in rural South Cameroon. Virology 504, 141–151. https://doi.org/10.1016/j.virol.2017.01.008

Rodgers, Mary A., Wilkinson, E., Vallari, A., McArthur, C., Sthreshley, L., Brennan, C.A., Cloherty, G., de Oliveira, T., 2017. Sensitive Next-Generation Sequencing Method Reveals Deep Genetic Diversity of HIV-1 in the Democratic Republic of the Congo. J. Virol. 91. https://doi.org/10.1128/JVI.01841-16

Roques, P., Robertson, D.L., Souquière, S., Apetrei, C., Nerrienet, E., Barré-Sinoussi, F., Müller-Trutwin, M., Simon, F., 2004. Phylogenetic characteristics of three new HIV-1 N strains and implications for the origin of group N. AIDS 18, 1371–1381. https://doi.org/10.1097/01.aids.0000125990.86904.28

Rousseau, C.M., Birditt, B.A., McKay, A.R., Stoddard, J.N., Lee, T.C., McLaughlin, S., Moore, S.W., Shindo, N., Learn, G.H., Korber, B.T., Brander, C., Goulder, P.J.R., Kiepiela, P., Walker, B.D., Mullins, J.I., 2006. Large-scale amplification, cloning and sequencing of near full-length HIV-1 subtype C genomes. J. Virol. Methods 136, 118–125. https://doi.org/10.1016/j.jviromet.2006.04.009

Saina, M.C., Bi, X., Lihana, R., Lwembe, R., Ishizaki, A., Panikulam, A., Palakudy, T., Musoke, R., Owens, M., Songok, E.M., Ichimura, H., 2015. Comparison of HIV-1 nef and gag Variations and Host HLA Characteristics as Determinants of Disease Progression among HIV-1 Vertically Infected Kenyan Children. PLoS ONE 10, e0137140. https://doi.org/10.1371/journal.pone.0137140

Sakamoto, Y., Takekawa, N., Ampofo, W., Sata, T., Tatsumi, M., 2006. Construction and characterization of HIV-1 CRF02_AG infectious molecular clones derived from Ghanaian patients. Unpublished direct submission to Genbank.

Sakamoto,Y., Takekawa, N, Tatsumi, M., 2006a. Construction and characterization of HIV-1 subtype A and CRF01_AE infectious molecular clones. Unpublished direct submission to Genbank.

Sakamoto,Y., Tatsumi,M., Takekawa,N., 2006b. Construction and characterization of naive HIV-1 subtype A infectious molecular clones derived from Rwanda patients for reference. Unpublished direct submission to Genbank.

Salazar-Gonzalez, J.F., Salazar, M.G., Keele, B.F., Learn, G.H., Giorgi, E.E., Li, H., Decker, J.M., Wang, S., Baalwa, J., Kraus, M.H., Parrish, N.F., Shaw, K.S., Guffey, M.B., Bar, K.J., Davis, K.L., Ochsenbauer-Jambor, C., Kappes, J.C., Saag, M.S., Cohen, M.S., Mulenga, J., Derdeyn, C.A., Allen, S., Hunter, E., Markowitz, M., Hraber, P., Perelson, A.S., Bhattacharya, T., Haynes, B.F., Korber, B.T., Hahn, B.H., Shaw, G.M., 2009. Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. J. Exp. Med. 206, 1273–1289. https://doi.org/10.1084/jem.20090378

Salminen, M.O., Carr, J.K., Robertson, D.L., Hegerich, P., Gotte, D., Koch, C., Sanders-Buell, E., Gao, F., Sharp, P.M., Hahn, B.H., Burke, D.S., McCutchan, F.E., 1997. Evolution and probable transmission of intersubtype recombinant human immunodeficiency virus type 1 in a Zambian couple. J. Virol. 71, 2647–2655.

Salminen, M.O., Ehrenberg, P.K., Mascola, J.R., Dayhoff, D.E., Merling, R., Blake, B., Louder, M., Hegerich, S., Polonis, V.R., Birx, D.L., Robb, M.L., McCutchan, F.E., Michael, N.L., 2000. Construction and biological characterization of infectious molecular clones of HIV-1 subtypes B and E (CRF01_AE) generated by the polymerase chain reaction. Virology 278, 103–110. https://doi.org/10.1006/viro.2000.0640

Salminen, M.O., Johansson, B., Sönnerborg, A., Ayehunie, S., Gotte, D., Leinikki, P., Burke, D.S., McCutchan, F.E., 1996. Full-length sequence of an ethiopian human immunodeficiency virus type 1 (HIV-1) isolate of genetic subtype C. AIDS Res. Hum. Retroviruses 12, 1329–1339. https://doi.org/10.1089/aid.1996.12.1329

Sanchez, A.M., DeMarco, C.T., Hora, B., Keinonen, S., Chen, Y., Brinkley, C., Stone, M., Tobler, L., Keating, S., Schito, M., Busch, M.P., Gao, F., Denny, T.N., 2014. Development of a contemporary globally diverse HIV viral panel by the EQAPOL program. J. Immunol. Methods 409, 117–130. https://doi.org/10.1016/j.jim.2014.01.004

Scriba, T.J., de Villiers, T., Treurnicht, F.K., zur Megede, J., Barnett, S.W., Engelbrecht, S., van Rensburg, E.J., 2002. Characterization of the South African HIV type 1 subtype C complete 5' long terminal repeat, nef, and regulatory genes. AIDS Res. Hum. Retroviruses 18, 149–159. https://doi.org/10.1089/08892220252779692

Serwanga, J., Mugaba, S., Pimego, E., Nanteza, B., Lyagoba, F., Nakubulwa, S., Heath, L., Nsubuga, R.N., Ndembi, N., Gotch, F., Kaleebu, P., 2012. Profile of T cell recognition of HIV type 1 consensus group M Gag and Nef peptides in a clade A1- and D-infected Ugandan population. AIDS Res. Hum. Retroviruses 28, 384–392. https://doi.org/10.1089/AID.2011.0175

Sherefa, K., Johansson, B., Salminen, M., Sönnerborg, A., 1998. Full-length sequence of human immunodeficiency virus type 1 subtype A, recombined with subtype C in the env V3 domain. AIDS Res. Hum. Retroviruses 14, 289–292. https://doi.org/10.1089/aid.1998.14.289

Simon, F., Mauclère, P., Roques, P., Loussert-Ajaka, I., Müller-Trutwin, M.C., Saragosti, S., Georges-Courbot, M.C., Barré-Sinoussi, F., Brun-Vézinet, F., 1998. Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. Nat. Med. 4, 1032–1037. https://doi.org/10.1038/2017

Sollerkvist, L.P., Gaseitsiwe, S., Mine, M., Sebetso, G., Mphoyakgosi, T., Diphoko, T., Essex, M., Ehrnst, A., 2014. Increased CXCR4 use of HIV-1 subtype C identified by population sequencing in patients failing antiretroviral treatment compared with treatment-naive patients in Botswana. AIDS Res. Hum. Retroviruses 30, 436–445. https://doi.org/10.1089/AID.2013.0203

Song,H., Cai,F., Chen,S., Hoper,J., Jiang,C., Gao,F., 2012. Analysis of recombinant viral genomes in individuals infected with multiple transmitted/founder HIV-1 strains. Unpublished direct submission to Genbank.

Spire, B., Sire, J., Zachar, V., Rey, F., Barré-Sinoussi, F., Galibert, F., Hampe, A., Chermann, J.C., 1989. Nucleotide sequence of HIV1-NDK: a highly cytopathic strain of the human immunodeficiency virus. Gene 81, 275–284. https://doi.org/10.1016/0378-1119(89)90188-1

Srinivasan, A., Anand, R., York, D., Ranganathan, P., Feorino, P., Schochetman, G., Curran, J., Kalyanaraman, V.S., Luciw, P.A., Sanchez-Pescador, R., 1987. Molecular characterization of human immunodeficiency virus from Zaire: nucleotide sequence analysis identifies conserved and variable domains in the envelope gene. Gene 52, 71–82. https://doi.org/10.1016/0378-1119(87)90396-9

Steain, M.C., Wang, B., Yang, C., Shi, Y.-P., Nahlen, B., Lal, R.B., Saksena, N.K., 2005. HIV type 1 sequence diversity and dual infections in Kenya. AIDS Res. Hum. Retroviruses 21, 882–885. https://doi.org/10.1089/aid.2005.21.882

Sutherland, K.A., Goodall, R.L., McCormick, A., Kapaata, A., Lyagoba, F., Kaleebu, P., Thiltgen, G., Gilks, C.F., Spyer, M., Kityo, C., Pillay, D., Dunn, D., Gupta, R.K., DART Virology Group, DART Trial Team, 2015. Gag-Protease Sequence Evolution Following Protease Inhibitor Monotherapy Treatment Failure in HIV-1 Viruses Circulating in East Africa. AIDS Res. Hum. Retroviruses 31, 1032–1037. https://doi.org/10.1089/aid.2015.0138

Takahoko, M., Tobiume, M., Ishikawa, K., Ampofo, W., Yamamoto, N., Matsuda, M., Tatsumi, M., 2001. Infectious DNA clone of HIV type 1 A/G recombinant (CRF02_AG) replicable in peripheral blood mononuclear cells. AIDS Res. Hum. Retroviruses 17, 1083–1087. https://doi.org/10.1089/088922201300343771

Takekawa, N., Sakamoto, Y., Kinomoto, M., Nii-Trebi, N., Tokunaga, K., Ampofo, W., Ofori-Adjei, D., Kurata, T., Sata, T., Tatsumi, M., 2009. A new generation of recombinants between CRF02_AG and subtype A detected in Ghana : analysis by infectious molecular clones. Unpublished direct submission to Genbank.

Takekawa, N., Tatsumi, M., 2009. A comprehensive panel of infectious molecular clones derived from HIV isolates of BBI subtype infectivity panel. Unpublished direct submission to Genbank.

Tatsumi, M., Hara, T., Sakamoto, Y., Terunuma, H., Ichiyama, H., Kasolo, M., Honda, M., Yamamoto, N., 2006. Construction and characterization of HIV-1 subtype C infectious molecular clones derived from Zambian isolates. Unpublished direct submission to Genbank.

Tatsumi, M., Kinomoto, M., Nii-Trebi, N., Sakamoto, Y., Tokunaga, K., Ampofo, W., Ofori-Adjei, D., Kurata, T., Sata, T., 2005. Infectious DNA clones of subtype G and CRF02_AG HIV-1 derived from Ghananian isolates. Unpublished direct submission to Genbank.

Tatsumi, M.,, Matsuda, M., 2002. Infectious DNA clones of HIV-1 subtype A. Unpublished direct submission to Genbank.

Tatsumi,M., Sakamoto,Y., Takekawa,N., 2006. Construction and characterization of HIV-1 subtype A infectious molecular clones. Unpublished direct submission to Genbank.

Tebit, D.M., Zekeng, L., Kaptué, L., Gürtler, L., Fackler, O.T., Keppler, O.T., Herchenröder, O., Kräusslich,

H.-G., 2004. Construction and characterization of an HIV-1 group O infectious molecular clone and analysis of vpr- and nef-negative derivatives. Virology 326, 329–339. https://doi.org/10.1016/j.virol.2004.05.027

Tebit, D.M., Zekeng, L., Kaptué, L., Kräusslich, H.-G., Herchenröder, O., 2003. Construction and characterisation of a full-length infectious molecular clone from a fast replicating, X4-tropic HIV-1 CRF02.AG primary isolate. Virology 313, 645–652. https://doi.org/10.1016/s0042-6822(03)00381-7

Tebit, D.M., Zekeng, L., Kaptué, L., Salminen, M., Kräusslich, H.-G., Herchenröder, O., 2002. Genotypic and phenotypic analysis of HIV type 1 primary isolates from western Cameroon. AIDS Res. Hum. Retroviruses 18, 39–48. https://doi.org/10.1089/088922202753394709

Theodore, T., Buckler-White, A.J., 1988. Human immunodeficiency virus type 1 (HIV-1), isolate Z2, unintegrated circular proviral DNA. Unpublished direct submission to Genbank.

Thomson, M.M., Casado, G., Posada, D., Sierra, M., Nájera, R., 2005. Identification of a novel HIV-1 complex circulating recombinant form (CRF18_cpx) of Central African origin in Cuba. AIDS 19, 1155–1163. https://doi.org/10.1097/01.aids.0000176215.95119.1d

Tongo, M., Dorfman, J.R., Abrahams, M.-R., Mpoudi-Ngole, E., Burgers, W.A., Martin, D.P., 2015. Near full-length HIV type 1M genomic sequences from Cameroon : Evidence of early diverging under-sampled lineages in the country. Evol Med Public Health 2015, 254–265. https://doi.org/10.1093/emph/eov022

Tongo, M., Martin, D.P., Zembe, L., Mpoudi-Ngole, E., Williamson, C., Burgers, W.A., 2013. Characterization of HIV-1 gag and nef in Cameroon: further evidence of extreme diversity at the origin of the HIV-1 group M epidemic. Virol. J. 10, 29. https://doi.org/10.1186/1743-422X-10-29

Toni, T., Adjé-Touré, C., Vidal, N., Minga, A., Huet, C., Borger, M.-Y., Recordon-Pinson, P., Masquelier, B., Nolan, M., Nkengasong, J., Fleury, H.J., Delaporte, E., Peeters, M., 2005. Presence of CRF09_cpx and complex CRF02_AG/CRF09_cpx recombinant HIV type 1 strains in Côte d'Ivoire, West Africa. AIDS Res. Hum. Retroviruses 21, 667–672. https://doi.org/10.1089/aid.2005.21.667

Toure-Kane, C., Montavon, C., Faye, M.A., Gueye, P.M., Sow, P.S., Ndoye, I., Gaye-Diallo, A., Delaporte, E., Peeters, M., Mboup, S., 2000. Identification of all HIV type 1 group M subtypes in Senegal, a country with low and stable seroprevalence. AIDS Res. Hum. Retroviruses 16, 603–609. https://doi.org/10.1089/088922200309025

Tovanabutra, S., Sanders-Buell, E., Arroyo, M.A., Maboko, L., Hoffman, O., Bose, M., Crossler, J., Bradfield, A., Geis, S., Birx, D.L., Robb, M.L., Michael, N.L., Hoelscher, M., McCutchan, F.E., 2009. HIV-1 subtype C gag gene from an incident infection identified in Tanzania. Unpublished direct submission to Genbank.

Treurnicht, F.K., Seoighe, C., Martin, D.P., Wood, N., Abrahams, M.-R., Rosa, D. de A., Bredell, H., Woodman, Z., Hide, W., Mlisana, K., Karim, S.A., Gray, C.M., Williamson, C., 2010. Adaptive changes in HIV-1 subtype C proteins during early infection are driven by changes in HLA-associated immune pressure. Virology 396, 213–225. https://doi.org/10.1016/j.virol.2009.10.002

Triques, K., Bourgeois, A., Vidal, N., Mpoudi-Ngole, E., Mulanga-Kabeya, C., Nzilambi, N., Torimiro, N., Saman, E., Delaporte, E., Peeters, M., 2000. Near-full-length genome sequencing of divergent African HIV type 1 subtype F viruses leads to the identification of a new HIV type 1 subtype designated K. AIDS Res. Hum. Retroviruses 16, 139–151. https://doi.org/10.1089/088922200309485

Trkola, A., Paxton, W.A., Monard, S.P., Hoxie, J.A., Siani, M.A., Thompson, D.A., Wu, L., Mackay, C.R., Horuk, R., Moore, J.P., 1998. Genetic subtype-independent inhibition of human immunodeficiency virus type 1 replication by CC and CXC chemokines. J. Virol. 72, 396–404.

Tzitzivacos, D.B., Tiemessen, C.T., Stevens, W.S., Papathanasopoulos, M.A., 2009. Viral genetic determinants of nonprogressive HIV type 1 subtype C infection in antiretroviral drug-naive children. AIDS Res. Hum. Retroviruses 25, 1141–1148. https://doi.org/10.1089/aid.2009.0080

Vallari, A., Bodelle, P., Ngansop, C., Makamche, F., Ndembi, N., Mbanya, D., Kaptué, L., Gürtler, L.G., McArthur, C.P., Devare, S.G., Brennan, C.A., 2010. Four new HIV-1 group N isolates from Cameroon: Prevalence continues to be low. AIDS Res. Hum. Retroviruses 26, 109–115.

https://doi.org/10.1089/aid.2009.0178
Vallari, A., Holzmayer, V., Harris, B., Yamaguchi, J., Ngansop, C., Makamche, F., Mbanya, D., Kaptué, L., Ndembi, N., Gürtler, L., Devare, S., Brennan, C.A., 2011. Confirmation of putative HIV-1 group P in Cameroon. J. Virol. 85, 1403–1407. https://doi.org/10.1128/JVI.02005-10
Vallejo, A., Heredia, A., Mas, A., Lee, S.F., Epstein, J.S., Soriano, V., Hewlett, I.K., 1998. Tropism, coreceptor use, and phylogenetic analysis of both the V3 loop and the protease gene of three novel HIV-1 group O isolates. J. Acquir. Immune Defic. Syndr. Hum. Retrovirol. 18, 417–425.
van Harmelen, J., Williamson, C., Kim, B., Morris, L., Carr, J., Karim, S.S., McCutchan, F., 2001. Characterization of full-length HIV type 1 subtype C sequences from South Africa. AIDS Res. Hum. Retroviruses 17, 1527–1531. https://doi.org/10.1089/08892220152644232
Vergne, L., Bourgeois, A., Mpoudi-Ngole, E., Mougnutou, R., Mbuagbaw, J., Liegeois, F., Laurent, C., Butel, C., Zekeng, L., Delaporte, E., Peeters, M., 2003. Biological and genetic characteristics of HIV infections in Cameroon reveals dual group M and O infections and a correlation between SI-inducing phenotype of the predominant CRF02_AG variant and disease stage. Virology 310, 254–266. https://doi.org/10.1016/s0042-6822(03)00167-3
Vergne, L., Peeters, M., Mpoudi-Ngole, E., Bourgeois, A., Liegeois, F., Toure-Kane, C., Mboup, S., Mulanga-Kabeya, C., Saman, E., Jourdan, J., Reynes, J., Delaporte, E., 2000. Genetic diversity of protease and reverse transcriptase sequences in non-subtype-B human immunodeficiency virus type 1 strains: evidence of many minor drug resistance mutations in treatment-naive patients. J. Clin. Microbiol. 38, 3919–3925.
Vidal, N., Bazepeo, S.E., Mulanga, C., Delaporte, E., Peeters, M., 2009. Genetic characterization of eight full-length HIV type 1 genomes from the Democratic Republic of Congo (DRC) reveal a new subsubtype, A5, in the A radiation that predominates in the recombinant structure of CRF26_A5U. AIDS Res. Hum. Retroviruses 25, 823–832. https://doi.org/10.1089/aid.2008.0283
Vidal, N., Frange, P., Chaix, M.-L., Mulanga, C., Lepira, F., Bazepeo, S.E., Goujard, C., Meyer, L., Rouzioux, C., Delaporte, E., Peeters, M., 2008. Characterization of an old complex circulating recombinant form, CRF27_cpx, originating from the Democratic Republic of Congo (DRC) and circulating in France. AIDS Res. Hum. Retroviruses 24, 315–321. https://doi.org/10.1089/aid.2007.0241
Vidal, N., Koyalta, D., Richard, V., Lechiche, C., Ndinaromtan, T., Djimasngar, A., Delaporte, E., Peeters, M., 2003. High genetic diversity of HIV-1 strains in Chad, West Central Africa. J. Acquir. Immune Defic. Syndr. 33, 239–246. https://doi.org/10.1097/00126334-200306010-00020
Vidal, N., Mulanga, C., Bazepeo, S.E., Lepira, F., Delaporte, E., Peeters, M., 2006. Identification and molecular characterization of subsubtype A4 in central Africa. AIDS Res. Hum. Retroviruses 22, 182–187. https://doi.org/10.1089/aid.2006.22.182
Vidal, N., Mulanga-Kabeya, C., Nzilambi, N., Delaporte, E., Peeters, M., 2000. Identification of a complex env subtype E HIV type 1 virus from the democratic republic of congo, recombinant with A, G, H, J, K, and unknown subtypes. AIDS Res. Hum. Retroviruses 16, 2059–2064. https://doi.org/10.1089/088922200750054800
Vidal,N., 2002a. Genetic variability of HIV-1 strains in Chad: high prevalence of recombinant strains and identifictauion [sic] of a new D' variant. Unpublished direct submission to Genbank.
Vidal,N., 2002b. Genetic variability of HIV-1 strains in chad: high prevalence of recombinant strains and identification of a new D" variant. Unpublished direct submission to Genbank.
Vinner, L., Holmgren, B., Jensen, K.J., Esbjornsson, J., Borggren, M., Hentze, J.L., Karlsson, I., Andresen, B.S., Gram, G.J., Kloverpris, H., Aaby, P., Da Silva, Z.J., Fenyö, E.-M., Fomsgaard, A., 2011. Sequence analysis of HIV-1 isolates from Guinea-Bissau: selection of vaccine epitopes relevant in both West African and European countries. APMIS 119, 487–497. https://doi.org/10.1111/j.1600-0463.2011.02763.x
Visawapoka, U., Tovanabutra, S., Currier, J.R., Cox, J.H., Mason, C.J., Wasunna, M., Ponglikitmongkol, M., Dowling, W.E., Robb, M.L., Birx, D.L., McCutchan, F.E., 2006. Circulating and unique recombinant forms of HIV type 1 containing subsubtype A2. AIDS Res. Hum. Retroviruses 22,

695–702. https://doi.org/10.1089/aid.2006.22.695
Wilbe, K., Casper, C., Albert, J., Leitner, T., 2002. Identification of two CRF11-cpx genomes and two preliminary representatives of a new circulating recombinant form (CRF13-cpx) of HIV type 1 in Cameroon. AIDS Res. Hum. Retroviruses 18, 849–856. https://doi.org/10.1089/08892220260190326
Wilbe, K., Salminen, M., Laukkanen, T., McCutchan, F., Ray, S.C., Albert, J., Leitner, T., 2003. Characterization of novel recombinant HIV-1 genomes using the branching index. Virology 316, 116–125. https://doi.org/10.1016/j.virol.2003.08.004
Wilkinson, E., Engelbrecht, S., 2009. Molecular characterization of non-subtype C and recombinant HIV-1 viruses from Cape Town, South Africa. Infect. Genet. Evol. 9, 840–846. https://doi.org/10.1016/j.meegid.2009.05.001
Wilkinson, E., Holzmayer, V., Jacobs, G.B., de Oliveira, T., Brennan, C.A., Hackett, J., van Rensburg, E.J., Engelbrecht, S., 2015. Sequencing and phylogenetic analysis of near full-length HIV-1 subtypes A, B, G and unique recombinant AC and AD viral strains identified in South Africa. AIDS Res. Hum. Retroviruses 31, 412–420. https://doi.org/10.1089/AID.2014.0230
Williamson, C., Morris, L., Maughan, M.F., Ping, L.-H., Dryga, S.A., Thomas, R., Reap, E.A., Cilliers, T., van Harmelen, J., Pascual, A., Ramjee, G., Gray, G., Johnston, R., Karim, S.A., Swanstrom, R., 2003. Characterization and selection of HIV-1 subtype C isolates for use in vaccine development. AIDS Res. Hum. Retroviruses 19, 133–144. https://doi.org/10.1089/088922203762688649
Wright, J.K., Brumme, Z.L., Carlson, J.M., Heckerman, D., Kadie, C.M., Brumme, C.J., Wang, B., Losina, E., Miura, T., Chonco, F., van der Stok, M., Mncube, Z., Bishop, K., Goulder, P.J.R., Walker, B.D., Brockman, M.A., Ndung'u, T., 2010. Gag-protease-mediated replication capacity in HIV-1 subtype C chronic infection: associations with HLA type and clinical parameters. J. Virol. 84, 10820–10831. https://doi.org/10.1128/JVI.01084-10
Wright, J.K., Naidoo, V.L., Brumme, Z.L., Prince, J.L., Claiborne, D.T., Goulder, P.J.R., Brockman, M.A., Hunter, E., Ndung'u, T., 2012. Impact of HLA-B*81-associated mutations in HIV-1 Gag on viral replication capacity. J. Virol. 86, 3193–3199. https://doi.org/10.1128/JVI.06682-11
Wright, J.K., Novitsky, V., Brockman, M.A., Brumme, Z.L., Brumme, C.J., Carlson, J.M., Heckerman, D., Wang, B., Losina, E., Leshwedi, M., van der Stok, M., Maphumulo, L., Mkhwanazi, N., Chonco, F., Goulder, P.J.R., Essex, M., Walker, B.D., Ndung'u, T., 2011. Influence of Gag-protease-mediated replication capacity on disease progression in individuals recently infected with HIV-1 subtype C. J. Virol. 85, 3996–4006. https://doi.org/10.1128/JVI.02520-10
Yamaguchi, J., Bodelle, P., Kaptué, L., Zekeng, L., Gürtler, L.G., Devare, S.G., Brennan, C.A., 2003. Near full-length genomes of 15 HIV type 1 group O isolates. AIDS Res. Hum. Retroviruses 19, 979–988. https://doi.org/10.1089/088922203322588332
Yamaguchi, J., Bodelle, P., Vallari, A.S., Coffey, R., McArthur, C.P., Schochetman, G., Devare, S.G., Brennan, C.A., 2004. HIV infections in northwestern Cameroon: identification of HIV type 1 group O and dual HIV type 1 group M and group O infections. AIDS Res. Hum. Retroviruses 20, 944–957. https://doi.org/10.1089/aid.2004.20.944
Yamaguchi, J., Coffey, R., Vallari, A., Ngansop, C., Mbanya, D., Ndembi, N., Kaptué, L., Gürtler, L.G., Bodelle, P., Schochetman, G., Devare, S.G., Brennan, C.A., 2006. Identification of HIV type 1 group N infections in a husband and wife in Cameroon: viral genome sequences provide evidence for horizontal transmission. AIDS Res. Hum. Retroviruses 22, 83–92. https://doi.org/10.1089/aid.2006.22.83
Yamaguchi, J., Ndembi, N., Ngansop, C., Mbanya, D., Kaptué, L., Gürtler, L.G., Devare, S.G., Brennan, C.A., 2009. HIV type 1 group M subtype G in Cameroon: five genome sequences. AIDS Res. Hum. Retroviruses 25, 469–473. https://doi.org/10.1089/aid.2008.0296
Yamaguchi, J., Vallari, A., Ngansop, C., Makamche, F., Ndembi, N., Mbanya, D., Kaptué, L., Gürtler, L.G., Devare, S.G., Brennan, C.A., 2010. Near full-length sequence of HIV type 1 subtype J strain 04CMU11421 from Cameroon. AIDS Res. Hum. Retroviruses 26, 693–697. https://doi.org/10.1089/aid.2009.0305

Yamaguchi, J., Vallari, A.S., Swanson, P., Bodelle, P., Kaptué, L., Ngansop, C., Zekeng, L., Gürtler, L.G., Devare, S.G., Brennan, C.A., 2002. Evaluation of HIV type 1 group O isolates: identification of five phylogenetic clusters. AIDS Res. Hum. Retroviruses 18, 269–282. https://doi.org/10.1089/088922202753472847

Yue, L., Pfafferott, K.J., Baalwa, J., Conrod, K., Dong, C.C., Chui, C., Rong, R., Claiborne, D.T., Prince, J.L., Tang, J., Ribeiro, R.M., Cormier, E., Hahn, B.H., Perelson, A.S., Shaw, G.M., Karita, E., Gilmour, J., Goepfert, P., Derdeyn, C.A., Allen, S.A., Borrow, P., Hunter, E., 2015. Transmitted virus fitness and host T cell responses collectively define divergent infection outcomes in two HIV-1 recipients. PLoS Pathog. 11, e1004565. https://doi.org/10.1371/journal.ppat.1004565

Zhao, J., Tang, S., Ragupathy, V., Carr, J.K., Wolfe, N.D., Awazi, B., Hewlett, I., 2010. Identification and genetic characterization of a novel CRF22_01A1 recombinant form of HIV type 1 in Cameroon. AIDS Res. Hum. Retroviruses 26, 1033–1045. https://doi.org/10.1089/aid.2009.0197

Zhao, J., Tang, S., Ragupathy, V., Gaddam, D., Wang, X., Zhang, P., Nyambi, P.N., Hewlett, I., 2012. CRF22_01A1 is involved in the emergence of new HIV-1 recombinants in Cameroon. J. Acquir. Immune Defic. Syndr. 60, 344–350. https://doi.org/10.1097/QAI.0b013e318258c7e3

Zhao, J., Tang, S., Viswanath, R., Wang, X., Awazi, B., Hewlett,I., 2008. Identification of new recombinant of CRF22 with subtype A1 or CRF02_AG of HIV-1 in Cameroon. Unpublished direct submission to Genbank.

Zheng, N.N., Daniels, R.S., 2004. Substitution of hiv type-1 non-B env genes in C2, a subtype B cassette system, results in functional chimeric viruses. AIDS Res. Hum. Retroviruses 20, 964–971. https://doi.org/10.1089/aid.2004.20.964

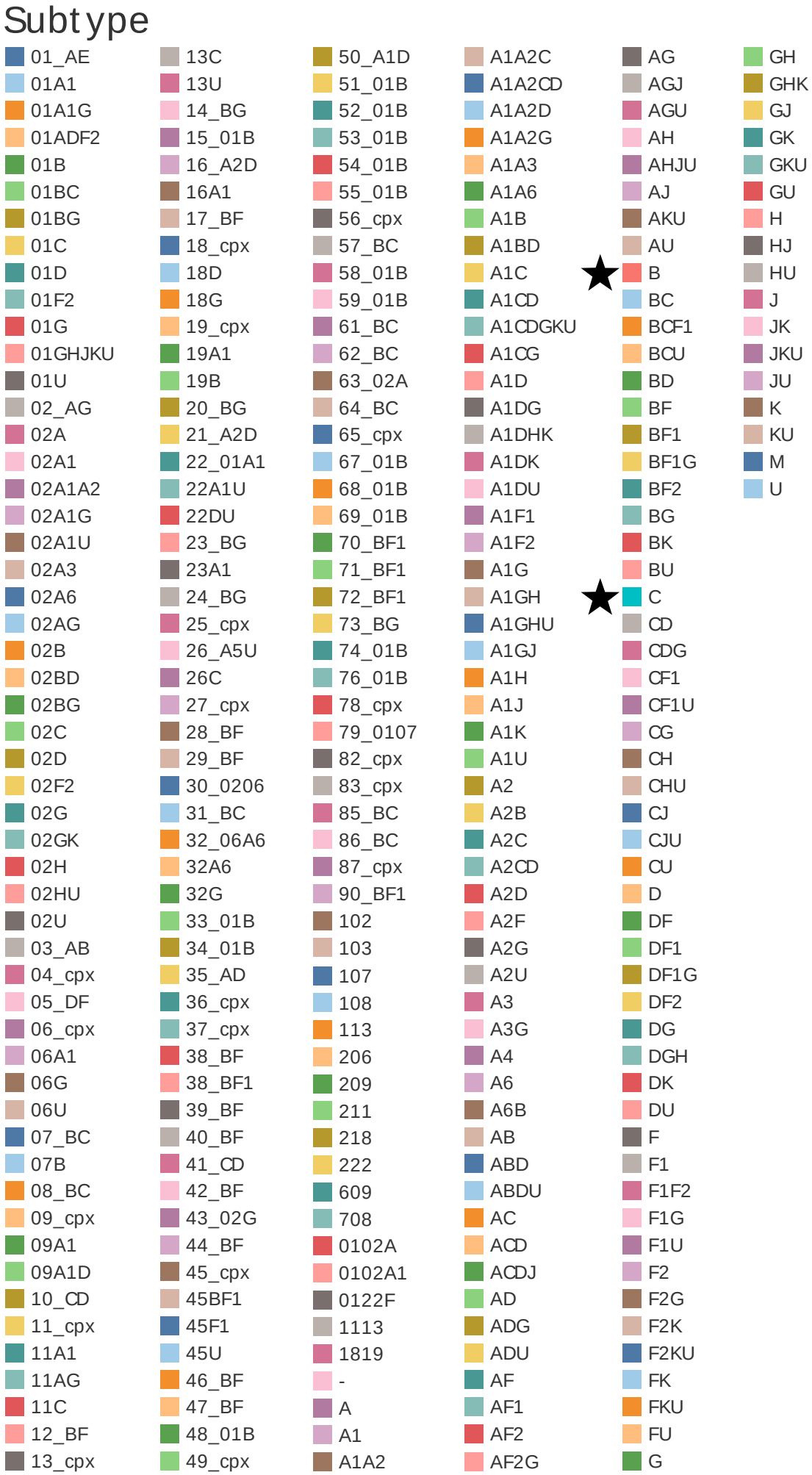
# Appendix 1B

*Table 3.2* ISO alpha-2 two-letter country codes and full names appearing in figures 3.2 and 3.3

| Country Abbreviation | Country Name |
|---|---|
| AO | Angola |
| BF | Burkina Faso |
| BI | Burundi |
| BJ | Benin |
| BW | Botswana |
| CD | Democratic Republic of the Congo |
| CF | Central African Republic |
| CG | Congo |
| CI | Ivory Coast |
| CM | Cameroon |
| DJ | Djibouti |
| ET | Ethiopia |
| GA | Gabon |
| GH | Ghana |
| GM | Republic of the Gambia |
| GN | Guinea |
| GQ | Equatorial Guinea |
| GW | Guinea Bissau |
| KE | Kenya |
| LR | Liberia |
| MG | Madagascar |
| ML | Mali |
| MW | Malawi |
| MZ | Mozambique |
| NE | Niger |
| NG | Nigeria |
| RW | Rwanda |
| SD | Sudan |
| SN | Senegal |
| SO | Somalia |
| TD | Chad |
| TG | Togo |
| TZ | United Republic of Tanzania |
| UG | Uganda |
| ZA | South Africa |
| ZM | Zambia |
| ZW | Zimbabwe |

**Complete key to figure 3.4**

List of subtypes appearing in figure 3.4a Subtypes B and C are marked with stars. List of ethno-linguistic groups and language families in Africa appearing in figure 3.4b (Center for Geographic Analysis, 2013). Some groups on the map do not have associated names in the original data. Bantu language families marked with stars.


Subtype
01_AE
01A1
01A1G
01ADF2
01B
01BC
01BG
01C
01D
01F2
01G
01GHJKU
01U
02_AG
02A
02A1
02A1A2
02A1G
02A1U
02A3
02A6
02AG
02B
02BD
02BG
02C
02D
02F2
02G
02GK
02H
02HU
02U
03_AB
04_cpx
05_DF
06_cpx
06A1
06G
06U
07_BC
07B
08_BC
09_cpx
09A1
09A1D
10_CD
11_cpx
11A1
11AG
11C
12_BF
13_cpx
13C
13U
14_BG
15_01B
16_A2D
16A1
17_BF
18_cpx
18D
18G
19_cpx
19A1
19B
20_BG
21_A2D
22_01A1
22A1U
22DU
23_BG
23A1
24_BG
25_cpx
26_A5U
26C
27_cpx
28_BF
29_BF
30_0206
31_BC
32_06A6
32A6
32G
33_01B
34_01B
35_AD
36_cpx
37_cpx
38_BF
38_BF1
39_BF
40_BF
41_CD
42_BF
43_02G
44_BF
45_cpx
45BF1
45F1
45U
46_BF
47_BF
48_01B
49_cpx
50_A1D
51_01B
52_01B
53_01B
54_01B
55_01B
56_cpx
57_BC
58_01B
59_01B
61_BC
62_BC
63_02A
64_BC
65_cpx
67_01B
68_01B
69_01B
70_BF1
71_BF1
72_BF1
73_BG
74_01B
76_01B
78_cpx
79_0107
82_cpx
83_cpx
85_BC
86_BC
87_cpx
90_BF1
102
103
107
108
113
206
209
211
218
222
609
708
0102A
0102A1
0122F
1113
1819
-
A
A1
A1A2
A1A2C
A1A2CD
A1A2D
A1A2G
A1A3
A1A6
A1B
A1BD
A1C
A1CD
A1CDGKU
A1CG
A1D
A1DG
A1DHK
A1DK
A1DU
A1F1
A1F2
A1G
A1GH
A1GHU
A1GJ
A1H
A1J
A1K
A1U
A2
A2B
A2C
A2CD
A2D
A2F
A2G
A2U
A3
A3G
A4
A6
A6B
AB
ABD
ABDU
AC
ACD
ACDJ
AD
ADG
ADU
AF
AF1
AF2
AF2G
AG
AGJ
AGU
AH
AHJU
AJ
AKU
AU
B
BC
BCF1
BCU
BD
BF
BF1
BF1G
BF2
BG
BK
BU
C
CD
CDG
CF1
CF1U
CG
CH
CHU
CJ
CJU
CU
D
DF
DF1
DF1G
DF2
DG
DGH
DK
DU
F
F1
F1F2
F1G
F1U
F2
F2G
F2K
F2KU
FK
FKU
FU
G
GH
GHK
GJ
GK
GKU
GU
H
HJ
HU
J
JK
JKU
JU
K
KU
M
U

# Ethnicity

ABABDAH
ABBALA
ABE
ABIDJI
ABINSI
ABO
ABRON
ACIPA
ACOLI
ADARAWA
ADJUKRU
AFAR
AFO
AGAR
AGATU
AGOI
AGUA
AHANTA
AIKE
AIT WARAIN
AJA
AKAN
AKANIKI
AKARE
AKEBON
AKOKO-EDO
AKOOSE
AKPA
AKPA-YACHE
AKPES
AKPINI
AKPOSO
AKYE
ALAGO
ALEGE
ALGERIAN
ALUR
AMARAR
AMBA
AMBO
AMBOIM
ANAANG
ANFILLO
ANGAS
ANGERE
ANIA
ANO
ANTANDROY
ANTANUSI
ANTEMURU
ANTESAKA
ANUAK
ANUFO
ANYI
AOUDJILA
AOUK
APAKIBETI
ARAD
ARGOBA
ARIAAL
ARUM
ARUSI
ASANFWE
ATEN
ATSAM
ATTA
ATUOT
AULYEHAN
AUSHI
AUYUKAWA
AVIKAM
AVOKAYA-AJIGU
AVOKAYA-OJILA
AWAK
AWIYA
AWUTO
AYIGHA
BA
BAALI
BABALE
BABANKI
BABILE
BABOLE
BADA
BADE
BAGA
BAGA MANDORI
BAGGARA
BAGHAP
BAGIRMI
BAHARIYA
BAI
BAILO
BAILUNDU
BAINUK
BAKA
BAKO
BAKPINKA
BAKUNDU
BAKWE
BAKWELE
BALANTA
BALONG
BAMBARA
BAMILEKE
BAMUN
BANDA
BANDI
BANGANTU
BANGBA
BANGU-BANGU
BANGWINJI
BANNA
BARA
BARABRA
BARAMBO
BAREA
BAREIN
BARI
BARIBA
BARWE
BASA
BASAA
BASHAR
BASONGO-MENO
BASSA
BASSOSSI
BATA
BATAHIN
BATI
BAULE
BAUSHI
BAYGO
BEBIL
BEDAWI
BEEMBE
BEKWARRA
BELANDA
BELANDA VIRI
BELLE
BEMBA
BEMBE
BENA
BENDE
BENGA
BENI GIL
BENI MENASSER
BENI MZAB
BENI-AMER
BERABISH
BERANDJOKO
BERGDAMA
BERIBERI
BEROM
BERTA
BERTI
BESOMBO
BETE
BETE-BENDE
BETI
BETSILEO
BETSIMISARAKA
BEYRU
BEZANUZANU
BIAFADA
BIALI
BIDEYAT
BIDIANKAMBA
BILE
BILEN
BIMOBA
BINDI
BINGA
BIRA
BIRGIT
BIRIFOR
BIRKED
BISA
BOBANGI
BOBILI
BOBO
BOBO-OULE
BODO
BOFI
BOGHOM
BOGURU
BOKALA
BOKIBA
BOKKOS
BOKO, BUSA
BOKONGO
BOKYI
BOLENDO
BOLENGE
BOLGO
BOLOKI
BOLOMBOKI
BOLON
BOLONGO
BOM
BOMASA
BOMU
BONDEI
BONDJO
BONDO
BONDONGO
BONGI
BONGILI
BONGO
BOR
BORORO
BOSOKU
BOZO
BUA
BUBIA
BUDAMONO
BUDDU
BUDJA
BUDUKWA
BUDUMA
BUILSA
BUKA
BULAANG
BULALA
BULLOM
BUMA
BUMAJI
BUMALI
BUNGU
BURA-PABIR
BURAK
BURJI
BURU
BURUN
BURUNGE
BUSHOONG
BUYU
BWAALA
BWARI
BWENDE
BWILE
BWISI
BYEP
CARA
CHAAMBA
CHAGGA
CHAKALI
CHAM
CHAMBA
CHAOUIA
CHE
CHISHINGA
CHIWERE'S NGUNI
CHOKWE
CHOKWE, LELE
CHOKWE, LUNDA
CHOPI
CIKUMA
CINGOLO
CIPALA
CISANJI
CITATA
CIVULA
COMO KARIM
CWAMHI-WURI
DAAROOD
DABA
DADIYA
DAGAARI
DAGOMBA
DAHALO
DAISU
DAJU
DAKAKARI
DAKHLA
DAMA
DAN
DAN-GEH
DANAKIL
DANGYO
DARI
DASS
DAY
DCIRIKU
DEG
DEGEMA
DEKAKIRE
DELIM
DEMBO
DENYA
DEY
DIA
DIDA
DIDINGA
DIGO
DIGODIA
DII
DIJIM
DIKIDIKI
DINKA
DIO
DIOBO
DIZI
DJAGADA
DJEBALA
DJEM
DJENNENKE
DJERBA
DODOTH
DOE
DOGON
DOKA
DOKO
DOKO-UYANGA
DONDO
DONGO
DRAWA
DUAISH
DUALA
DUI-MENIA
DUKA
DULBU
DULI
DUMBULE
DURUMA
DUUPA
DUWAI
DYALONKE
DYE
DYULA
DZA
DZING
EASTERN LUBA, KUNDA
EBANGA
EBIRA
EDO
EFIK
EFUTOP
EGGON
EGYPTIAN
EJAGHAM
EJAMAT
EKIT
ELEKU
ENDO
ENYELLE
EOTILE
EPEITA
ESAN
ESELE
ESHIRA
ESIMBI
ETON
ETULO
EWE
EWONDO
FALI
FANYAN
FEZARA
FEZZAN
FIA
FIGIG
FIHERENA
FILALA
FIPA
FODONON
FON-GBE
FONGE
FONGORO
FOODO
FUFULDE
FULANI
FULANKRIABE
FULBE
FULILWA
FULIRU
FUNGWA
FUNGWE
FUR
FURU
FYAM
FYER
GA'ANDA
GAAJAK
GAAJOK
GAALIN
GAAM
GADAMES
GADE
GAFSA
GAGU
GALA
GALAMBU
GANAGA
GANDA
GANJA
GARREH-AJURAN
GAYA
GBAGA
GBAGYI
GBANZIRI
GBARI
GBAYA
GBERI
GBII
GBINDA
GBIRI
GBOMA
GEJI
GENYA
GERAWA
GERI
GHANGALA
GHULFAN
GIBE
GIDAR
GIIWO
GIMIRO
GIMMA
GIMNIME
GIMR
GINGO
GISU
GIZIGA SOUTH
GOBU
GODIE
GOEMAI
GOGO
GOKANA
GOLA
GOMA
GOMANI'S NGUNI
GONDO
GONJA
GOROWA
GOUIN
GOWA
GREBO
GUA
GUDE
GUDU
GUENZE
GULA
GULA KARA OF SUDAN
GULE
GURAGE
GURENSI
GURMA
GURMANA
GURO
GURUNTUM
GWA
GWANDARA
GWE
GWENO
GWERE
GWOMU
GYEM
HA
HADENDOWA
HADIYYA
HAM
HAMAMA
HAMBA
HAMYAN
HANDA
HANGAZA
HANYA
HARARGE OROMO
HARARI
HAVU
HAWIYA
HAYA
HEHE
HEMBA
HEMBA, BANGU-BANGU
HEMBA, BUYU
HENGA
HERERO
HIMA
HINGA
HOLO-HOLO
HOLOHOLO
HOLU
HOMBO
HOROM
HUBA
HUM
HUNDE
HUNGANA
HUNGO
HUNGWORO
HUTU
IBAJI
IBIBIO
IBINO
ICEN
ICHEVE
IDOMA
IDON
IFUMBA
IGALA
IGBO
IGEDE
IGUTA
IJAW
IJE
IKA
IKASSA
IKENGA
IKIZU
IKOMA
IKONGO
IKPESHI
IKULU
ILA
ILA-TONGA
ILEBO
ILWANA
IMBANGALA
IMOMA
IMRAGEN
INJOLO
IPANGA
IPULO
IRANGI
IRAQW
IRIGWE
ISAAQ
ISANZU
ISHA
ISOKO
ISSA
ISUWU
ITSEKIRI
ITU
IVBIOSAKON
IWA
IWELLEMMEDEN
IYEMBE
IZDRA
IZI
JA' ALIYYIN GAALIIN
JAMALA
JANJERO
JANJI
JARA
JARAWA
JARSO
JEKAING
JERA
JERID

# Ethnicity (cont.)

JIBU
JIDA
JIE
JIJI
JIMI
JIMINI
JIRU
JITA
JIU
JONGA
JOOLA
JOPADHOLA
JUKUN
JUKUN KONA
JUKUN WASE
JUKUN WURKUM
JUNGO
JUR
JUR MODO
JWIRA-PEPESA
KAAMBA
KABA
KABA DEMI
KABA NAA
KABA SO
KABONGA
KABRE
KABYE
KABYLE
KADARA
KADARU
KAFFA
KAGOMA
KAGORO
KAGURU
KAJAKSE
KAKO
KAKONDA
KALAMSE
KALANGA
KALIKO
KALUNDWE
KAM
KAMANGA
KAMANTAN
KAMBA
KAMBARI
KAMBATA
KAMIR
KAMKAM
KAMO
KAMUKU
KAMWE
KANAKURU
KANDAWIRE
KANDERMA
KANEMBU
KANGO
KANTANA
KANU
KANUFI
KANURI
KANYOK
KAONDE
KAONDE, SANGA
KARA
KARABORO
KARANGA
KARE
KAREKARE
KASEM
KASENA
KASONGI
KASONGO
KASONKE
KATAB
KATYA
KAWENDI
KAYAMBA
KAYLA
KEDERU
KEENGE
KEL
KEL AHAGGAR
KEL AYR ASBEN
KEL GRES
KEL IFFORA
KELE
KENGA
KENYANG
KENYI
KERA
KERARISH
KEREWE
KETE
KETE EAST
KETE LULUA
KETE NORTH
KETE SOUTH
KEYO
KHARGA
KHUMALO
KIBET
KIGA
KIKUYU
KIM
KIMBU
KIMBUNDU
KINGA
KIONG
KIPSIKIS
KIR-BALAR
KISAMA
KISSI
KITA
KIYAK
KLAOH
KO
KOBIANA
KOFFA
KOFYAR
KOHUMONO
KOKE
KOKO
KOL BIKELE
KOLBILA
KOM
KOMA
KOME
KONGO DINGA
KONJI
KONJO
KONKOMBA
KONO
KONONGO
KONSO
KONYANKE
KOONI
KORO
KOROP
KOTA
KOTE
KOTOKO
KOTOPO
KOUYA
KPELLE
KPESSI
KRAN, NGERE
KRESH
KRIM
KUANG
KUDU
KUFRA
KUKA
KUKELE
KUKU
KUKUYA
KULANGO
KULERE
KULUNG
KUMU
KUNAMA
KUNDA
KUNDU
KUNTA
KUNUZ
KUNYI
KUPTO
KURAMA
KURANKO
KURFEI
KURI
KURUMFE
KURYA
KUSAAL
KUSHI
KUSU
KUTEP
KUTIN
KUTU
KUTURMI
KUYERI
KUYU
KWAAMI
KWAFI
KWAKUM
KWAME
KWANDI
KWANGARI
KWANGWA
KWANJA
KWANKA
KWAYA
KWERE
KWESE
LAALI
LABWOR
LADO
LAGBA
LAGUAT
LAGWAN
LAKA
LALA
LALA ROBA
LALIA
LAMA
LAMBA
LAMBYA
LAMJA
LANDUMA
LANGA
LANGO
LARU
LEELAU
LEGA
LEGBO
LEKA
LELE
LEMBWE
LEMORO
LENDU
LENGI
LENGOLA
LENJE
LESE
LIA
LIBOLO
LIGBI
LIGI
LIJILI
LIKO
LIKWALA
LIMA
LIMO
LO
LOBI
LOBI LORHON
LOGO
LOGORIK
LOI
LOKALO
LOKELE
LOKO
LOKORO
LOKOYA
LOMA
LOMBI
LOMOTWA
LOMWE
LONDO
LONGARIM
LONGTO
LOPA
LOSAKANYI
LOTSU-PIRI
LOVALE
LOVEDU
LOZI
LUANO
LUBA
LUBA EASTERN
LUBA KASAYI
LUBA SHANKADI
LUBA UPEMBA
LUBA, LUNDA
LUBA, NDEMBU
LUBA, NDEMBU, LUNDA
LUBILA
LUCHAZI
LUGBARE
LUGURU
LULA
LULUBA
LULUWA
LUMBU
LUNDA
LUNDA, CHOKWE
LUNDA, KETE
LUNDA, NDEMBU
LUNGU
LUNTU
LURI
LUSHANGI
LUUNDA
LUYIA
LWALU
LWENA
LWER
LWIMBI
LYANGALILE
LYELE
MA
MAAKA
MAASAI
MAAZA
MABA
MABENDI
MABEZA
MABIHA
MABINZA
MACHINGA
MADA
MADI
MAGHARBA
MAGUZAWA
MAHAFALY
MAHONGWE
MAK
MAKAA
MAKERE
MAKOMA
MAKONDE
MAKUA
MAKWA
MALINKE
MALWAL
MAMBAI
MAMBILA
MAMBWE
MAMPRUSI
MAMVU
MANALA
MANDA
MANDARI
MANDINKA
MANDYAK
MANGAS
MANGAYAT
MANGBELE
MANGBETU
MANGBUTU
MANINKA
MANKANYA
MANO
MANYANGA
MANYIKA
MANZA
MAO SOUTH
MAPOPOI
MARARIT
MARBA
MARFA
MARGHI
MARGHI SOUTH
MARKA
MASAKIN
MASALIT
MASANA
MASHASHA
MASHI
MASMAJE
MASONGO
MASSA
MASSALAT
MATENGO
MATUMBI
MAU
MAURI
MAYOGO
MBA
MBAAMBA
MBAATI
MBAGANI
MBAI BEDIONDO
MBALA
MBANGWE
MBANZA
MBATA
MBEKO
MBELIME
MBELO
MBEMBE
MBEMBRE
MBENE
MBERE
MBESA
MBINSA
MBO
M'BOKO
MBOLE
MBOMOTABA
MBONGO
MBOSHI
MBUGWE
MBUKUSHU
MBULI
MBULU
MBUM
MBUNDA
MBUNGA
MBUUN
MBWELA
MBWELA KOLWE
MBWERA
MDUNDULU
MEDJE
MEDOGO
MENDE
MENING
MERINA
MERU
MESME
MFINU
MFUNU
MIDOB
MIGAAMA
MILTU
MIMI
MINIANKA
MINUNGO
MINUNGU
MITSOGHO
MITTU
MITUKU
MIYA
MOBA
MOBANGO
MOBENGE
MOKOLE
MOLO
MONDOMBE
MONDZOMBO
MONGO
MONO
MONTOL
MOROCCAN
MORU
MOSSI
MOTEMBO
MOYE
MPAMA
MPANGU
MPE
MPEZENI'S NGUNI
MPONGMPONG
MPOTO
MPUT
MPUTU
MPYEMO
MSER
MUBI
MUCHIKONGO
MUKULU
MUMUYE
MUN
MUNDANG
MUNDU
MUNGA
MURLE
MUSGU
MUSSEY
MUSSUMBE
MVELE
MWAGHAVUL
MWENYI
MYANG
MYENE
NABAN
NAFANA
NAGUMI
NALU
NAMBA
NAMJI
NANCERE
NANDE
NANDI
NANDU-TARI
NANDYA
NATENI
NATIORO
NAWDM
NBWERA
NDAAKA
NDALI
NDAM
NDAMBA
NDAU
NDEBELE
NDEMBU
NDEMLI
NDENDEULE
NDENGEREKO
NDENGESE
NDIBU
NDO
NDOGO
NDOLO
NDOOLA
NDULU
NDUNGA
NDZABI
NDZUNDZA
NEFUSA
NEYO
NGABA
NGALA
NGALANGI
NGAM
NGAM GIR BOR
NGAMBAI
NGAMO
NGANDA
NGANDO
NGANDU
NGANDYERA
NGANGELA
N'GARE
NGASA SHAKA
NGATA
NGBAGA
NGBAKA
NGBANDI
NGEENDE
NGENGE
NGENGELE
NGENZA
NGGWAHYI
NGHWELE
NGINDO
NGIZIM
NGOLA
NGOMBE
NGONGI
NGONGO
NGOUNGWONI
NGUL
NGULU
NGUMBA
NGUMBO
NGUNDI
NGUNGULU
NGUNI
NGURIMI
NGWABA
NHAM
NIABWA
NIELLIM
NIMBARI
NINZAM
NJAJGULGULE
NJAMUS
NKANGALA
NKANSI
NKANU
NKHUMBI
NKOLE
NKOMI
NKOYA
NKUKOLI
NKUTSHU
NKWE
NOBIIN

# Ethnicity (cont.)

NOLE
NOWOLO
NSAMBA
NSAPO
NSENGA
NSONGO
NTANDU
NTCHAM
NTOMBA
NUBA
NUER
NUMANA
NUNA
NUNGU
NUNU
NUNUMA
NUPE
NWENSHI
NWERA
NYAKYUSA
NYALA
NYALI
NYAMWANGA
NYAMWEZI
NYANEKA-HUMBE
NYANGA
NYANGATOM
NYANGBARA
NYANJA
NYARAFOLO
NYATURU
NYEMBA
NYENGO
NYIHA
NYILAMBA
NYIMANG
NYINDU
NYONG
NYORO
NYULI
NZAKARA
NZANYI
NZIMA
OBOLO
OBULKOM
ODUT
OKAK
OKO-ENI-OSAYEN
OKPE
OKPE-AKUKU
OKPELA
OMBO
OMETO
OMONO
OOLI
OPA
OPAMERI
OPUUO
ORANA
ORING
OROMO
ORON
OTANG
OTUHU
OUASSA
PAI
PALLAKA
PAMBIA
PANA
PANDE
PANI
PAPEL
PARE
PARE, SAA
PATU
PAYE
PENDE
PERE
PERO
PEVE
PHOKA
PIMBWE
PINDA
PINDI
PITI
PIYA
PODJULU
PODZO
POGORO
POKOMO
POMBO
POMO
PONGWE
POTO
PUGULI
PUKU
PUNU
PYAANG
QUILENGUE
RASHAD
REGEIBAT
RENDILL
RIF
RIYAH
RONGA
RUARHA
RUBASA
RUFIJI
RUNGA
RUNYANKOLE
RURI
RUSHA
RUUND
RUWENG
SAAB
SAADI
SABA
SAFWA
SAGALA
SAGARA
SAHEL
SAHO
SAKA
SAKALAVA
SAKATA
SALA
SALA MPASU
SAMBA
SAMBA LEKO
SAMBO
SAMBU
SAMBURU
SAN
SANGA
SANGO
SANGU
SANUSI
SAPAUT
SAR
SARA
SARA GULA
SARA GULAY
SARO
SARUA
SASARU
SAYA
SEBA
SEEKU
SEKE
SENA
SENGA
SENGELE
SERE
SERER
SESE
SH
SHAGAWU
SHAIKIA
SHALL
SHAMBAA
SHANJO
SHATT
SHEBELLE
SHERBRO
SHI
SHIKI
SHILA
SHILLUK
SHINJI
SHIRAWA
SHLUH
SHONA
SHOO
SHUBI
SHUKRIA
SHUWA
SIDAMO
SIDI
SIHANAKA
SIMAA
SINASHA
SINGA
SINYAR
SIRTICAN
SISAALA
SISYA
SIWA
SIZAKI
SO
SOKORO
SOLA
SOLI
SOLONGO
SOLU
SOMRAI
SOMYEWE
SONGHAI
SONGO
SONGOLA
SONGYE
SONINKE
SONJO
SOONDE
SOSSO
SOTHO
SOUTHERN BANGANTU
SUAFA
SUBA
SUBIYA
SUGA
SUK
SUKU
SUKUMA
SUKWA
SUMA
SUMBWA
SUNDI
SUNGOR
SURI
SURUBU
SUSU
SWAKA
SWAZI
SYEMU
TABWA
TAFIRE
TAGWANA
TAHOU
TAJAKANT
TAJUASO
TALE
TALENSI
TAMA
TAMAZIGHT
TAMBAHUAKA
TAMBAS
TAMBERAMA
TAMBO
TAMEZRET
TAMPOLENSE
TANALA
TANDA
TANGA
TANGALE
TANGBAGO
TANKARA
TAPSHIN
TAROK
TASUMSA
TATOG
TAWANA
TAZARAWA
TCHWABO
TEDA
TEFASI
TEGE
TEITA
TEM
TEMNE
TENDA
TEPETH
TERA
TERE
TESO
TETELA
TEVUNDRU
TIBA
TIBEA
TIE
TIENE
TIGRAY
TIGRINYA
TIKAR
TIKUU
TIO
TITU
TIV
TOBANGA
TOGBO
TOKA
TOMA
TONGA
TONGA-INHAMBANE
TONGWE
TOPOKE
TOPOSA
TORAM
TORO
TOROBE
TOTELA
TOUGOURT
TOUYO
TOW
TRARZA
TRIBOUE
TRIPOLITANIAN
TSAAM
TSHIAKA
TSIENIMBALALA
TSIMIHETY
TSONG
TSONGA
TUAT
TUKEN
TUKULOR
TULA
TULAMA
TUMBUKA
TUMBWE
TUMTUM
TUNDJUR
TUNGU
TUNISIAN
TUPURI
TURA
TURKA
TURKANA
TURUMBU
TUSIAN
TUTSI
TYEMBARA
TYENGA
UBANG
UHAMI-IYAYU
UJOGOMA
UKAAN
UKPE
UKPET
UKWUANI
ULED NAIL
UMON
UNGA
UREGU
URHOBO
UTUGWANG
UZEKWE
VAGALA
VAI
VEKHEE
VENDA
VERE
VEZO
VIDUNDA
VILI
VIN
VINZA
VIRA
VIYE
VUMA
VUTE
WAAMA
WADA
WAJA
WALA
WALLAGA
WAMBU
WANDYA
WARGLA
WARJI
WASA
WASI ALAWA
WE
WENYA
WILE
WILLE
WINIAMA
WINJI
WOLLO
WOLOF
WOM
WONGO
WOYO
XHOSA
YAA
YAELIMA
YAGOUTE
YAHI
YAKA
YAKA, SUKU
YAKOMA
YALIWASA
YAMAIE
YAMANDUNDU
YAMONGO
YANA
YANZI
YAO
YARSE
YASA
YASAMA
YEKE
YELA
YESKWA
YEYE
YIWOM
YOKO
YOMBE
YORUBA
YOWA
YUKUTARE
YUNGUR
ZAGHAWA
ZANAKI
ZANDE
ZANDE ABANDIA
ZANDE AVUNGARA
ZARAMO
ZARI
ZARMA
ZEEM
ZEKARA
ZELA
ZENAGA
ZIBAN
ZIGUA
ZILMANU
ZIMBA
ZINZU
ZOMBO
ZULA


Language Family
Adamawa-Ubangian
Adamawa-Ubangian / Chari-Nile
Bantoid
Bantu
Bantu / Bantu
Berber
Chadic
Chadic / Cushitic
Chadic / Fufulde
Chari-Nile
Chari-Nile / Nilotic
Cushitic
Fufulde
Fufulde / Adamawa-Ubangia
Fur
Gbaya
Khoi: Nama, Bergdama
Kordofanian
Kru
Kwa
Maban
Malagasy
Miscellaneous / Unclassified
Nilotic
Nilotic / Bantoid
Nilotic / Bantu
Northern Mande
Other
Saharan
Saharan / Cushitic
Saharan / Nilotic
San
Sandawe
Semitic: Arab, Bedouin
Songhai
Southern Mande
Voltaic
West Atlantic

# Appendix 1C

Paper published in Virus Evolution, where this work appears.

Note: As this is an unaltered published manuscript, figures in it and its accompanying supplementary materials are numbered according to the published work and not by thesis chapter. Only page numbers have been added for ease of navigation and reference. Bold monospace font at the bottom centre of each page of the article follow page numbering of the thesis. Variable space font at the bottom right of the first page, and top left of subsequent pages is the original unaltered page numbering (starting at one). The article's online supplementary material has page numbers added according to the thesis.

*Virus Evolution*, 2020, 6(2): veaa085

doi: 10.1093/ve/veaa085
Research Article

# HIV-1 p24Gag adaptation to modern and archaic HLA-allele frequency differences in ethnic groups contributes to viral subtype diversification

Nicolaas C. Kist,[1,2,†] Ben Lambert,[2,3] Samuel Campbell,[1,2] Aris Katzourakis,[2,‡] Daniel Lunn,[4] Philippe Lemey,[5,§] and Astrid K.N. Iversen[1,*,**]

[1]Division of Clinical Neurology, Nuffield Department of Clinical Neurosciences, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, UK, [2]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK, [3]Department of Infectious Disease Epidemiology, School of Public Health, Faculty of Medicine, Imperial College London, Medical School Building St Mary's Campus, Norfolk Place, London W2 1PG, UK, [4]Department of Statistics, University of Oxford, St Giles', Oxford OX1 3LB, UK and [5]Department of Microbiology and Immunology, Rega Institute for Medical Research, KU Leuven - University of Leuven, Leuven B-3000, Belgium

*Corresponding author: E-mail: astrid.iversen@ndcn.ox.ac.uk

[†]https://orcid.org/0000-0002-1081-6594

[‡]https://orcid.org/0000-0002-5665-4041

[§]https://orcid.org/0000-0003-2826-5353

[**]https://orcid.org/0000-0002-3507-5956

## Abstract

Pathogen-driven selection and past interbreeding with archaic human lineages have resulted in differences in human leukocyte antigen (HLA)-allele frequencies between modern human populations. Whether or not this variation affects pathogen subtype diversification is unknown. Here we show a strong positive correlation between ethnic diversity in African countries and both human immunodeficiency virus (HIV)-1 p24*gag* and subtype diversity. We demonstrate that ethnic HLA-allele differences between populations have influenced HIV-1 subtype diversification as the virus adapted to escape common antiviral immune responses. The evolution of HIV Subtype B (HIV-B), which does not appear to be indigenous to Africa, is strongly affected by immune responses associated with Eurasian HLA variants acquired through adaptive introgression from Neanderthals and Denisovans. Furthermore, we show that the increasing and disproportionate number of HIV-infections among African Americans in the USA drive HIV-B evolution towards an Africa-centric HIV-1 state. Similar adaptation of other pathogens to HLA variants common in affected populations is likely.

**Key words:** HIV-1; adaptation; subtype; diversification; HLA; phylodynamics; phylogenetics.

## 1. Introduction

Differences in human leukocyte antigen (HLA) frequencies between geographically dispersed populations are primarily the result of pathogen-driven selection (Doherty and Zinkernagel 1975) and past interbreeding with archaic human lineages (Segurel and Quintana-Murci 2014; Quintana-Murci 2019). Although some gene-flow from modern human into Neanderthals appears to have occurred in Europe as early as 100,000–200,000 years before present (YBP) (Chen et al. 2020), the major dispersal of modern humans from Africa into Eurasia took place ~60,000 YBP (Reich 2018). These modern humans first mated with Neanderthals in the Levant and thousands of years later with at least three distinct Denisovan lineages in Western Asia, East Asia, and Oceania (Reich 2018; Jacobs et al. 2019). Because these groups of archaic humans had already adapted to Eurasian pathogens for ~700,000–450,000 years, the introgression (the gaining of genetic material through interbreeding) of archaic HLA variants likely provided modern humans with pre-adapted HLA variants that were crucial to their survival in the new environments. This survival advantage is evident today as HLA Class I alleles of archaic origin comprise at least 50–85 per cent of the HLA variants in Eurasian populations, even though the overall archaic genomic inheritance is only 1–6 per cent (Abi-Rached et al. 2011). Despite some back-migration of Eurasians to Africa over the last ~20,000 years (Chen et al. 2020), these archaic HLA variants are the key difference between the HLA profile of populations in Africa and Eurasia. Statistical analyses of African whole-genome sequencing data provide evidence of interbreeding with other lineages of archaic humans within Africa, but the results are controversial because of a lack of fossil DNA (Hsieh et al. 2016; Nielsen et al. 2017). What impact the HLA variation between populations over time has had on the subtype diversification of both old and novel pathogens such as human immunodeficiency virus (HIV)-1 is largely unknown.

The cross-species transmission of simian immunodeficiency virus that gave rise to the HIV-1 group responsible for the majority of the global HIV-1 epidemic (Group M), likely occurred ~1900 in Cameroon or the Democratic Republic of the Congo (DRC; Worobey et al. 2008). However, the HIV-1 pandemic only ignited in Kinshasa in the DRC in the 1920s and spread over the following decades via transport networks and migrant labour throughout the country and beyond (Faria et al. 2014). HIV-1 started to diversify into subtypes around 1960 concurrent with the transition from a slow to a much faster transmission phase (Faria et al. 2014).

HIV-1 subtypes are defined using phylogenetic analyses and the genetic distance between them is comparable; they are labelled alphabetically, and each subtype consensus sequence carries distinct combinations of subtype-specific amino acids in conserved subtype-specific positions in every viral protein in addition to changes at other positions (Foley et al. 2018). These subtype-specific amino acid differences are particularly striking in the conserved, clinically important, highly immunogenic cytotoxic T lymphocyte (CTL) target regions of p24Gag (the capsid protein) where they constitute the only difference between the subtype consensus sequences (Supplementary Fig. S1; Kiepiela et al. 2007; Matthews et al. 2008; Goulder and Walker 2012).

The highest number of HIV-1 subtypes and circulating and unique recombinant forms (CRFs and URFs) of various HIV subtypes is found within Africa (Chang et al. 2015; Foley et al. 2018). Whereas subtype-precursor lineages to a high degree are embedded within the larger HIV-1 diversity in the DRC, which also includes examples of most global subtypes, only distinct subtypes predominate outside the country (Rambaut et al. 2001; Faria et al. 2014). The most geographically widespread subtype is HIV-B (Hemelaar et al. 2011). The ancestral HIV-B virus originated in Africa and spread to Haiti around 1965–70 (Gilbert et al. 2007). The earliest HIV-B sequences in the USA were detected in stored serum samples collected from men-who-have-sex-with-men (MSM) in 1978–79 (Gilbert et al. 2007; Worobey et al. 2016). HIV-B spread within the USA and beyond and now accounts for ~11 per cent of worldwide infections, although no indigenous (i.e. not imported) HIV-B subtype has ever been identified in Africa (Hemelaar et al. 2011). In contrast, an HIV-C-precursor lineage spread successfully from the southern DRC Katanga province to most of the Southern and Eastern parts of Africa, and today HIV-C is responsible for ~50 per cent of all infections (Hemelaar et al. 2011; Faria et al. 2014).

Upon infection, HLA Class I (A–C) molecules on the cell surface present peptide fragments (epitopes) from viral proteins that have been digested by intra-cellular proteasomes. This presentation is a central point in the interplay between HIV-1 and the host's immune system as it triggers CTL and Natural Killer cell responses that can kill HIV-infected cells. The HLA genes are highly polymorphic, and an HLA molecule has specific binding motifs that allow only some of the produced epitopes to be presented by each HLA variant. Because the combination of HLA variants (or the HLA profile) varies between individuals, the selective pressures on the virus vary accordingly (Tenzer et al. 2014). The variation in HLA frequencies between populations and ethnic groups ensures that the sum of the HLA-associated selective pressures on the virus will also differ.

HLA-associated polymorphisms have primarily been examined within distinct HIV-1 subtypes in combined Canadian/US HIV-B sequence sets (Cotton et al. 2014; Kinloch et al. 2016) after Bhattacharya et al. (2007) reported that viral subtype effects confounded previously reported HLA associations (Moore et al. 2002). Many of the polymorphisms are intra-epitope CTL escape mutations, and studies have focussed on the presence, accumulation, and possible reversion of small subsets of these mutations (Supplementary Fig. S2A; Leslie et al. 2004, 2005; Kawashima et al. 2009; Fryer et al. 2012) and their impact on viral fitness and viral load set point (Matthews et al. 2008; Carlson et al. 2015). Although HLA-associated polymorphisms are found in all HIV-1 proteins, analyses of the specificity and breadth of CTL responses associated with immune control of HIV-1 have demonstrated that responses targeting p24Gag were almost exclusively associated with lowering viremia (Kiepiela et al. 2007; Matthews et al. 2008; Goulder and Walker 2012).

We experimentally identified a second form of HLA-associated selection which occur at subtype-specific positions in p24Gag and affects proteasomal production of all epitopes encoded in epitope-clusters up- and down-stream of the subtype-specific position (illustrated in Supplementary Fig. S2B and explained in detail in the legend; Tenzer et al. 2014). These epitope-clusters typically contain ~20–50 epitopes presented by approximately as many HLA variants and proteasomal production of each epitope is either increased, unaffected, decreased, or eliminated depending on the nature of the amino acid at the subtype-specific position. We found that intra-host selection on HIV-1 favoured an amino acid at the subtype-specific position which thwarted or limited production of HIV-1 epitopes that that individual's HLA variants could present regardless of the infecting subtype. At the population level, we observed an inverse relationship between the amount of each epitope produced following processing of the HIV-B or -C consensus

sequences and the frequencies of the presenting HLA variants in the populations in which these subtypes circulated (Supplementary Fig. S2C and D; Tenzer et al. 2014).

Intra-host selection on HIV-1 to escape anti-viral CTL responses through intra-epitope escape-mutations and changes at subtype-specific sites that limit epitope-processing likely co-occurs. However, within the conserved, highly immunogenic, p24Gag regions (Supplementary Fig. S1), subtype-specific positions are subjected to the combined pressure by >80 HLA variants (HIV Molecular Immunology Compendium 2016) and changes at these positions at the HIV-1 consensus level are likely observed faster than changes within single epitopes restricted by one, or a few, HLA variants (Kawashima et al. 2009).

The current model for HIV-1 subtype diversification is that HIV-1 subtypes are the results of multiple, selectively neutral, random 'founder events', whereby individuals with distinct viral lineages moved to new regions and established local epidemics (Pybus and Rambaut 2009; Peeters, Jung, and Ayouba 2013). HLA variation and hence variation in immune response targeting between populations are generally not assumed to play a significant role in this process (Pybus and Rambaut 2009; Peeters, Jung, and Ayouba 2013). Here we considered the effect of HLA variation between populations and HLA-associated selection at subtype-specific sites through modification of epitope production (Tenzer et al. 2014).

First, we investigated whether differences in ethnic diversity in African countries was associated with HIV-1 p24Gag diversity (on the amino-acid level ignoring subtype) and subtype diversity, respectively, and observed robust positive linear relationships. Second, we explored the HLA profile patterns in all HIV-B and HIV-C-infected patients with linked HLA information using an HLA-based principal component analysis (PCA). We found that the most influential drivers in principal component (PC)1 were Eurasian HLA variants introgressed from archaic humans, whereas HLA variants associated with the two major haplotypes in South Africa strongly influenced PC2. Third, we estimated which of the >80 epitope-presenting HLA variants in the clinically important HIV-B and HIV-C p24Gag regions had the most substantial effect on the evolution of the five subtype-specific sites within these regions. Fourth, we examined if the continuously increasing proportion of African Americans in the USA HIV-B infected population over time affected the evolution of these subtype-specific sites. We found that the subtype-specific sites in HIV-B-p24Gag in the USA over time increasingly incorporated amino acids more commonly found in African HIV-1 subtypes than in the HIV-B consensus sequence.

Collectively, our results suggest a modified model for HIV-1 subtype diversification. We propose that variation at subtype-specific sites in HIV-1 p24Gag results from a combination of global HIV-1 spread through random founder events and the subsequent continuous granular viral adaptation to the HLA profile of the infected population.

## 2. Results

### 2.1 African HIV-1 p24Gag and subtype diversity correlate with ethnic diversity

We hypothesized that HLA-mediated selection drove HIV-1 p24Gag diversification and that subtype-specific amino acid differences were the result of selection for HIV-1 sequences that limited or abrogated processing of the epitopes presented by the most common HLA variants in each population. If HLA-mediated selection drove HIV-1 diversification, we should find greater HIV-1 diversity in countries with greater HLA diversity.

To investigate, we examined HIV-1 diversity within Africa where modern humans originated and have maintained relatively large effective population sizes for ~300,000 years resulting in a high level of genetic diversity within and between the ~2000 distinct African ethnic groups (Felix and Meur 2001; Patin et al. 2017; Tishkoff et al. 2009; Hershkovitz et al. 2018). Unfortunately, HLA data from Africa is limited as gold standard HLA A, B, and C data (i.e. all HLA types being reported as four-digit HLA variants) from only eight Sub-Saharan African populations/ethnic groups are available in the HLA-allele database (Gonzalez-Galarza et al. 2015; Supplementary Table S1). As a proxy for HLA diversity we used the Shannon entropy of each country's ethnic demographics reported by Alesina et al. (2003) and Busby et al. (2017), which is based on ethnic, linguistic, cultural, and religious fractionalization.

Using this measure of ethnic diversity as a proxy for HLA diversity is justified by multiple studies that have demonstrated that African population structure largely mirrors linguistic and geographic similarity (Tishkoff et al. 2009; Busby et al. 2016, 2017) and have provided evidence of a strong link between genetic diversity and ethno-linguistic diversity (Busby et al. 2016, 2017; Excoffier et al. 1987, 1991; Tishkoff et al. 2009; Pagani et al. 2012; Patin et al. 2017). This genetic diversity is even higher in the HLA region as local adaptation at HLA loci generally are more pronounced than in the rest of the genome due to strong selective pressures by local pathogens and adaptive admixture between expanding populations and adapted local groups (Cao et al. 2004; Busby et al. 2017; Patin et al. 2017; Meyer et al. 2018). The strong link between ethnicity, genetics, and linguistics is not found on all continents as, for example, in South America, ethnic groups are largely defined by racial or physical criteria, not linguistics (Alesina et al. 2003).

To further justify using ethnic diversity of African countries as a proxy for HLA diversity, we performed simulations showing a linear relationship between different mixtures of fictive ethnic groups and HLA diversity using the HLA A, B, and C data from the eight Sub-Saharan African populations/ethnic groups (Fig. 1).

Pooling our results across all replicates and numbers of ethnic mixing populations, we found a strong and statistically significant relationship between ethnic diversity and HLA diversity ($R^2 = 0.41$, $F_{1,\ 1,748} = 1{,}190$, $P < 2 \times 10^{-16}$). Furthermore, this result held for subsets of the data comprising simulations of populations from a given number of ethnicities. Collectively, these simulations underscored that ethnic diversity in Africa is a valid proxy for HLA diversity.

Next, we examined the relationship between the actual combined measure of ethnic diversity (Alesina et al. 2003) and HIV-1 p24Gag diversity using all HIV-1 p24Gag sequences in the HIV database regardless of subtype (Foley et al. 2018). We observed a highly statistically significant relationship between ethnic diversity and p24Gag diversity ($F_{1,13} = 15.53$, $P = 0.0017$; Fig. 2A). To examine if the combined measure of ethnic diversity also affected HIV-1 subtype diversity, we counted the number of subtypes, URFs and CRFs in each country using the HIV database sequences and allocation of subtype data ($n = 36{,}044$) (Foley et al. 2018). The total number of subtypes, URFs and CRFs per country was plotted against the ethnic diversity for each country weighted by the number of sequences per country and examined using linear regression. Similar to the result for HIV-1 p24Gag diversity, a significant relationship was found between ethnic diversity and subtype diversity ($P < 2 \times 10^{-5}$; Fig. 2B).

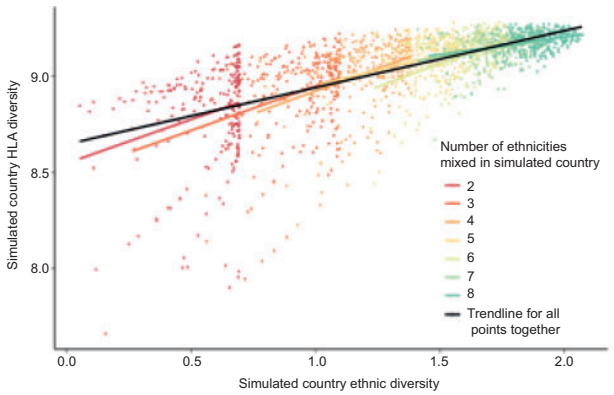



**Figure 1.** The simulated ethnic diversity of a fictitious country correlates with its simulated HLA diversity. Eight 'populations' were mixed in arbitrary proportions to simulate the mixture of ethnic groups within different fictive African 'countries'. We then mixed the real gold standard HLA-allele frequency data from the eight populations in the HLA-allele database (Gonzalez-Galarza et al. 2015; Supplementary Table S1) according to the simulated proportions and performed linear regressions of the HLA diversity of these 'simulated countries' versus the simulated countries' ethnic diversity calculated from the simulated proportions. Every point represents a simulated country; 500 countries were simulated for each number (k) of ethnic groups between 2 and 8 within each country. Each country consists of k randomly chosen populations/ethnic groups from Supplementary Table S1. Black line: best-fit line following a linear regression analysis of all simulated countries ($P < 2 \times 10^{-16}$). Coloured lines: best-fit line following a separate linear regression analysis of simulated countries consisting of k 'ethnic groups'.

To further test robustness, we excluded Cameroon and the DRC where HIV-1 first circulated (Worobey et al. 2008; Faria et al. 2014), which still resulted in significant relationships (ethnic diversity versus HIV-1 p24*gag* variability ($P = 0.0134$) and subtype diversity ($P < 2 \times 10^{-5}$)) indicating that the timing of the epidemic in each African country was not the sole determinant of HIV-1 diversification. Collectively, these results suggest that the association between a country's HIV-1 diversity and its ethnic diversity might be driven by the granular adaptation of HIV-1 to the HLA variants common in particular ethnic groups as countries with more ethnic groups end up with greater HIV-1 variation.

Geographically, the highest concentration of ethno-linguistic (Felix and Meur 2001), ethnic (Alesina et al. 2003) and HIV-1 subtype diversity was found along the equator, with, for example, ~110 and ~370 ethno-linguistic groups within Cameroon and the DRC, respectively (Felix and Meur 2001; Fig. 2C and D). As ethnic differences typically correlate with language differences in Africa (Tishkoff et al. 2009; Busby et al. 2016, 2017), we examined if HIV-1 subtypes clustered within specific language families or sub-groups. Most African ethno-linguistic groups belong to one of five language families (Niger-Congo, Nilo-Saharan, Afro-Asiatic, Austronesian, and Khoe, Fig. 2E), which each can be subdivided into several language groups, for example, the Niger-Congo family includes Bantu-speaking people that account for approximately one-third of all Sub-Saharan Africans (Patin et al. 2017) and non-Bantu-speaking people in West Africa.

We observed more similar HIV-1 subtype patterns in related ethno-linguistic groups (Tishkoff et al. 2009; Busby et al. 2016, 2017; Fig. 2D and E). Notably, HIV-C predominated in countries with the highest percentage of Bantu-speaking people (Patin et al. 2017). One exception was the many HIV-1 subtypes found in the DRC where adaptive introgression of HLA from rainforest hunter-gatherers during the Bantu expansion ~3,000 years ago significantly increased the HLA variability of western and central Bantu-speaking people relative to that of southern Bantu-speaking people (Patin et al. 2017; Foley et al. 2018). Moreover, HIV-C was only introduced into the Horn of Africa around 1980–85 (Simmons et al. 2009), and has since, for example, in Ethiopia evolved into two distinct Ethiopian HIV-C lineages (Amogne et al. 2016). We explored whether countries which had at least 50 per cent Bantu First Language Speakers were more likely to have at least 50 per cent subtype C in their HIV-infected population. This analysis included all African countries regardless of epidemiological history and demonstrated a link between high levels of Bantu language speaking inhabitants and high levels of HIV-C (Fisher's test, $P = 0.037$). Niger-Congo non-Bantu language groups reside in West Africa where CRF_02AG and HIV-G predominate, and HIV-C is rare. Nilo-Saharan, Afro-Asiatic and Bantu language groups predominate in East Africa where HIV-A, -D, and -C are frequently found (Fig. 2D and E).

Together with the explosion of HIV-1 subtypes in the 1960s (Faria et al. 2014), our results suggest that ethnic diversity, and hence HLA diversity, was a significant driver of HIV-1 diversification, and that neutral evolution and/or differences in the timing of local African HIV-1 epidemics was not solely responsible for the creation of HIV-1 subtypes as suggested previously (Pybus and Rambaut 2009; Peeters et al. 2013).

## 2.2 Spatio-ethnic HLA profiles associated with distinct HIV-1 subtypes

We next asked which HLA variants contributed most to the variance–covariance structure of the worldwide HIV-infected patient population by performing a PCA on all publicly available HLA Class I data from patients infected with HIV-B ($n = 318$) or HIV-C ($n = 704$; Foley et al. 2018). Because of the history of HIV-1 research, similar data on patients infected with other subtypes are rare, and most African data derive from South Africa, especially the KwaZulu-Natal province (Foley et al. 2018). We assumed that the distribution of HLA variants and HLA haplotypes (a series of HLA genes on one chromosome that is found together more often than expected due to linkage disequilibrium (LD)) in geographically dispersed populations reflected the sum of past infectious challenges and adaptive introgression from archaic humans (Doherty and Zinkernagel 1975; Segurel and Quintana-Murci 2014; Quintana-Murci 2019). For each member of the population, we proposed that their genetic 'HLA profile' abstractly could be considered as a point in the population's multidimensional HLA profile space. A novel pathogen such as HIV-1 would rapidly spread and evolve within this HLA space. The space itself would only change slowly due to the added selective effect of HIV-1 infection on HLA frequencies and largely outside of our current observation time.

The individual HLA variants which contributed most to the first two PCs clustered in three groups, reflecting three common haplotypes (Fig. 3A): one group containing Eurasian variants (HLA-A*02:01, HLA-B*07:02, and HLA-C*07:02) and two groups of HLA variants more common in Africa (HLA-B*58:02, HLA-C*06:02, and HLA-A*30:01, HLA-B*42:01, HLA-C*17:01; Fig. 3B and Supplementary Table S6). The first two PCs explained 14.5 per cent of the variance and the first 12 explained 50 per cent. PC1 revealed four features: (i) the HLA separation correlated with the geographical origin of the patients with Africans and Eurasians at the extremes (Fig. 3C), (ii) the Caribbean patients were near the centre, reflecting the history of admixture in that population (Fig. 3C; Benn-Torres et al. 2008), (iii) the separation correlated with the distribution of the patients' HIV-1 subtypes

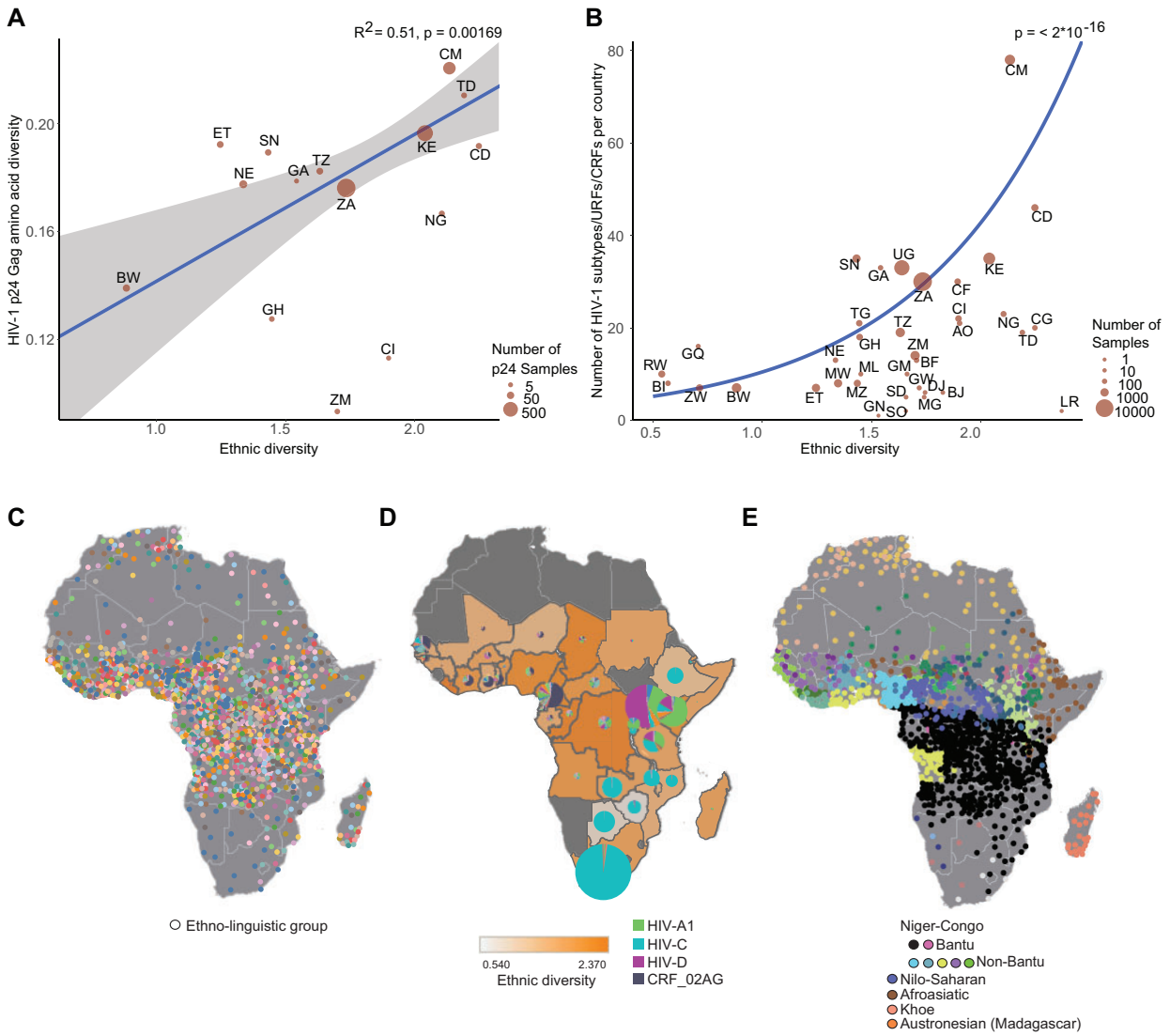



**Figure 2.** Ethnic diversity drives HIV-1 p24Gag and subtype diversification. (A) Weighted linear regression of HIV-1 p24Gag diversity versus ethnic diversity (Alesina et al. 2003; each determined using Shannon entropy). All Sub-Saharan African countries with ≥ 6 p24Gag sequences were included (Foley et al. 2018) (the number of sequences and the ISO alpha-2 two-letter country codes are shown in Supplementary Table S2). Grey-shaded area represents the 95 per cent confidence interval (CI). (B). Weighted linear regression of the count of HIV-1 subtypes, URFs and CRFs (Foley et al. 2018) versus ethnic diversity (Alesina et al. 2003) within Sub-Saharan Africa (illustrated in D) using a generalized linear model fit with a Poisson error distribution with a log link function (the number of subtypes, CRFs, and URFs and the ISO alpha-2 two-letter country codes are shown in Supplementary Table S2). Grey-shaded area represents the 95 per cent CI (too small to see). (C) Map illustrating the geographical location of ~2,000 ethno-linguistic groups in Africa (modified from Felix and Meur 2001). Key in Supplementary Table S3. (D). Map showing the ethnic diversity (Alesina et al. 2003) within Sub-Saharan Africa (orange shaded background) overlaid with pie charts demonstrating HIV-1 subtype and CRF diversity within each country (Foley et al. 2018). Missing countries (dark grey) lacked either HIV-1 or ethnic fractionalization data. The pie charts are scaled according to the number of unique patient HIV-1 sequences; a not scaled version is shown in Supplementary Fig. S3; complete key in Supplementary Table S4. (E) Map illustrating the current location of the five major language families within Africa with predominant groups indicated (modified from Felix and Meur 2001); complete key in Supplementary Table S5.

(Fig. 3D), and, (iv) the three HLA variants that contributed most to the separation were all inherited by Eurasians from Neanderthals and Denisovans (Abi-Rached et al. 2011) with HLA-A*02:01 explaining most of the variance (Fig. 3E).

PC2 robustly distinguished between two common Southern African HLA haplotypes with opposing effects on HIV-1 disease progression; HLA-B*58:02-HLA-C*06:02, linked to fast progression (found in ~20% of the population), and HLA-A30:01-HLA-B*42:01-HLA-C*17:01, linked with slow progression through the accumulation of fitness-reducing CTL escape-mutations (found in ~13%; (Goulder and Walker 2012; Gonzalez-Galarza et al. 2015; Fig. 3F). In Zambia, HLA-B*42:01 is also linked to a faster HIV-1 acquisition (allele frequency ~30%) (Song et al. 2011; Gonzalez-Galarza et al. 2015). We calculated a haplotype score, which started at zero and was incremented or reduced by one for each allele in these two haplotypes, revealing the near-complete stratification of PC2 (Supplementary Fig. S4). Although these HLA variants can occur in other combinations, the LD is almost complete in Bantu-speaking South Africans and other Southern Bantu-speaking people in countries with the highest levels of HIV-C infection (Fig. 3D, F, and Supplementary Fig. S5; Paximadis et al. 2012; Gonzalez-Galarza et al. 2015).

These results suggest that the HLA variants that contributed most to the variance–covariance structure within the HIV-1

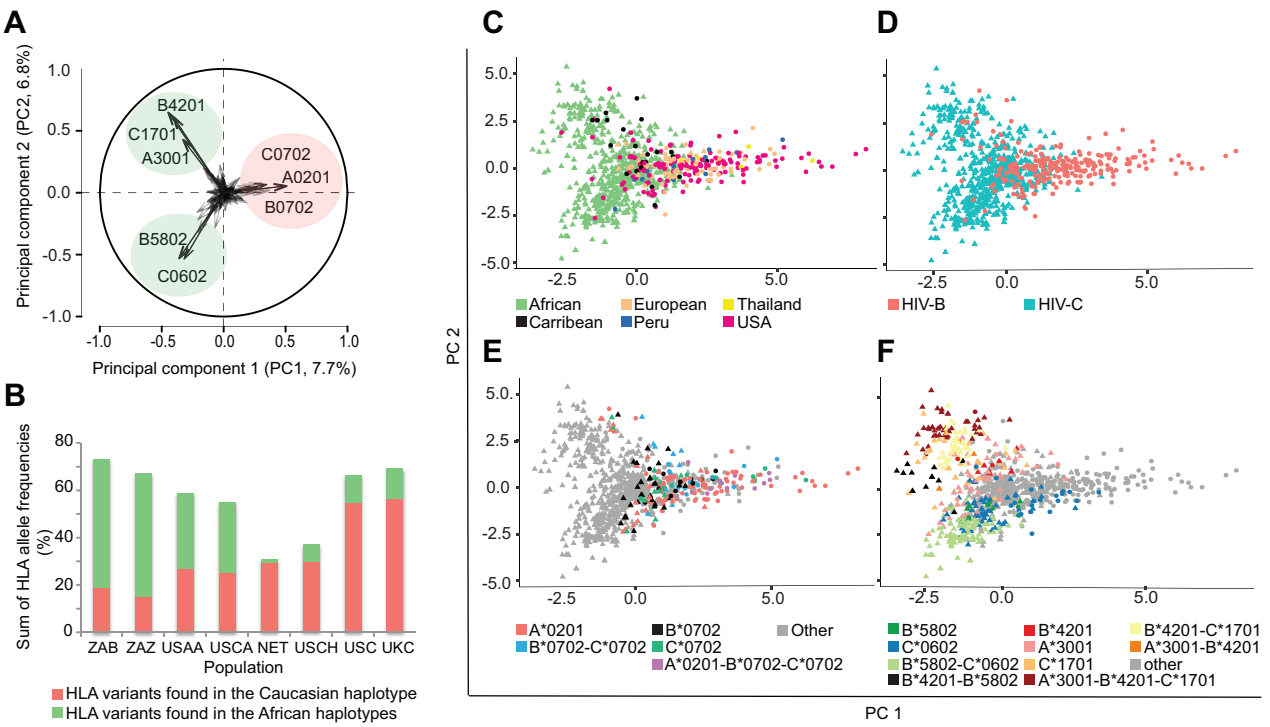



**Figure 3.** HLA PCA results coloured by country/region, HIV-subtype, and HLA-allele variants. (A) Loading plot of the direction vectors associated with each HLA type. The closer the value is to 1 (or −1) the greater the effect of the component on the variable. A patient's position in (C–F) is obtained by summing the values for the patient's HLA-alleles. HLA variants within African (green circle) and Caucasian (red circle) haplotypes are indicated. (B) The combined HLA-allele frequency of the individual HLA variants found in the African haplotypes (green) and the Caucasian haplotype (red) in eight populations: South African Black (ZAB), South African Zulu (ZAZ), US African American (USAA), US Caribbean (USCA), North-East Thai (NET), US Chinese (USCH), US Caucasian (USC), and, UK Caucasian (UKC). Individual HLA-allele frequencies are reported in Supplementary Table S6. Scatter plot of HLA PC1 and PC2 coloured according to: (C) geographical origin of the patient, (D) whether the patient was infected by HIV-B (red circle) or HIV-C (turquoise triangle), (E) the position of Caucasian, and (F) of predominately African HLA variants that explained the majority of variation in PCs 1 and 2.

patient population worldwide originated from interbreeding between the ancestors to modern Eurasians and Neanderthals and Denisovans. An African HIV-1 lineage imported to Eurasia would, therefore, circulate in a partly new HLA profile space. Within Africa, we observed that the HLA variants that explained most of the variance were part of common haplotypes, which are associated with opposing HIV-1 disease progression patterns and facilitated HIV-1 acquisition in some regions.

### 2.3 Distinct ethnic HLA variants affect HIV-B and -C evolution

To estimate the selective pressure of individual HLA Class I variants on the diversification of subtype-specific amino acids in a population, we analysed all p24Gag HIV-B and -C sequences linked to an HLA profile (Foley et al. 2018) using a Bayesian phylogenetic HLA selection model. The analyses were performed on each subtype separately. We focussed on five subtype-specific positions (27, 41, 116, 120, and 128) situated within the two conserved p24Gag regions that dominate the clinically effective HIV-specific CTL response in patients infected by either HIV-B or -C (Goulder and Walker 2012; Fig. 4A and B and Supplementary Fig. S1). Because each subtype-specific position is independently influenced by the phylogeny (Supplementary Figs S6 and S7), the model was corrected for phylogenetic co-variation using the BEAST-derived phylogenetic variance–covariance matrix for each subtype-specific position. This correction adjusted for potential viral population structure and neutral evolution. Because some subtype-specific positions vary between subtypes, but only have limited variation within the given subtype (position 128 in HIV-B, and positions 27, 41, 116, and 128 in HIV-C), or too much variation for these analyses given our dataset (position 120, HIV-B; Supplementary Fig. S1), HLA associations could only be robustly determined for three positions in HIV-B and one in HIV-C. At the HIV-B positions 27, 41, and 116, selection typically favours one of two amino acids, and at position 120 in HIV-C, one of three (Supplementary Fig. S1). The total HLA pressure on each subtype-specific position could be considered to be the joint pressure exerted by all the HLA variants able to present epitopes encompassing (Tenzer et al. 2014) or overlapping that subtype-specific position (Foley et al. 2018; Fig. 4A and B), and by HLA variants associated through structural codon-co-variation (Carlson et al. 2008; Crawford et al. 2011). Significant statistical links between an HLA variant and selection for either HIV-B or non-HIV-B-like amino acids at the four subtype-specific positions were expressed as odds ratios (ORs).

Overall, ~50 per cent of the Caucasian/Eurasian HLA A, B, and C variants that shaped HIV-B evolution originated from Neanderthals and/or Denisovans, which is similar to the frequencies of these variants in people of European descent (Abi-Rached et al. 2011; Fig. 4C and D). A minority of HLA variants appeared to be footprints of the on-going HIV-B epidemic in the Caribbean and in African-Americans (e.g. HLA A*23:01, HLA B*57:03, and HLA C*16:01; Fig. 4D). Although most HLA variants selected for HIV-B-like subtype-specific amino acids, some selected for non-B likeness (e.g. HLA C*16:01 (Position 27) and HLA A*31:01 (Position 41)). These results are in line with previous

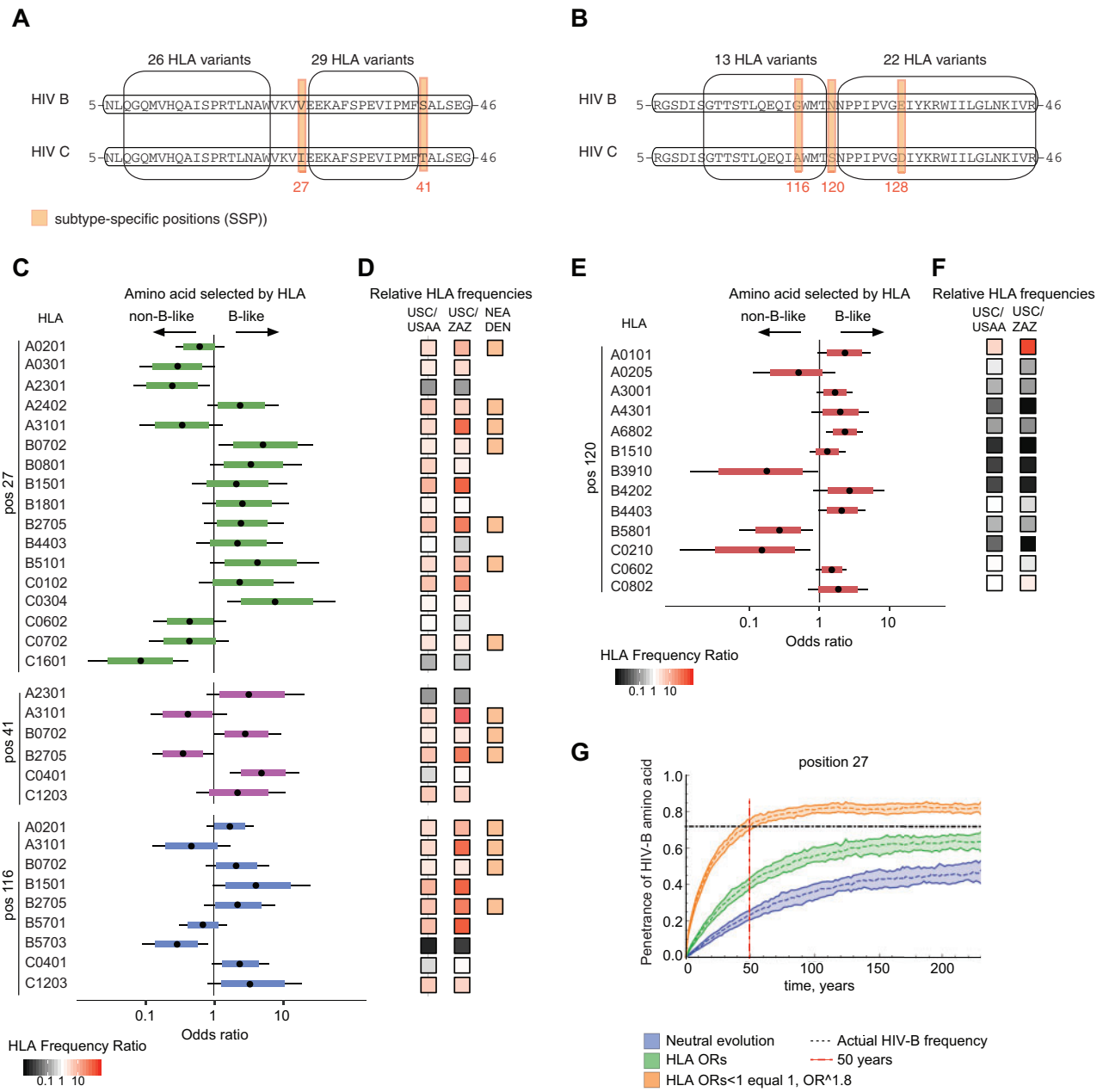



**Figure 4.** HLA-associated selective pressure on HIV-B and HIV-C. (A and B) The two conserved p24Gag HIV-1 regions. Ellipses indicate epitope-clusters with the approximate number of presenting HLA variants noted above. Orange rectangles signify subtype-specific positions (Supplementary Fig. S1). (C) HIV-B: The posterior probability distributions on HLA-associated ORs; ORs greater, or less than one, represent the belief that an HLA variant is associated with the HIV-B or non-HIV-B (i.e. African) consensus amino acid in that position, respectively. The central point represents the median, the coloured region is the 80 per cent Bayesian credible interval (BCI), and the whiskers represent the 95 per cent BCI. (D) The relative HLA-allele frequencies in US Caucasians (USC) compared with African-Americans (USAA; left hand column), and versus South African Zulus (ZAZ) (a Bantu-speaking people; middle column). HLA variants identified in DNA samples from three Neanderthals (NEA) and one Denisovan (DEN; Abi-Rached et al. 2011) are indicated (orange squares). (E) HIV-C: The posterior probability distributions on HLA-associated ORs. As in (C). (F) As in (D), none of these HLA variants were identified in Neanderthals or the Denisovan. (G) An agent-based model of neutral and HLA-mediated selection effects on HIV-C p24Gag, Position 27. The effects of neutral (blue) and HLA-mediated selection using the estimated HIV-B HLA ORs (green), and the increased HLA ORs for HIV-B-likeness (orange) are shown.

analyses of intra-host adaptation of HIV in (Tenzer et al. 2014), which demonstrated that the consensus HIV-B (or HIV-C) sequence was often observed in carriers of common HLA variants whereas carriers of rare HLA variants frequently selected for non-consensus amino acids. The data are consistent with the variation within our patients (illustrated in Supplementary Fig. S8) and in HIV-B sequences worldwide (Foley et al. 2018). In contrast, HIV-C evolution was shaped by African HLA variants common in Bantu-speaking people and one ancient HLA variant (HLA-A*01:01) found in parts of Africa (Abi-Rached et al. 2011; Fig. 4E and F), consistent with the spread of the HIV-C epidemic (Gonzalez-Galarza et al. 2015).

Our estimated HLA ORs (Fig. 4C and E) do not reflect the HLA selection pressure acting on the HIV-B-precursor virus during

the first 30–40 years of the HIV-1 epidemic outside of Africa, because >80 per cent of our data were from 2001 or later (Foley et al. 2018), and because the epidemiological demographics have changed from predominately Caucasian until the mid-nineties to one with mixed ethnicities (Worobey et al. 2016).

To examine how the current HLA ORs and neutral selection would affect HIV-B evolution we used an agent-based epidemic model. As the first HIV-1 lineage in Haiti is unknown (Rambaut et al. 2004), we chose to test an HIV-C-like virus, which differs from HIV-B at subtype-specific positions 27, 41 and 116 (Supplementary Fig. S1). We found that neutral evolution of HIV-C would take ~250 years to approach a 50 per cent HIV-B subtype-specific amino acid frequency (position 27, Fig. 4G, blue line; positions 41 and 116, Supplementary Fig. S9), whereas HLA-mediated selection would reach this threshold more than twice as fast (Fig. 4G, green line). This result demonstrates that the estimated HLA ORs are not only statistically significant but also biologically relevant.

Next, we tried to imitate the HLA selective pressure acting on the HIV-B precursor early in the epidemic. We ignored the selective pressure from African HLA variants and modelled how much the HLA ORs selecting for HIV-B-likeness should be increased to reach the current consensus subtype-specific amino acid frequencies within 50–55 years, a time span similar to that of HIV-B evolution (Worobey et al. 2016). We found that the HLA ORs had to be raised by a factor of 1.8–2.5 depending on subtype-specific position (Position 27, Fig. 4G, orange line; Positions 41 and 116, Supplementary Fig. S9). This result suggests that the initial HLA-mediated selective pressure on the HIV-B-precursor was either stronger than our estimates and/or that we are missing the effect of some HLA variants due to viral adaptation before 2001. Alternatively, or additionally, some of the subtype-specific amino acids in the HIV-B-precursor might have been more similar to HIV-B than HIV-C, and/or structural co-evolutionary effects (Carlson et al. 2008; Crawford et al. 2011) might be stronger than expected.

Collectively, our analyses reveal that HIV-B and HIV-C circulate in populations with distinct, and partly non-overlapping, HLA profiles and that the difference in population-specific HLA-mediated selective pressures affected HIV-1 evolution.

### 2.4 The emergence of an African American-adapted HIV-B

If the subtype-specific differences between HIV-B and HIV-C are the result of different HLA-allele frequencies between Eurasian and African populations, changes in the ethnic demographics of an HIV-infected population should affect the selective pressure at the subtype-specific positions. To investigate this hypothesis, we examined the HIV-1 epidemic in the USA. HIV-B was likely introduced in the 1970s (Worobey et al. 2016) and initially circulated primarily amongst Caucasian MSM. However, the epidemic gradually spread to other groups, and, regardless of sexual orientation, African and Hispanic Americans have increasingly been disproportionally affected (Fig. 5A). A complex set of socioeconomic factors drive risk to these groups, including discrimination, stigma, poverty, lack of access to care, and cultural differences combined with in-group sexual bias (Centers for Disease Control and Prevention 2020, www.cdc.gov). Currently, African and Hispanic Americans account for ~42 and 20 per cent of all new infections despite making up only about 13 and 16 per cent of the US population, respectively (Hall et al. 2008; Prejean et al. 2011).

Based on our previous HLA diversity analysis (Fig. 1), we proposed that the increased fraction of non-Caucasians in the US HIV-infected population would increase its HLA diversity. To investigate, we performed simulations using the HLA A, B, and C frequency data for US Caucasians and African Americans reported by Gragert et al. (2013). Briefly, we created mixed populations of the two groups and calculated the HLA Shannon entropy (diversity) of the simulated mixed populations (HLA B is shown in Fig. 5B, all HLA simulations and plots of HLA frequencies are shown in Supplementary Fig. S10). The results demonstrated that the HLA B and HLA C diversity peaked when the two groups were present in approximately equal proportions, whereas the HLA A diversity continuously increased as the proportion of African Americans grew. These differences were due to the relatively lower diversity of US Caucasian HLA A alleles and the more similar diversity of HLA B and C alleles between the groups (Supplementary Fig. S10). For simplicity, we did not add HLA data from US Hispanics, which are more like those of US Caucasians. Adding these would only increase the overall diversity.

Because 73–78 per cent of the African American genetic heritage originates from Africa (~30% from Bantu-speaking people and ~70% from regions of West Africa; Patin et al. 2017), we hypothesized that the HIV-B subtype-specific amino acids would evolve towards amino acids more commonly found in African HIV-1 strains than in HIV-B (Fig. 5C). We estimated the ethnic demographics of the HIV-infected population in the USA from (Hall et al. 2008) and (Prejean et al. 2011) and used that demographic data to calculate the frequency of African Americans in the US HIV-infected population over time (Supplementary Fig. S11; Gonzalez-Galarza et al. 2015).

We compiled all dated HIV-B p24Gag sequences associated with a patient ID (to prevent oversampling sequences from single patients) and used these in a Hierarchical Bayesian multilevel model that incorporated phylogenetic information from BEAST to prevent phylogenetic relationships from biasing the analysis, thereby correcting for viral population structure and neutral evolution (Supplementary Fig. S12). The posterior distributions over the regression coefficients of the phylogenetic model robustly demonstrated that the probability that the 41, 116, and 120 subtype-specific position was occupied by an HIV-B subtype-specific amino acid was inversely associated with the frequency of African HLA variants in the infected population (Fig. 5D). Although the pattern was similar for Position 27 and the 80 per centBCI excluded zero, the 95 per cent BCI did not quite exclude zero.

Collectively, our analyses demonstrated that the subtype-specific positions in HIV-B over the last 20 years in the USA have evolved away from the original HIV-B consensus sequence and increasingly incorporate amino acids found in African HIV-1 subtypes (Fig. 5E, Supplementary Fig. S12, and Table S7). These results show that HIV-1 diversification is an on-going process that can be influenced by changes in the ethnic composition of the HIV-infected population within a country. As African Americans continue to suffer a disproportionate number of HIV-infections, HIV-B is likely to continue to adapt to the HLA variants common in this group.

## 3. Discussion

In this study, we found that HLA Class I allelic frequency differences and archaic HLA Class I variants combine to create population-specific HLA-associated selective pressures that uniquely shape HIV-1 p24Gag variation and subtype

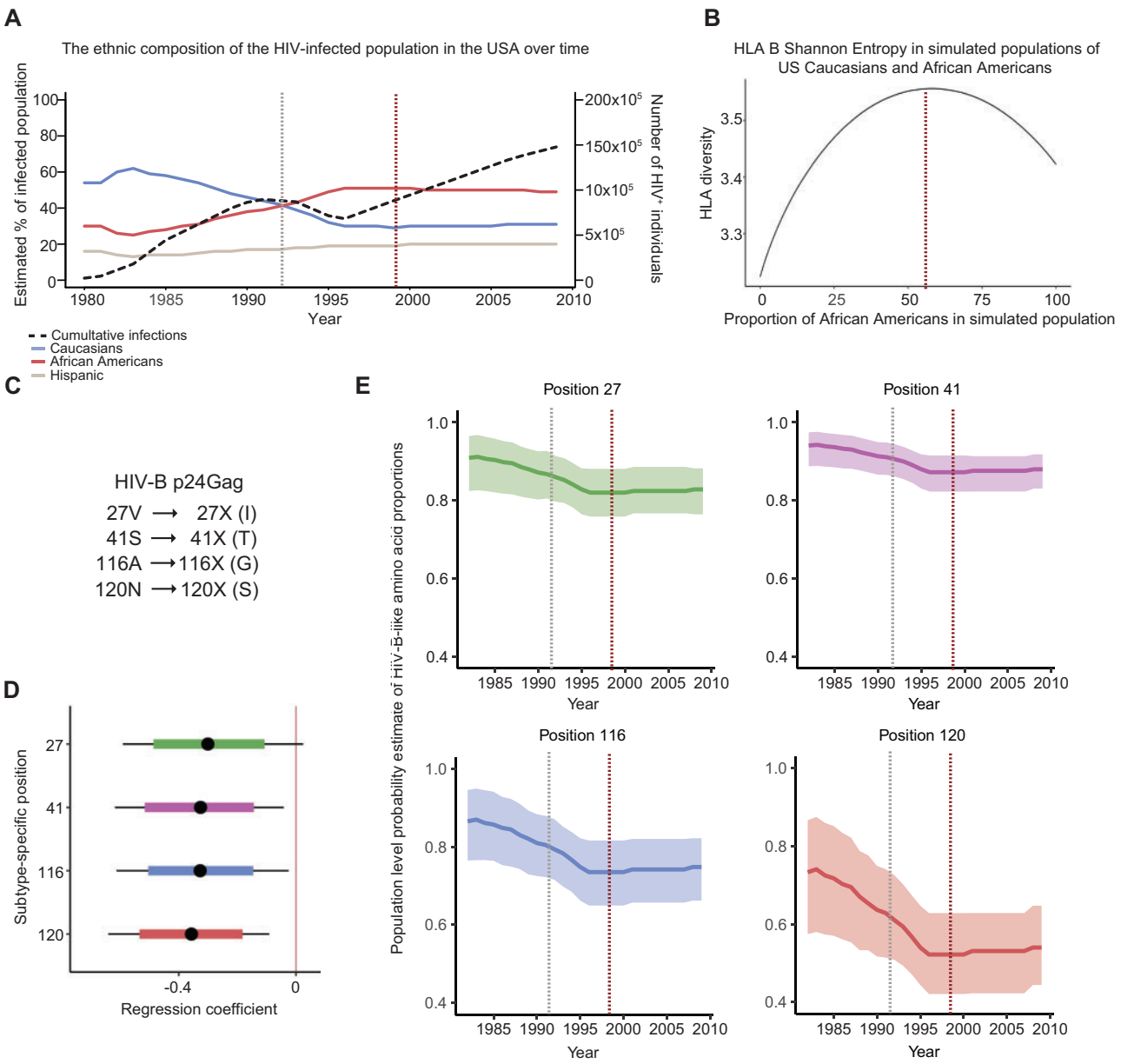



**Figure 5.** HIV-B adaptation to the proportion of African Americans in the US HIV-infected population over time. (A) The estimated ethnic composition of the HIV-infected population in the USA over time. The grey vertical line indicates the time of equal proportions of Caucasians and African Americans, and the dark red vertical line when the proportion of African Americans had stabilized. (B) Simulation of the relationship between the proportion of African Americans in a mixed African American and US Caucasian population and the HLA B Shannon entropy (HLA B diversity) of those mixed populations. Red dotted line highlights the maximum HLA diversity (all HLA simulations and HLA frequencies in Supplementary Fig. S10). (C) Overview of the changing subtype-specific positions; HIV-C subtype-specific amino acids are shown in parentheses and represent the most common change. (D) Posterior distributions over the regression coefficients of our model. The central point represents the median, the coloured region is the 80 per cent BCI, and the whiskers represent the 95 per cent BCI. The 95 per cent BCI exclude 0 for Positions 41, 116, and 120, whereas the 95 per cent BCI for Position 27 slightly overlap zero. (E) Population-level estimates of the probability a subtype-specific position is occupied by an HIV-B subtype-specific amino acid as explained by the fraction of African Americans plotted against time (with the 95% BCI).

diversification. This observation suggested that these population-specific HLA-associated selective pressures might also shape the evolution of other novel pathogens in emerging epidemics, for example, Zika virus and Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and might be partly responsible for the subtype diversification of viruses that have a longstanding infection history in humans (e.g. Hepatitis B, Hepatitis C, and Morbilli). Similarly, these population-specific selective pressures could affect the evolution of old pathogens when these spread into previously naïve populations in new geographical locations due to human activities, and/or climate change (e.g. dengue virus and West Nile virus). This granular adaptation of pathogens to a population-specific HLA composition mirrors the already accepted granular adaptation of human populations to different combinations of infectious challenges (Doherty and Zinkernagel 1975).

The adaptive introgression of Neanderthal and Denisovan HLA variants in Eurasians, and of HLA B*73:01 (a unique major histocompatibility complex (MHC)-BII allele more closely related to chimpanzee and gorilla MHC-B alleles than to other

modern human B alleles; Abi-Rached et al. 2011), and more recently of rainforest hunter-gatherers HLA variants in Bantu-speaking people (Patin et al. 2017) collectively suggest that introgression of pre-adapted, advantageous HLA variants might have occurred throughout hominid evolution and modern human history as invading populations mixed with adapted locals. Such adaptive introgression events are likely not limited to primate evolution. The repeated introgression of immune response-related genes would expose locally established pathogens to a continuous selective pressure for longer than the length of time that the most recently immigrated host lineage has been in a place.

In Africa, we found significant positive relationships between ethnic diversity and HIV-1 p24Gag sequence and subtype diversity, respectively. These robust associations suggested that ethnic diversity, and by extension HLA diversity and differences in HLA-associated selective pressures, helps drive HIV-1 evolution and might result in subtype diversification. It will be important to confirm and extend the findings reported here across other antigenic regions of HIV-1. The spread and adaptation of HIV-1 to different ethnic groups might initially have been facilitated by, for example, a combination of the socioeconomic and cultural changes induced by the colonization of the DRC in the early 1900s, for example, the rise of cities, extended transport networks, and centralized camps for treatment of sleeping sickness (Worobey et al. 2008; Faria et al. 2014).

Our results, and previous experimental work (Frahm et al. 2006; Tenzer et al. 2009, 2014), suggest that the local HIV-1 consensus sequence represent the most common adaptions to escape effective immune responses in the majority of people in the population in which it circulates. Consequently, the local HIV-1 consensus sequence is the least likely sequence combination to evoke effective immune responses upon transmission. This proposal is in line with reports that the virus closest to the consensus sequence in a population is most readily transmitted (Carlson et al. 2014; Song et al. 2011). Because the transmitted virus only needs to be present in small amounts in the infectious dose (Keele et al. 2008), these observations together might help the understanding of why the HIV-1 evolutionary rate at the within-host compared with the population scale is ~5-fold faster (Raghwani et al. 2018). The predominant viral variants in the host are typically those best adapted to escape that particular host's specific immune responses, while those viruses that are most likely to be transmitted are those most like the local consensus sequence. These 'consensus-like' sequences are likely to be found as a minority population and to be more closely related to the initially infecting viral variants, compatible with both the 'store and retrieve' and 'adapt and revert' scenarios suggested by Raghwani et al. (2018).

Between 1962 and 1970, HIV-1 spread beyond Africa as a Haitian aid-worker returned from the DRC (Gilbert et al. 2007). Whether this person was infected by HIV-B or a Subtype-B/D ancestor lineage is unknown, but the latter is likely given the lack of reports of indigenous HIV-B epidemics in Africa (van Harmelen et al. 1997; Hemelaar et al. 2011; Foley et al. 2018). The Haitian HIV-B epidemic is characterized by being more genetically diverse than elsewhere, and the predominant viral lineages are distinct from the pandemic HIV-B lineage (Gilbert et al. 2007). Despite ~70,000 American tourists visiting Haiti every year throughout the 1970s, and a rapidly developing hetero- and homosexual sex industry (Giraud 2010), it took ~10–12 years before HIV-B spread within the USA to a sufficient extent for the epidemic to be detected (Gilbert et al. 2007; Worobey et al. 2016). It is also noteworthy that no HIV-1 outbreak was observed in Belgium following the return of the colonialists when Congo achieved independence in 1960. The causes of this lack of spread or these interludes are debateable and likely multiple, partly dissimilar, and complex. However, we propose that one shared contributing factor could be lower transmissibility of African HIV-1 lineages to and between Europeans, perhaps because these HIV-1 lineages were not pre-adapted to escape immune responses associated with HLA variants common in Caucasians. This hypothesis is indirectly supported by reports of; (i) faster disease progression in HLA-matched recipients of pre-adapted HIV-1 (Carlson et al. 2016), (ii) less polyfunctional and more narrow CTL responses and higher viral loads in infected vaccine recipients with higher HLA I adaptation to the Gag vaccine insert (Boppana et al. 2019), and (iii) studies in Zambia demonstrating an association between specific, common HLA variants and faster HIV-1 acquisition (Song et al. 2011) and a more readily transmission of the predominant HIV-C consensus sequence (Carlson et al. 2014). In combination, these observations suggest that the increasing HIV-B adaptation to African-Americans, and the on-going high infection rates, could result in faster transmission and disease progression in this group.

Our results, and the many HIV-1 vaccine trial failures (Barouch and Korber 2010), suggest that the starting point for HIV-1 vaccine design should be the HLA composition of the target population, not the HIV-1 subtype infecting that population as in, for example, the unsuccessful Step and Phambili trials (Corey, McElrath, and Kublin 2009), because the circulating subtype will have adapted to avoid evoking strong immune responses in the majority of the population (Tenzer et al. 2009, 2014). One vaccine design option would be to select in-natural-infection subdominant epitope sequences from vulnerable parts of HIV-1 (Dahirel et al. 2011) that are presented by HLA variants common in the target population and embed these in artificial contexts that promote high epitope production. The resulting abundant epitope production will help prime anti-viral CTL responses (Faroudi et al. 2003), while even limited production of these epitopes after HIV-1 infection will allow the primed CTLs to recognize and kill infected cells (Purbhoo et al. 2004).

## 4. Materials and methods

### 4.1 Simulation of the relationship between ethnic diversity and HLA diversity

To investigate the relationship between ethnic diversity and HLA diversity we compiled HLA data from all available African populations/ethnic groups that had been HLA typed to a 'Gold standard' (i.e. all HLA types were reported as four-digit HLA variants) ($n = 8$; detailed in Supplementary Table S1). Although three of the seven African populations came from Kenya, they belonged to distinct ethnic groups and were, therefore, all included. The eight populations were mixed in arbitrary proportions to simulate the mixture of ethnic groups within different fictive African 'countries'. The proportions were assigned as follows: k random draws were taken from a uniform distribution. These were turned into proportions by dividing each random draw by the sum of all random draws. HLA types were mixed proportionally into each simulated 'country', and Shannon entropies were calculated both for the simulated ethnic diversity and the simulated HLA diversity. Linear regressions were performed of the Shannon entropy of the HLA diversity calculated from these mixes versus the population Shannon entropy calculated from the simulated mix. Statistically significant results

were obtained for linear regression analyses of simulated countries consisting of n 'ethnic groups' and for all simulated countries (not shown).

### 4.2 Shannon entropy analyses of African HIV-1 p24Gag diversity

To calculate the Shannon entropy (diversity) of HIV-1 group M p24Gag for each country we searched the LANL HIV database (Foley et al. 2018) for one non-problematic HIV-1 p24Gag sequence per patient from Sub-Saharan Africa (each sequence was $\geq$ 684 nucleic acids in length ($\sim$99% of the $HXB_2$ p24 sequence (GenBank K03455))) using the integer patient ID tool. Subsequently, we removed all sequences from HIV-1 Groups N, O, and P, and URFs and CRFs thereof, as well as sequences without a unique patient ID to avoid oversampling individual patients. We screened the remaining sequences for those originating from uncultured peripheral blood mononuclear cells, and those originated from cultured virus by examining all paper references. All sequences originating from cultured virus were discarded because culturing likely obscures the sequence footprints of the intra-host immune response. The remaining 2,274 sequences were aligned using the LANL HIV Align program, after which the sequences were stratified by country. Minor manual corrections were made to each country's alignment as needed, and the mean amino acid Shannon Entropy of each of these alignments was calculated using the R (version 3.4.3) package bio3d (version 2.3-3).

### 4.3 Shannon entropy analysis of ethnic diversity

As a proxy for each country's HLA diversity, we used 'ethnic diversity', which was based on the combined measure of ethnic, linguistic, cultural, and religious fractionalization defined by Alesina et al. (2003). This combined measurement is based on data from several sources, including Encyclopaedia Britannica, The CIA World Factbook, Minority Rights Group International, and (Levinson 1998). The ethnic diversity for each country was estimated by calculating the Shannon entropy for the ethnic proportions data in (Alesina et al. 2003) as follows:

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

where $n$ indicates the number of ethnicities recorded with the database, and $p_i$ is the proportion of ethnicity $i$ within the country's population.

### 4.4 Examination of ethnic versus HIV-1 p24Gag diversity

Ethnic diversity was plotted against the mean HIV-1 p24Gag amino acid entropy, weighted by the number of p24Gag sequences per country, using ggplot2 (version 2.2.1). The size of each dot was scaled logarithmically by the number of sequences to indicate graphically how this weighting was occurring. A weighted linear model was fit in R (version 3.4.3) to assess statistical significance.

### 4.5 African HIV-1 sequences and subtype analysis

We searched the LANL HIV database (Foley et al. 2018) for all non-problematic HIV-1 sequences originating in Africa (176,071 sequences). HIV-1 from Groups N, O, and P, and URFs and CRFs thereof were removed, and all HIV-1 group M pure subtypes, URFs and CRFs were filtered for one sequence per patient according to the integer patient ID (Foley et al. 2018). Sequences were then filtered for those originating in Sub-Saharan African (as defined by the United Nations Statistics Division, not LANL) yielding 36,044 sequences. From these, the number of subtypes, URFs, and CRFs per country and the number of sequences of each subtype, URF, and CRF per country was calculated. The total number of subtypes, URFs and CRFs per country was plotted against the ethnic diversity for each country weighted by the number of sequences per country using ggplot2 (version 2.2.1). The size of each point was scaled logarithmically by the number of sequences. A weighted generalized linear model was fit using R (version 3.4.3) to determine the significance. We generated a map with pie charts showing the relative amounts of each subtype in the LANL database for each country, using Tableau 10.4 (Fig. 2D; a non-weighted version is shown in Supplementary Fig. S3). The ethnic diversity for each country was also displayed on this map. We generated a map showing the African ethnolinguistic groups (Fig. 2E). For each African country we obtained the proportion of First Language Users of Bantu languages in the general population from Ethnologue (2020; www.ethnologue.com). For ease of analysis we dichotomized this variable using 50 per cent as a threshold value. Similarly, we divided the African countries into two groups by the prevalence of HIV-C in their HIV-infected population.

### 4.6 Worldwide HIV-infected patient information

To study potential HLA driven selection on HIV subtype-specific motifs, we searched the LANL HIV database (Foley et al. 2018) for all HLA typed HIV-1-infected patients from which HIV-1 p24Gag had been sequenced. For each patient, we obtained the majority rule consensus for five Gag subtype-specific positions (Positions 27, 41, 116, 120, and 128, Supplementary Fig. S1). The LANL HLA annotations were manually curated to ensure a consistent HLA nomenclature (e.g. 'B2705', 'B*27:05', and 'HLA-B*27:05' were all rewritten as 'B2705'). We used HLA annotations with four or more digits to generate an HLA presence–absence matrix to avoid pooling HLA variants with different effects on HIV evolution (e.g. HLA-B*35-Py/Px and HLA-B*58:01/02 (Goulder and Watkins 2008; Goulder and Walker 2012)). However, this matrix was extended by two-digit HLA annotations if the four-digit HLA type could be inferred based on (i) the patients HLA variants (typical for either Caucasians or African Americans), or (ii) the escape mutation pattern in their HIV-1 sequences (see Supplementary Table S8 for patient ID). For example, patients with HLA-B*27:05 often target the KK10 epitope, and the R264K, R264T, or L268M substitutions are often present within the epitope, but this is not the case in patients with HLA-B*27:02. In total, four-digit HLA types were inferred for 23 patients with HLA-B*27 (Goulder et al. 2001; Goulder and Watkins 2008; Goulder and Walker 2012). Likewise, 29 B*57 patients were reclassified as B*57:01 based on the associated paper reference (Migueles et al. 2003) and two patients were classified as B*57:01 carriers based on their other HLA types and cohort demographics (Goulder and Watkins 2008; Goulder and Walker 2012). We reassigned six American patients with HLA type B*35 to B*35:01 in African-American patients, as the frequency of HLA B*35:01 in this ethnic group is far higher than that of HLA B*35:02 or HLA B*35:03 (Gonzalez-Galarza et al. 2015). After assessing all the publicly available data, we limited ourselves to HIV-B ($n = 318$) and HIV-C ($n = 704$), the only subtypes for which there were enough patients to perform our analyses.

### 4.7 HLA presence–absence matrix

We generated a binary HLA presence–absence matrix; the rows represented patients ($n = 1{,}022$), and the columns the HLA Class I variants (the total number of HLA A, B, and C variants was 78). The entries were 1 when a patient was annotated with an HLA variant in the LANL database, and 0 otherwise resulting in an HLA profile for each patient (the table is available upon request).

### 4.8 Population HLA Class I information

The HLA frequencies of all populations were derived from the HLA-allele database (Gonzalez-Galarza et al. 2015) except for the South African Zulu population. These HLA frequencies were obtained as part of the Females Rising through Education, Support, and Health project and were shared by Professor Bruce Walker, Harvard University, and Professor Mary Carrington, Frederick National Laboratory for Cancer Research.

### 4.9 Exploratory PCA

To investigate the variance–covariance structure of the HLA presence–absence matrix, we performed a PCA (Menozzi, Piazza, and Cavalli-Sforza 1978) and created a scatterplot of the first two PC scores (Fig. 3 and Supplementary Fig. S4). Briefly, PCA uses an orthogonal transformation to convert the variables in the HLA presence–absence matrix into a smaller set of linear combinations called PCs and determines the minimum number of factors that will account for the maximum variance in the data. The first PC accounts for as much of the variability in the data as possible, and each succeeding PC accounts for as much of the remaining variability as possible.

### 4.10 Accounting for phylogenetic variance

In our statistical model, we explicitly accounted for the phylogenetic covariance between individuals infected by closely related viruses by including a parameter ($u_{ij}$) for every patient ($i$) and amino acid position ($j$) combination. Every column in this matrix (representing the phylogenetic effect for one subtype-specific position) was modelled by a multivariate normal distribution over patients using a phylogenetically informed covariance matrix, thereby representing the similarity between patients through the relatedness of their viruses. This phylogenetic correction was necessary because, for example, antenatal transmission could lead to closely related viruses (with very similar amino acids in the subtype-specific positions) in patients with similar HLA variants even in the absence of HLA-associated selection. We chose to use a separate phylogenetically informed multivariate normal for each subtype-specific position because using just a single multivariate normal (i.e. a single parameter per patient, shared over all $k$ subtype-specific positions, from a phylogenetically informed multivariate normal distribution) would fail to capture the independence of each subtype-specific position.

This model can be expressed as follows:

$$\text{logit}(p_{ij}) = \alpha_j + \boldsymbol{\beta}_{j,1:l_j} \cdot \mathbf{X}_{i,j,1:l_j} + u_{ij}$$

$$u_{*j} \sim \mathcal{N}_n(0, \sigma\mathbf{C})$$

$$y_{ij} \sim \text{Bernoulli}(p_{ij})$$

$n$ is the number of patients.$k$ is the number of amino acid positions.$l_j$ is the number of HLA types used as explanatory variables for subtype-specific position $j$.

$\mathbf{Y}$ is a $n \times k$ matrix (with individual elements $y_{ij}$) of binaries representing whether the majority rule amino acid in a patient is the same as the HIV-B Majority Rule consensus amino acid.

$\mathbf{X}$ is a $n \times k \times l_j$ array of binaries representing the selected HLA types for the $j$th subtype-specific position. This ragged 3D construction is necessary because we use different sets of HLA types as explanatory variables for each amino acid; hence a separate 2D HLA matrix is needed for each position.

$\boldsymbol{\beta}$ is a $k \times \max(l_j)$ matrix representing the HLA coefficients for the $k$ different amino acids. Vectors $\boldsymbol{\beta}_{j*}$ are not used in the model beyond index $l_j$; this is because ragged arrays are not available in Stan (the software we used to do Markov chain Monte Carlo (MCMC) analysis; Carpenter et al. 2017).

We define $\mathbf{P}$ as an $n \times k$ matrix containing the inferred probabilities $p_{ij}$ of amino acid $j$ in patient i being B-like.

$\mathbf{C}$ is a covariance matrix representing genetic relatedness between the viruses infecting the patients in our sample (see below).

The parameterization $\boldsymbol{\beta}_{j,1:l_j} \cdot \mathbf{X}_{i,j,1:l_j}$ was chosen because there is little overlap between the HLA types that selected for each amino acid position. The $l_j$ variable is necessary because different numbers of HLA types are appropriate for the different subtype-specific positions.

Gelman's recommendation (Gelman et al. 2008) for weakly informative priors appropriately scaled for logistic regression was adopted for the alpha and beta parameters; namely Cauchy distributions with a location parameter of 0 and a scale parameter of 2.5. The $\sigma$ parameter was given a half-normal prior with a 0 mean and a standard deviation of 1. Because this analysis is highly parameterized (one $u_{ij}$ per data point), allowing $\sigma$ to vary resulted in high sampling variance in $\sigma$ and the $u_{ij}$ parameters until the inverse logit function collapses to 0 or 1 for all data points, leading to instability of the $\sigma$ parameter and non-convergence due to non-identifiability. To prevent this, we set $\sigma = 1$. This choice of parameter is reasonable in the context of a logit link function, which rapidly approaches 0 and 1 outside the range $[-2.71, 2.71]$. This method is based on a previous Bayesian model which only allowed for linear regressions (de Villemereuil et al. 2012), which we here extend to allow the fitting of generalized linear models. We do not incorporate phylogenetic uncertainty, but instead use a point-estimate of the phylogenetic correlation matrix C.

We used simulations to justify our choice for an independent phylogenetically informed multivariate Gaussian prior for each amino acid position. We simulated each amino acid separately using a Bernoulli distribution where $p$ is a draw from the multivariate normal distribution using a phylogenetic variance–covariance matrix and the logistic link function; $x$ was a draw from a multivariate Gaussian distribution using the same phylogenetic variance–covariance matrix. We varied the number of amino acids simulated between 1 and 9 and generated 2,000 simulated datasets for each amino acid count. We fitted a model with a single phylogenetic multivariate Gaussian distributed random effect shared between all amino acids in a patient, and a model with a separate phylogenetic multivariate Gaussian distributed prior for each amino acid. We found that the model with the shared multivariate Gaussian was prone to increase false positive rates (FPRs) and that the FPR increased with increasing numbers of simulated amino acids. The FPR in the model with a separate phylogenetical multivariate normal distribution for each amino acid was stable. Stan code is available upon request.

### 4.11 Generating the phylogeny and the phylogenetic correlation matrix

For each subtype, we built manually curated alignments with a single representative sequence per patient and used the LANL HYPERMUT tool to remove hypermutated sequences. To avoid biasing the analysis by including HLA selective effects in the correlation matrix C, which could occur since CTL responses can drive convergent evolution leading to clustering in phylogenetic trees (Matthews et al. 2009), we removed from the alignment the five subtype-specific columns we use as response variables in the multilevel model, and columns subject to CTL escape mutation known to skew phylogenetic analysis. BEAST was used to sample from the posterior distributions of phylogenies by running 10 chains for 200 million steps with an HKY+G model of nucleotide evolution, a lognormal clock, and a logistic growth coalescent tree prior. The initial portions of the chains were discarded as burn-in (50 million for HIV-B; 100 million for HIV-C, except one HIV-C chain for which we discarded 150 million). We evaluated the posterior using CODA (Plummer et al. 2006), and ensured that all effective sample sizes (ESS) for continuous parameters were larger than 200. We used TreeAnnotator to build a Maximum Clade Credibility (MCC) tree for each subtype, keeping the heights of the maximum posterior tree to prevent visualization artefacts. For every tree sampled from the posterior, we calculated a phylogenetic variance–covariance matrix (Felsenstein 1985; Freckleton, Harvey, and Pagel 2002). By averaging over all sampled trees, we estimate the posterior phylogenetic variance–covariance matrix scaled in years, which was transformed into the phylogenetic correlation matrix C.

### 4.12 Cross-validation

To avoid overfitting the high number of HLA types relative to the number of patients (HIV-B: 26 HLAs/292 patients; HIV-C: 45 HLAs/687 patients) we performed a 5-fold cross-validation procedure (Kohavi 1995). This procedure incorporated both feature selection and model fitting, which allowed us to optimize the bias-variance trade-off. Within each training set, we calculated the correlation coefficients between each subtype-specific position and all HLA types; we then constructed Bayesian models with the $l$ best-correlating HLA types for every subtype-specific position and used the inferred parameters to predict the excluded validation sets. From these predictions, we calculated the area under the curve (AUC) of the receiver operator characteristic curves (ROC) values for each subtype-specific position (Fawcett 2006). We chose this metric because it is invariant to the calibration of the latent probabilities in the Bernoulli model. The training and validation sets were constructed on the patient level to avoid upwardly biasing the AUC (Saeb et al. 2016).

This procedure was repeated for every $l$ until all HLA variants were included in the model. After 10 replications, to reduce cross-validation error, a LOESS (LOcally Estimated Scatterplot Smoothing) curve of the results was calculated for each subtypes-specific site. In the cross-validation model fitting procedure on the training set, all $l_j$ were identical. For each subtype-specific position, the $l_j$ associated with the highest AUC was taken as the starting point for the parsimonious model. Cross-validation was performed on HIV-B and HIV-C separately for both the patient-parameter and the phylogenetic model. The numbers of HLA variants vary at each position in Fig. 5 as they reflect the point when adding additional HLA variants ceased to improve our predictions for that particular subtype-specific position.

### 4.13 Parsimonious model

For the parsimonious models, we took the $l_j$ HLA types for each position which had the highest correlations with the amino acids, calculating the correlations using the entire dataset, where $l_j$ was the optimum number of HLA types for a subtype-specific position $j$ according to the cross-validations. We then fitted a multilevel Bayesian model using these predictors and refitted it after removing any HLA types where the associated OR 80 per cent BCI failed to exclude 1. We obtained posterior distributions from the model using the Stan MCMC sampler (Carpenter et al. 2017). We ran four chains for 2,000 generations each, discarded the default 50 per cent burn-in and confirmed convergence by ESS $>$500 for all parameters, $\hat{R}<1.01$ and satisfactory Gelman diagnostics (Gelman and Rubin 1992; Brooks and Gelman 1998; Gelman et al. 2008). Pairs plots of the beta posteriors were examined to check for collinearity between the HLA types, which could arise if two or more included HLA types were part of a common haplotype. Pair plots of the MCMC traces were examined at every step to search for potential interactions.

### 4.14 An agent-based method to investigate HIV-1 subtype diversification

To create our model, we combined a transmission dynamics model with a model for within-host HIV-1 evolution (both described below). We examined the effect of HLA selection on HIV-1 evolution in two ways. First, we used the HLA ORs estimated from the HIV-B HLA regression dataset and evaluated how these would affect the modelled amino acid frequencies (Fig. 4G and Supplementary Fig. S9) and second, we tested how much the HLA ORs would have to be increased to result in the HIV-B amino acid frequencies observed today. To more adequately capture the dynamics of the early, primarily Caucasian, HIV-1 epidemic outside of Africa, we adjusted all HLA ORs from HLA variants common in AfricanAmericans to 1. Because only non-B like subtype-specific amino acids are present initially in our model, there would be no selective pressure against HIV-B-like amino acids as this selective pressure would originate from the CTL response against epitopes that are only processed, or processed to a greater degree, when an HIV-B-like amino acid is present in a subtype-specific position. To reflect this, we restricted all odds $<1$ (i.e. those which drive evolution away from subtype-B) to equal 1.

#### 4.14.1 Model part 1: transmission dynamics

We assumed a human population of a fixed size into which an initial infection of an HIV-C-like virus was introduced to a small subset of the population at time zero. The infection subsequently spreads stochastically through the population according to the following processes:

- Random mating between an infected person and one of the susceptible individuals (S) occurs at a rate $\beta$. This assumption implies there is no spatial structure to the community that is, it is panmictic.
- Death of a susceptible individual occurs at a rate $\mu_S$.
- Death of an infected individual occurs at a rate $\mu_I > \mu_S$.

To maintain a constant population size, whenever an individual died they were replaced with a susceptible individual. We assumed that all the individuals in the population were

susceptible to HIV-1 before the infection was introduced. After the introduction, the population was composed of a mix of susceptible and infected individuals. In this model, individuals could not recover from their HIV-1 infection, and so the infection dynamics followed that of a stochastic 'susceptible-infectious' model. To simulate the dynamics we used the continuous time Gillespie algorithm (see Erban, Chapman, and Maini 2017).

#### 4.14.2 Model part 2: within-host evolution of the HIV sequences

We recognize that in reality, an individual is usually infected with a limited number of HIV-1 variants that are not specifically adapted to the recipient. In our simulations, however, we suppose that within-host evolution of HIV-1 is sufficiently rapid to ensure that the HIV-1 population within a patient during most of the infectious period (i.e. after ~9 months; Karlsson et al. 2007; Goulder and Walker 2012; Abrahams et al. 2013) becomes dominated by HIV-1 that has adapted to that particular host.

We modelled the observed HIV-B and HIV-C subtype-specific sites in HIV-1 p24Gag and assumed that the specific amino acid present at each of these positions evolved in accordance with the HLA profile of the host to be either HIV-B- or HIV-C-like. When individuals in our model are 'born' they are endowed with an HLA profile, which is drawn from an underlying prespecified distribution for a collection of HLA types for which we have determined the selective ORs for selecting for an HIV-B-like amino acid if HIV-1 infected. When each individual is created, they can have a maximum of two (and a minimum of zero) alleles for each of the HLA A, B, and C with associated ORs. We recognize that in real life, an individual will possess two alleles for each of the three HLA classes. In our model, if an individual has, for example, zero HLA A variants associated with HLA ORs, then this means that their two HLA Class A alleles are not amongst those we model (and have an OR of one; see below).

For a given amino acid position $j$ we calculated the combined odds across all HLA classes belonging to the host by taking the product of the corresponding odds,

$$\overline{\eta}_j = \prod_{i=1}^{K} \eta_{ij}, \tag{4}$$

where $K \in (3, 4, 5, 6)$ is the number of HLA classes that we model, and $\eta_{ij}$ is the estimated odds for HLA class $i$ and amino acid position $j$. We used an independent continuous time Markov Chain to model the evolution of each amino acid position. Specifically, we assume a transmission matrix of the form,

$$\Gamma = \begin{bmatrix} (1-\gamma)\overline{\eta}_j & \gamma \\ \overline{\eta}_j\gamma & (1-\gamma) \end{bmatrix},$$

where $\gamma$ is the mutation rate of an amino acid. The matrix is orientated so that the diagonal entries determine the rate at which an amino acid position remains at its current state (top-left corresponding to the rate of remaining in an HIV-B state, bottom-right to remaining in an HIV-C state), and the off-diagonal elements correspond to the rate of mutation. In the matrix, the left-hand column is multiplied by the estimated odds which result in neutrality if $\overline{\eta}_j = 1$, a bias towards an HIV-B if $\overline{\eta}_j > 1$, or a bias towards HIV-C if $\overline{\eta}_j < 1$. If one individual infects a susceptible (multiple HIV inoculations within one individual are not allowed in our model) then they are given an HIV variant with a sequence that corresponds to the current sequence (i.e. the result of the evolving Markov Chain after the time since infection) within the infected person.

#### 4.14.3 Model sequence

In the simulations, we first created a population of $N$ individuals, and then randomly infect an initial number $I_0$ of the populace with HIV-1 with one particular sequence (in our simulations variants with all amino acids in the HIV-C positions). We then iterate the following until a given time $T$ has elapsed,

1. Determine the time step $t$ until the next 'event' and the type of event (an infection, a death of a susceptible, or a death of an infected) to occur by the Gillespie algorithm.
2. Implement the event.
3. For each infected individual in the population evolve each HIV amino acid position as a continuous time Markov Chain for time $t$ with rates determined by their HLA type.

We used the parameters shown in Table 1 in the simulations. The transmission rate at the start of the epidemic was obtained from Varghese et al. (2002).

### 4.15 HIV-infected US patient sequences and HLA information

To assess HIV-1 evolution over time in the USA, we searched the LANL HIV sequence database (Foley et al. 2018) for all dated HIV-B sequences that spanned HXB2 nucleotide coordinates 1,252–1,569 (a fragment of p24Gag) and were sampled in the USA and annotated with a patient ID ($n = 9{,}729$ HIV-B sequences from 437 patients). This limitation was necessary to avoid pseudo-replication through the inclusion of multiple sequences from the same patient as unique data points. In contrast to the compilation of the worldwide HIV-infected patient sequences, we also included patients without HLA information and allowed for multiple sequences to be used per patient rather than taking the majority rule amino acid.

### 4.16 Building a posterior distribution of phylogenies from US sequences

We ran 10 BEAST chains for 100 million steps using a single representative sequence per patient and a GTR+G model of nucleotide evolution, using a log-normal clock and a logistic coalescent; this has previously been shown to be appropriate for similar data (Worobey et al. 2016). To promote model convergence, we fully incorporated ambiguous sites in this model and did not treat them as missing. We removed an appropriate 20 per cent burn-in and confirmed ESS > 200 on all continuous parameters in CODA (Plummer et al. 2006). We used TreeAnnotator to calculate a MCC summary tree.

**Table 1.** The parameters used in the agent-based agent-based epidemic model

| Parameter | Interpretation | Value |
|---|---|---|
| $n$ | Population size | 1,000 |
| $I_0$ | Initial infected group size | 500 |
| $T$ | Total time | 230 years |
| $\beta$ | Transmission rate | 0.0002 per infected/susceptible pair |
| $\mu_S$ | Death rate of susceptibles | 0.02 per year |
| $\mu_I$ | Death rate of infected | 0.1 per year |
| $\gamma$ | Amino acid mutation rate | 0.06 per year |

### 4.17 Bayesian model comparison with BEAST from US HIV-1 sequences

We also used the posterior distribution of trees to perform some exploratory analyses. It was not possible to use BEAST to analyse multiple sequences sampled from a patient. Instead, we limited our analysis to whichever sequence was most recently sampled from the patient, to allow the virus the most time to adapt to its host. We treated the amino acids as binary traits, which can be either positive if the amino acid in the patient's subtype-specific position is the same as the HIV-B consensus, or negative if it is different. We fitted four models of Discrete Trait evolution, varying the parameter configuration. Because we estimated that most transitions would occur from the HIV-B consensus amino acids to non-HIV-B consensus amino acids, we investigated if a unidirectional model, which did not allow the reverse transition, would fit better than a bi-directional model. Because we believe that these four subtype-specific positions are undergoing roughly the same process, we specified a hierarchical prior on their transition rates. Implementing the hierarchical prior was trivial in the unidirectional case, as there we could set the B→non-B transition rates to 1, and the non-B→B transition rates to 0, and put a hierarchical prior on the clock rates. The bi-directional models were parameterized by setting the B→non-B transitions to 1 and allowing the non-B→B transitions to vary as a relative rate; this model was made hierarchical by allowing the clocks to vary freely and independently, while hierarchical priors were placed on the relative rate for each subtype-specific site. We used the stepping stone sampling procedure implemented in BEAST to obtain estimates of the marginal likelihood for each of the four models, which were converted to Bayes Factors for model comparison by taking the difference between the marginal likelihoods of two models on the log scale (Baele et al. 2012, 2013). XML files are available upon request.

### 4.18 Calculating the fraction of African Americans in the US HIV-infected population over time

To estimate the ethnic demographics of the HIV-infected population in the USA, we converted existing estimates of the ethnic demographics of HIV incidence (new infections) to prevalence (number of living hosts infected). We obtained incidence data from the literature (Hall et al. 2008; Prejean et al. 2011) and used the following assumptions:

1. In the beginning of the epidemic (when no drugs were available) the average survival time was 10 years.
2. Starting from 1996 the availability of combination therapy increased the survival time to at least 25 years.
3. Starting from 2002 further improvement in the standard of care extended this to at least 32 years.

These assumptions were the same for all ethnic groups.

We then calculated the fraction of African Americans in the US HIV-1 infected population at each time point.

### 4.19 Regression of the US HIV-Infected population's African American frequencies on time

To smooth the data and to incorporate uncertainty in the estimates, we performed a Bayesian regression of time on the estimated African American proportions of HIV-1 infected Americans (Supplementary Fig. S11). Initial observations revealed a change (or a change-point) between growth and decay of the proportion of African Americans around 1998 and we used a Bayesian change-point analysis to model this. To prevent discontinuity at the change-point, it was required that the function was continuous with a continuous first derivative, and we, therefore, chose the following form:

$$f(x) = \alpha_0 + \alpha_1(x-p) + \alpha_2(x-p)^2 + \alpha_3(x-p)^3 + \alpha_4(x-p)^4$$

$$g(x) = c + de^{\gamma(x-p)}$$

$$y = \begin{cases} f(x) & \text{if } x \leq p \\ g(x) & \text{otherwise.} \end{cases}$$

where:

$$\alpha_0 = c + d$$

$$\alpha_1 = d\gamma$$

We used normal priors of mean 0 and standard deviation 10 on all parameters, except the $p$ change-point parameter (normal with mean corresponding to the year 1998 and SD 1), because the approximate position of the change-point is clearly between 1996 and 2000, and the likelihood function precision parameter (gamma with shape = 0.001 and rate = 0.001). To stabilize the variance and thereby ensure homoscedasticity, a regression model was fitted, which was weighted by the total number of patients alive according to the simple population model; this had the further advantage of allowing some uncertainty around the earlier time points. The calculated frequency of African Americans in the US HIV-infected population was standardized, and 1980 was taken as Year 0. Because the model uses latent binary variables to distinguish between the $f$ and $g$ functions we used the JAGS Gibbs sampler (Plummer 2003). Because of strong correlations between the parameters (due to collinearity of the $(x–p)$ polynomials), it was necessary to run JAGS for 100,000 steps. The median fitted values from this regression were used as the explanatory variable in the subsequent Bayesian Multilevel Model.

### 4.20 Bayesian multi-level model incorporating phylogenetic information

We included a phylogenetic signal in much the same way as in the previous analysis, by fitting parameters for every position/patient combination. These parameters were sampled from a single multivariate normal distribution for every subtype-specific position, using the phylogenetic correlation matrix to relate patients according to the evolutionary history of the virus. This can be expressed as follows:

$$y_{ij} \sim \text{Bernoulli}(p_{ij})$$

$$\text{logit}(p_{ij}) = \alpha_j + \beta_j x_i + u_{z_i j}$$

The (hyper-)priors are:

$$\alpha_j \sim \mathcal{N}(0, 2)$$

$$\beta_j \sim \mathcal{N}(\lambda, \sigma)$$

$$\lambda \sim \mathcal{N}(0, 2)$$

$$\sigma \sim \text{Half} - \text{Normal}(0, 1)$$

$$u_{*j} \sim \mathcal{N}_m(0, \mathbf{C})$$

This creates a multi-level model where each data point informs both the estimated probability of HIV-B-likeness for a subtype-specific position in an individual patient (through the $u_{ij}$ parameters) and the population level estimates (through $\beta_j x_i$). Additionally, we used a hierarchical prior on the slopes to relate the four subtype-specific sites to one another, creating what would be in classical statistical terms a random slope model, which allows the partial pooling of information to further protect against phylogenetic covariance as well as giving us additional power. This construction is justified by our prior belief that these four positions are special in the same way, and by our explorations of the data set when testing the phylogenetic model, which favoured a hierarchical prior on the transition rates. A hierarchical prior on the transition rates is the phylogenetic analogue to the hierarchical prior on the regression coefficients implemented here. More generally, a hierarchical prior can also be seen as a conservative choice due to the protection it confers against multiple hypothesis testing (Gelman 2009). As in the previous worldwide analyses, we used the correlation matrix as a variance–covariance matrix to prevent non-identifiability, implicitly setting the $\sigma$ in $\sigma C$ to 1 (not shown in Equation (4) to avoid confusion with the $\beta$ variance $\sigma$). The model was fitted using Stan and ran for 5,000 steps using a non-centred parameterization and the MCMC chains converged on the posterior distributions (all $\hat{R} \leq 1.01$). To validate the generative properties of the model, we performed a posterior predictive test, which confirmed that the actual data could have been generated by our model. Due to the binary nature of our data, we took the proportion of HIV-B consensus amino acids in each position for every patient in each year and calculated the yearly average for each subtype-specific position separately.

### 4.21 Population-level estimates of HIV-B consensus amino acid frequencies

Using our multilevel model, we reconstructed the population-level proportion estimates of the HIV-B consensus amino acid over time. We set x' as the inferred African American frequency in every year between 1982 and 2009 and used the following formula, which only included the population level parameters in the multilevel model and left out the patient-specific adjustments:

$$\text{logit}(q_{ij}) = \alpha_j + \beta_j x'_i \quad (6)$$

By plotting these population-level estimates ($q_{ij}$) against time, we recovered the underlying pattern from the data, which was the population level estimate that can be visualized together with the uncertainty inherent in the estimate (Fig. 5).

## 5. Data sets and code availability

The data were obtained from publicly available databases (the LANL HIV sequence database (Foley et al. 2018) and the Allele Frequency Net database (Gonzalez-Galarza et al. 2015)), and the South African Zulu HLA data were kindly shared by Professor Bruce Walker, Harvard University, and Professor Mary Carrington, Frederick National Laboratory for Cancer Research. All new code and algorithms are available upon request .

## Supplementary data

Supplementary data are available at *Virus Evolution* online.

## Author contributions

A. K. N. I. conceived and designed the overall study. N. C. K. and P. L. conducted the phylodynamic analyses. N. C. K., D. L., B. L, S. C., and A. K. N. I. performed data analyses and modelling and A. K. and P. L. contributed intellectual input. A. K. N. I. and N. C. K. wrote the paper; all authors commented on and approved the final article.

## Acknowledgements

We are deeply indebted to our co-author, mentor, and dear friend Dr Daniel Lunn, who sadly passed away before the acceptance of this publication. We thank the patients, the researchers whose publicly available data made this work possible, Mark Pagel, and Paul Klenerman for helpful discussions, and Bruce Walker and Mary Carrington for HLA data.

## Funding

Funding was obtained from the Medical Research Fund, Oxford University (MRF 1090) (A.K.N.I.), the EPSRC (1365928) and the Hendrik Muller Fund (N.C.K), the Special Research Fund, KU Leuven (OT/14/115) (P.L.), and the Research Foundation – Flanders ('Fonds voor Wetenschappelijk Onderzoek – Vlaanderen', G066215N, G0D5117N, and G0B9317N; P.L.). N.C.K was the holder of a Scatcherd Scholarship.

**Conflict of interest:** None declared.

## References

Abi-Rached, L. et al. (2011) 'The Shaping of Modern Human Immune Systems by Multiregional Admixture with Archaic Humans', *Science*, 334: 89–94.

Abrahams, M. R. et al. (2013) 'Rapid, Complex Adaptation of Transmitted HIV-1 Full-Length Genomes in Subtype C-Infected Individuals with Differing Disease Progression', *Aids (London, England)*, 27: 507–18.

Alesina, A. et al. (2003) 'Fractionalization', *Journal of Economic Growth*, 8: 155–94.

Amogne, W. et al. (2016) 'Phylogenetic Analysis of Ethiopian HIV-1 Subtype C near Full-Length Genomes Reveals High Intrasubtype Diversity and a Strong Geographical Cluster', *AIDS Research and Human Retroviruses*, 32: 471–4.

Baele, G. et al. (2012) 'Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty', *Molecular Biology and Evolution*, 29: 2157–67.

—— et al. (2013) 'Accurate Model Selection of Relaxed Molecular Clocks in Bayesian Phylogenetics', *Molecular Biology and Evolution*, 30: 239–43.

Barouch, D. H., and Korber, B. (2010) 'HIV-1 Vaccine Development after STEP', *Annual Review of Medicine*, 61: 153–67.

Benn-Torres, J. et al. (2008) 'Admixture and Population Stratification in African Caribbean Populations', *Annals of Human Genetics*, 72: 90–8.

Bhattacharya, T. et al. (2007) 'Founder Effects in the Assessment of HIV Polymorphisms and HLA Allele Associations', *Science*, 315: 1583–6.

Boppana, S. et al. (2019) 'HLA-I Associated Adaptation Dampens CD8 T-Cell Responses in HIV Ad5-Vectored Vaccine Recipients', *The Journal of Infectious Diseases*, 220: 1620–8.

Brooks, S. P., and Gelman, A. (1998) 'General Method for Monitoring Convergence of Iterative Simulations', *Journal of Computational and Graphical Statistics*, 7: 434–55.

Busby, G. B. et al. (2017), 'Inferring adaptive gene-flow in recent African history', *bioRxiv*. doi: 10.1101/205252.

—— et al.; Malaria Genomic Epidemiology Network. (2016) 'Admixture into and within Sub-Saharan Africa', *eLife*, 5:

Cao, K. et al. (2004) 'Differentiation between African Populations is Evidenced by the Diversity of Alleles and Haplotypes of HLA Class I Loci', *Tissue Antigens*, 63: 293–325.

Carlson, J. M. et al. (2008) 'Phylogenetic Dependency Networks: Inferring Patterns of CTL Escape and Codon Covariation in HIV-1 Gag', *PLoS Computational Biology*, 4: e1000225.

—— et al. (2014) 'HIV Transmission. Selection Bias at the Heterosexual HIV-1 Transmission Bottleneck', *Science*, 345: 1254031.

—— et al. (2015), 'HIV-1 Adaptation to HLA: A Window into Virus-Host Immune Interactions', *Trends in Microbiology*, 23: 212–24.

—— et al. (2016) 'Impact of Pre-Adapted HIV Transmission', *Nature Medicine*, 22: 606–13.

Carpenter, B. et al. (2017) 'Stan: A Probabilistic Programming Language', *Journal of Statistical Software*, 76: 1–32.

Centers for Disease Control and Prevention. <www.cdc.gov> accessed 24 October 2020.

Chang, C. H. et al. (2015) 'HIV-Infected Sex Workers with Beneficial HLA-Variants Are Potential Hubs for Selection of HIV-1 Recombinants That May Affect Disease Progression', *Scientific Reports*, 5: 11253.

Chen, L. et al. (2020) 'Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals', *Cell*, 180: 677–87.e16.

Corey, L., McElrath, M. J., and Kublin, J. G. (2009) 'Post-Step Modifications for Research on HIV Vaccines', *AIDS*, 23: 3–8.

Cotton, L. A. et al. (2014) 'Genotypic and Functional Impact of HIV-1 Adaptation to Its Host Population during the North American Epidemic', *PLoS Genetics*, 10: e1004295.

Crawford, H. et al. (2011) 'The Hypervariable HIV-1 Capsid Protein Residues Comprise HLA-Driven CD8+ T-Cell Escape Mutations and Covarying HLA-Independent Polymorphisms', *Journal of Virology*, 85: 1384–90.

Dahirel, V. et al. (2011) 'Coordinate Linkage of HIV Evolution Reveals Regions of Immunological Vulnerability', *Proceedings of the National Academy of Sciences of the United States of America*, 108: 11530–5.

de Villemereuil, P. et al. (2012) 'Bayesian Models for Comparative Analysis Integrating Phylogenetic Uncertainty', *BMC Evolutionary Biology*, 12: 102.

Doherty, P. C., and Zinkernagel, R. M. (1975) 'Enhanced Immunological Surveillance in Mice Heterozygous at the H-2 Gene Complex', *Nature*, 256: 50–2.

Erban, R., Chapman, J. M., and Maini, P. (2017), 'A Practical Guide to Stochastic Simulations of Reaction-Diffuson Processes', *https://arxiv.org/abs/0704.1908*.

Ethnologue. <https://www.ethnologue.com/> accessed 24 October 2020.

Excoffier, L. et al. (1987) 'Genetics and the History of Sub-Saharan Africa', *American Journal of Physical Anthropology*, 30: 151–94.

—— et al. (1991) 'Spatial Differentiation of RH and GM Haplotype Frequencies in Sub-Saharan Africa and Its Relation to Linguistic Affinities', *Human Biology*, 63: 273–307.

Faria, N. R. et al. (2014) 'HIV Epidemiology. The Early Spread and Epidemic Ignition of HIV-1 in Human Populations', *Science*, 346: 56–61.

Faroudi, M. et al. (2003) 'Lytic versus Stimulatory Synapse in Cytotoxic T Lymphocyte/Target Cell Interaction: Manifestation of a Dual Activation Threshold', *Proceedings of the National Academy of Sciences of the United States of America*, 100: 14145–50.

Fawcett, T. (2006) 'An Introduction to ROC Analysis', *Pattern Recognition Letters*, 27: 861–74.

Felix, M., and Meur, C. (2001) 'People's Atlas of Africa', <https://worldmap.harvard.edu/maps/africa-language-cs> accessed 25 October 2020.

Felsenstein, J. (1985) 'Phylogenies and the Comparative Method', *The American Naturalist*, 125: 1–15.

Foley, B. et al. (2018), *HIV Sequence Compendium 2018 (Theoretical Biology and Biophysics)*, Los Alamos, NM: Los Alamos National Laboratory.

Frahm, N. et al. (2006) 'Control of Human Immunodeficiency Virus Replication by Cytotoxic T Lymphocytes Targeting Subdominant Epitopes', *Nature Immunology*, 7: 173–8.

Freckleton, R. P., Harvey, P. H., and Pagel, M. (2002) 'Phylogenetic Analysis and Comparative Data: A Test and Review of Evidence', *The American Naturalist*, 160: 712–26.

Fryer, H. R. et al. (2012) 'Cytotoxic T-Lymphocyte Escape Mutations Identified by HLA Association Favor Those Which Escape and Revert Rapidly', *Journal of Virology*, 86: 8568–80.

Gelman, A. (2009), 'Why We (Usually) Don't Have to Worry About Multiple Comparisons', *https://arxiv.org/abs/0907.2478*.

——, and Rubin, D. B. (1992) 'Inference from Iterative Simulation Using Multiple Sequences', *Statistical Science*, 7: 457–72.

—— et al. (2008) 'A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models', *The Annals of Applied Statistics*, 2: 1360–83.

Gilbert, M. T. et al. (2007) 'The Emergence of HIV/AIDS in the Americas and Beyond', *Proceedings of the National Academy of Sciences of the United States of America*, 104: 18566–70.

Giraud, P. (2010), *The Tumultous History - From Pearl of the Caribbean to Broken Nation*. New York: Palgrave Macmillan.

Gonzalez-Galarza, F. F. et al. (2015) 'Allele Frequency Net 2015 Update: New Features for HLA Epitopes, KIR and Disease and HLA Adverse Drug Reaction Associations', *Nucleic Acids Res*, 43: D784–8.

Goulder, P. J., and Walker, B. D. (2012) 'HIV and HLA Class I: An Evolving Relationship', *Immunity*, 37: 426–40.

——, and Watkins, D. I. (2008) 'Impact of MHC Class I Diversity on Immune Control of Immunodeficiency Virus Replication', *Nature Reviews Immunology*, 8: 619–30.

—— et al. (2001) 'Evolution and Transmission of Stable CTL Escape Mutations in HIV Infection', *Nature*, 412: 334–8.

Gragert, L. et al. (2013) 'Six-Locus High Resolution HLA Haplotype Frequencies Derived from Mixed-Resolution DNA Typing for the Entire US Donor Registry', *Human Immunology*, 74: 1313–20.

Hall, H. I. (2008) 'Estimation of HIV Incidence in the United States', *JAMA*, 300: 520–9.

Hemelaar, J. et al.; WHO-UNAIDS Network for HIV Isolation and Characterisation. (2011) 'Global Trends in Molecular Epidemiology of HIV-1 during 2000-2007', *AIDS (London, England)*, 25: 679–89.

Hershkovitz, I. et al. (2018) 'The Earliest Modern Humans outside Africa', *Science*, 359: 456–59.

HIV Molecular Immunology Compendium, eds K. Yusim. et al. (2016) <http://www.hiv.lanl.gov/content/immunology> accessed 24 October 2020.

Hsieh, P. et al. (2016) 'Model-Based Analyses of Whole-Genome Data Reveal a Complex Evolutionary History Involving Archaic Introgression in Central African Pygmies', *Genome Research*, 26: 291–300.

Jacobs, G. S. et al. (2019) 'Multiple Deeply Divergent Denisovan Ancestries in Papuans', *Cell*, 177: 1010–21 e32.

Karlsson, A. C. et al. (2007) 'Sequential Broadening of CTL Responses in Early HIV-1 Infection is Associated with Viral Escape', *PLoS One*, 2: e225.

Kawashima, Y. et al. (2009) 'Adaptation of HIV-1 to Human Leukocyte Antigen Class I', *Nature*, 458: 641–5.

Keele, B. F. et al. (2008) 'Identification and Characterization of Transmitted and Early Founder Virus Envelopes in Primary HIV-1 Infection', *Proceedings of the National Academy of Sciences*, 105: 7552–7.

Kiepiela, P. et al. (2007) 'CD8+ T-Cell Responses to Different HIV Proteins Have Discordant Associations with Viral Load', *Nature Medicine*, 13: 46–53.

Kinloch, N. N. et al. (2016) 'Population-Level Immune-Mediated Adaptation in HIV-1 Polymerase during the North American Epidemic', *Journal of Virology*, 90: 1244–58.

Kohavi, R. (1995), 'A Study of Cross-Validation and Bootstrap for Accuracy Estimationand Model Selection', in C., Mellish (ed.), *International Joint Conference on Artificial Intelligence*, Vol. 14. USA: Morgan Kaufmann Publishers Inc.

Leslie, A. J. et al. (2004) 'HIV Evolution: CTL Escape Mutation and Reversion after Transmission', *Nature Medicine*, 10: 282–9.

Leslie, A. et al. (2005) 'Transmission and Accumulation of CTL Escape Variants Drive Negative Associations between HIV Polymorphisms and HLA', *Journal of Experimental Medicine*, 201: 891–902.

Levinson, D. (1998), *Ethnic Groups Worldwide, a Ready Reference Handbook*. Phoenix: Oryx Press.

Matthews, P. C. et al. (2008) 'Central Role of Reverting Mutations in HLA Associations with Human Immunodeficiency Virus Set Point', *Journal of Virology*, 82: 8548–59.

—— et al. (2009) 'HLA Footprints on Human Immunodeficiency Virus Type 1 Are Associated with Interclade Polymorphisms and Intraclade Phylogenetic Clustering', *Journal of Virology*, 83: 4605–15.

Menozzi, P., Piazza, A., and Cavalli-Sforza, L. (1978) 'Synthetic Maps of Human Gene Frequencies in Europeans', *Science*, 201: 786–92.

Meyer, D. et al. (2018) 'A Genomic Perspective on HLA Evolution', *Immunogenetics*, 70: 5–27.

Migueles, S. A. et al. (2003) 'The Differential Ability of HLA B5701+ Long-Term Nonprogressors and Progressors to Restrict Human Immunodeficiency Virus Replication is Not Caused by Loss of Recognition of Autologous Viral Gag Sequences', *Journal of Virology*, 77: 6889–98.

Moore, C. B. (2002) 'Evidence of HIV-1 Adaptation to HLA-Restricted Immune Responses at a Population Level', *Science*, 296: 1439–43.

Nielsen, R. et al. (2017) 'Tracing the Peopling of the World through Genomics', *Nature*, 541: 302–10.

Pagani, L. et al. (2012) 'Ethiopian Genetic Diversity Reveals Linguistic Stratification and Complex Influences on the Ethiopian Gene Pool', *The American Journal of Human Genetics*, 91: 83–96.

Patin, E. et al. (2017) 'Dispersals and Genetic Adaptation of Bantu-Speaking Populations in Africa and North America', *Science*, 356: 543–46.

Paximadis, M. et al. (2012) 'Human Leukocyte Antigen Class I (A, B, C) and II (DRB1) Diversity in the Black and Caucasian South African Population', *Human Immunology*, 73: 80–92.

Peeters, M., Jung, M., and Ayouba, A. (2013) 'The Origin and Molecular Epidemiology of HIV', *Expert Review of anti-Infective Therapy*, 11: 885–96.

Plummer, M. (2003), 'JAGS; A program for analysis of Bayesian graphical models using Gibbs sampling', *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003*, Vienna, Austria.

—— et al. (2006) 'CODA: Convergence Diagnosis and Output Analysis for MCMC', *R News*, 6: 7–11.

Prejean, J. et al.; for the HIV Incidence Surveillance Group. (2011) 'Estimated HIV Incidence in the United States, 2006-2009', *PLoS One*, 6: e17502.

Purbhoo, M. A. et al. (2004) 'T Cell Killing Does Not Require the Formation of a Stable Mature Immunological Synapse', *Nature Immunology*, 5: 524–30.

Pybus, O. G., and Rambaut, A. (2009) 'Evolutionary Analysis of the Dynamics of Viral Infectious Disease', *Nature Reviews Genetics*, 10: 540–50.

Quintana-Murci, L. (2019) 'Human Immunology through the Lens of Evolutionary Genetics', *Cell*, 177: 184–99.

Raghwani, J. et al. (2018) 'Evolution of HIV-1 within Untreated Individuals and at the Population Scale in Uganda', *PLoS Pathogens*, 14: e1007167.

Rambaut, A. et al. (2001) 'Human Immunodeficiency Virus. Phylogeny and the Origin of HIV-1', *Nature*, 410: 1047–8.

—— et al. (2004) 'The Causes and Consequences of HIV Evolution', *Nature Reviews Genetics*, 5: 52–61.

Reich, D. (2018) *Who We Are and How we Got Here* United Kingdom: Oxford University Press, Oxford.

Saeb, S. et al. (2016) 'Vodoo Machine Learning for Clinical Predictions', https://www.biorxiv.org/content/10.1101/059774v1.

Segurel, L., and Quintana-Murci, L. (2014) 'Preserving Immune Diversity through Ancient Inheritance and Admixture', *Current Opinion in Immunology*, 30: 79–84.

Simmons, N. et al. (2009), 'HIV/AIDS in the Horn of Africa', in *Public Health Aspects of HIV/AIDS in Low and Middle Income Countries*, pp. 331–64. Springer-Verlag, New York, NY, USA.

Song, W. et al. (2011) 'Disparate Associations of HLA Class I Markers with HIV-1 Acquisition and Control of Viremia in an African Population', *PLoS One*, 6: e23469.

Tenzer, S. et al. (2009) 'Antigen Processing Influences HIV-Specific Cytotoxic T Lymphocyte Immunodominance', *Nature Immunology*, 10: 636–46.

—— et al. (2014) 'HIV-1 Adaptation to Antigen Processing Results in Population-Level Immune Evasion and Affects Subtype Diversification', *Cell Reports*, 7: 448–63.

Tishkoff, S. A. et al. (2009) 'The Genetic Structure and History of Africans and African Americans', *Science*, 324: 1035–44.

van Harmelen, J. et al. (1997) 'An Association between HIV-1 Subtypes and Mode of Transmission in Cape Town, South Africa', *AIDS (London, England)*, 11: 81–7.

Varghese, B. et al. (2002) 'Reducing the Risk of Sexual HIV Transmission: Quantifying the per-Act Risk for HIV on the Basis of Choice of Partner, Sex Act, and Condom Use', *Sexually Transmitted Diseases*, 29: 38–43.

Worobey, M. et al. (2008) 'Direct Evidence of Extensive Diversity of HIV-1 in Kinshasa by 1960', *Nature*, 455: 661–4.

—— et al. (2016) '1970s and "Patient 0" HIV-1 Genomes Illuminate Early HIV/AIDS History in North America', *Nature*, 539: 98–101.

**Supplementary data:**

# HIV-1 p24Gag adaptation to modern and archaic HLA-allele frequency differences in ethnic groups contributes to viral subtype diversification

Nicolaas C. Kist, Ben Lambert, Samuel Campbell, Aris Katzourakis, Daniel Lunn, Philippe Lemey, Astrid K.N. Iversen

**Overview:**

**1) Supporting figures 1 to 12:**



**2) Supporting tables 1 to 8:**

**Table S1:** African populations with HLA A, B, and C data

**Table S2:** The number of p24 Gag sequences per country and the number of HIV-1 subtypes, CRFs and URFs per country and the country code abbreviation key

**Table S3:** Complete key to the African ethnicities in **Fig. 3C**

**Table S4:** Complete key over HIV-1 subtypes shown in **Fig. 3D**

**Table S5:** Complete key over African languages shown in **Fig. 3E**

**Table S6:** HLA class I allele frequencies in worldwide populations

**Table S7:** Overview of the amino acid variation at SSP in HIV-1 subtype consensus sequences

**Table S8:** LANL patient IDs of patients with imputed four digit HLAs

**Figure S1**

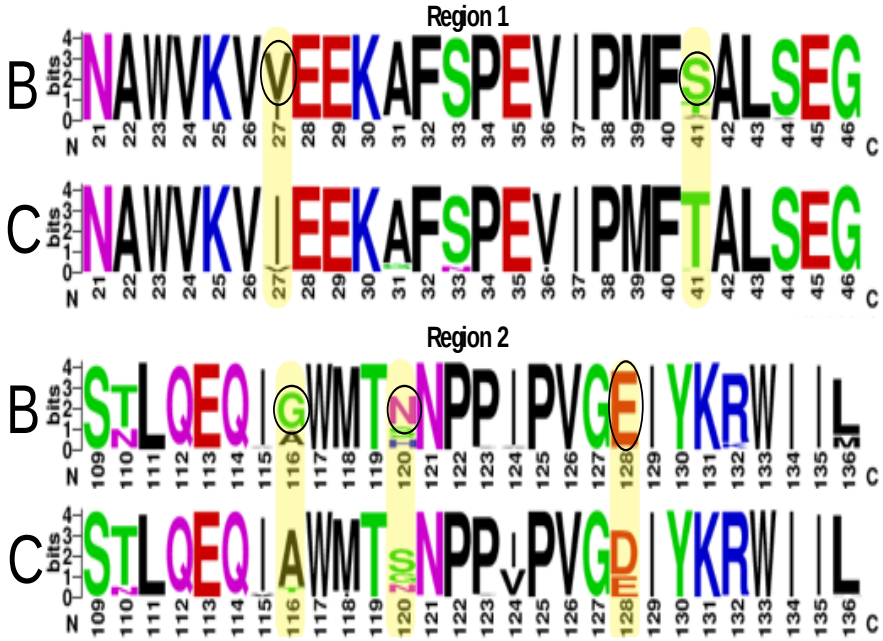



**Fig. S1: Subtype-specific differences between the two conserved, clinically-important HIV Gag Regions in HIV-B (B) and HIV-C (C)**

The subtype-specific positions (SSP) 27, 41, 116, 120, and 128 are highlighted in yellow, and the letters representing site-specific amino acids are sized according to their associated frequencies. HIV-B position 128 and HIV-C positions 27, 41, and 116 are almost always occupied by the consensus amino acids and cannot be used in the analysis. The HIV-B consensus amino acids are circled. This figure was generated using a single sequence from each HLA annotated patient in the HIV database (Foley et al. 2018).

**Figure S2**

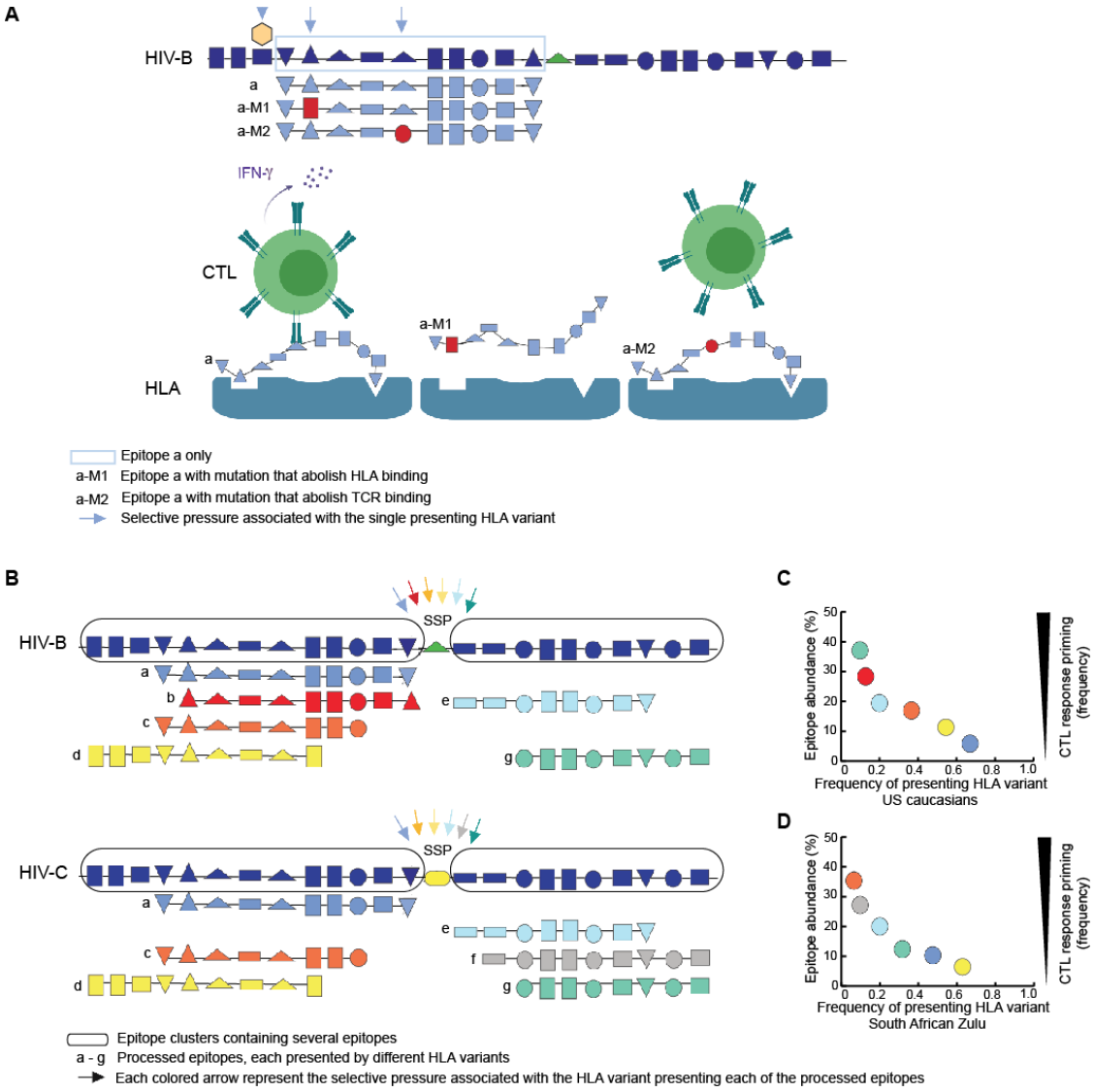



**Fig. S2. Outline of HLA-associated single-epitope and multiple-epitope selective pressures**.

**(A).** Three common forms of HLA-associated single epitope selective pressures. CTL responses targeting an epitope might select for an upstream mutation (indicated by the orange polygon) that disrupts ERAP trimming (as in (Draenert et al. 2004)) or intra-epitope mutations that abrogates epitope binding to the presenting HLA molecule or T cell receptor (TCR) recognition of the epitope (reviewed in (Goulder and Watkins 2008; Goulder and Walker 2012)). Arrows signify different forms of CTL selective pressures, geometric symbols indicate amino acids, and the green triangle represents an HIV-B subtype-specific amino acid (as in (**B**)).

**(B).** Schematic representation of a conserved p24Gag region in HIV-B and HIV-C with a subtype-specific amino acid position (SSP; HIV-B, green triangle, HIV-C, yellow ellipse) and down- and upstream epitope clusters; epitopes a-g indicate epitopes processed by intra-cellular proteasomes. When a patient presents any of these epitopes, additional HLA-associated selective pressures might select for the intra-epitope CTL-escape mutations described in **Figure S2A**. The nature of the amino acid in the subtype-specific position controls intra-cellular proteasomal production of the surrounding epitope-clusters (Tenzer et al. 2014), which each can contain 20-50 epitopes presented by approximately as many HLA variants (Foley et al. 2018). The outlined structure consisting of partly overlapping epitope sequences in clusters in hydrophobic regions that are separated by a subtype-specific position is also common in other HIV-1 proteins (Foley et al. 2018; Lucchiari-Hartz et al. 2003).

**(C), (D)**. Schematic outline of the outcome of the HLA-associated selective pressure on the subtype-specific positions (experimentally demonstrated in (Tenzer et al. 2014)). The combined HLA-associated selective pressure on the SSP in an HIV-infected population results in an inverse relationship between the abundance of a processed epitope and the frequency of the presenting HLA allele in the population in which the virus circulates (Tenzer et al. 2014). In this hypothetical example, the grey epitope is not produced when the proteasome processes HIV-B because the HLA variants that can present the grey epitope is very common in US Caucasians. In contrast, the red epitope is not produced when the proteasome digests HIV-C because the restricting HLA variant is very common in the South African Zulu population. The likelihood of CTL priming will increase with the amount of presented epitope on the infected cell surface (Faroudi et al. 2003; Tenzer et al. 2009).

**Figure S3**

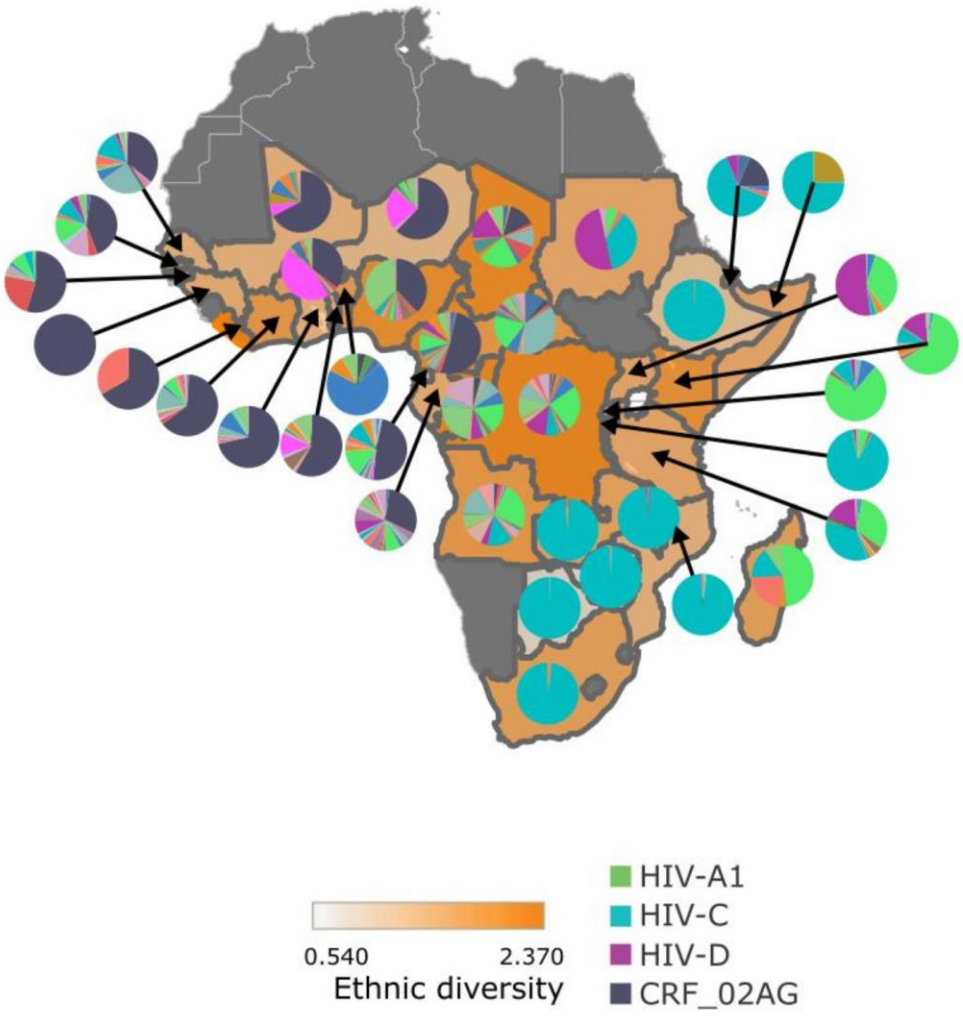



**Fig. S3: HIV-1 subtype and CRF distribution in Sub-Saharan Africa**

This figure is similar to **Fig. 3d**, but the pie charts have not been scaled according to the number of sequences from each country. The map showing the ethnic diversity within Sub-Saharan Africa (orange shaded background) is overlaid with pie charts demonstrating HIV-1 subtype diversity within each country. Missing countries (dark grey) lacked either HIV-1 or ethnic fractionalization data; complete subtype key can be found in **Table S4**.

**Figure S4**

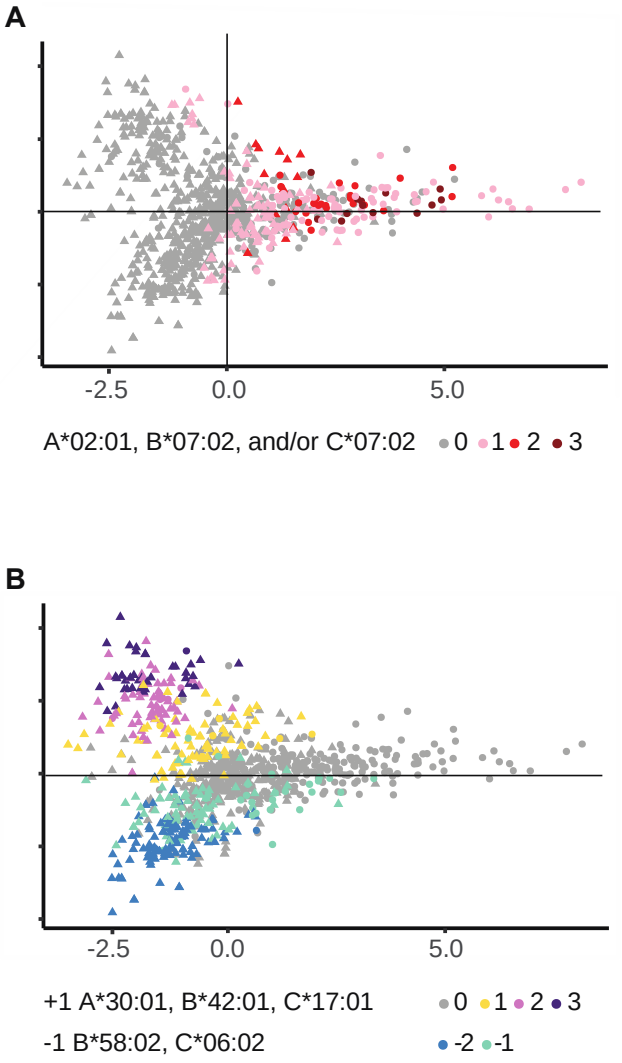



**Fig. S4: The patients' haplotype score in PC1 and PC2**

**(A).** The first two PCs explained 14.5% of the variance (PC1 = 7.7%, PC2 = 6.8%). The haplotype score is not stratified for Caucasian HLA variants as the HLA A*02:01 contributes most to the variance. HLA B*07:01 and HLA C*07:01 are in LD in Caucasians (a haplotype inherited from Neanderthals (Abi-Rached et al. 2011)), but not in Africans (Gonzalez-Galarza et al. 2015).

**(B).** The patient's haplotype score is incremented by one for each allele from haplotype HLA A*30:01-B*42:01-C*17:01 and reduced by one for each allele from haplotype HLA B*58:02-C*06:02 showing near-complete stratification of PC2 due to these haplotypes.

**Figure S5**

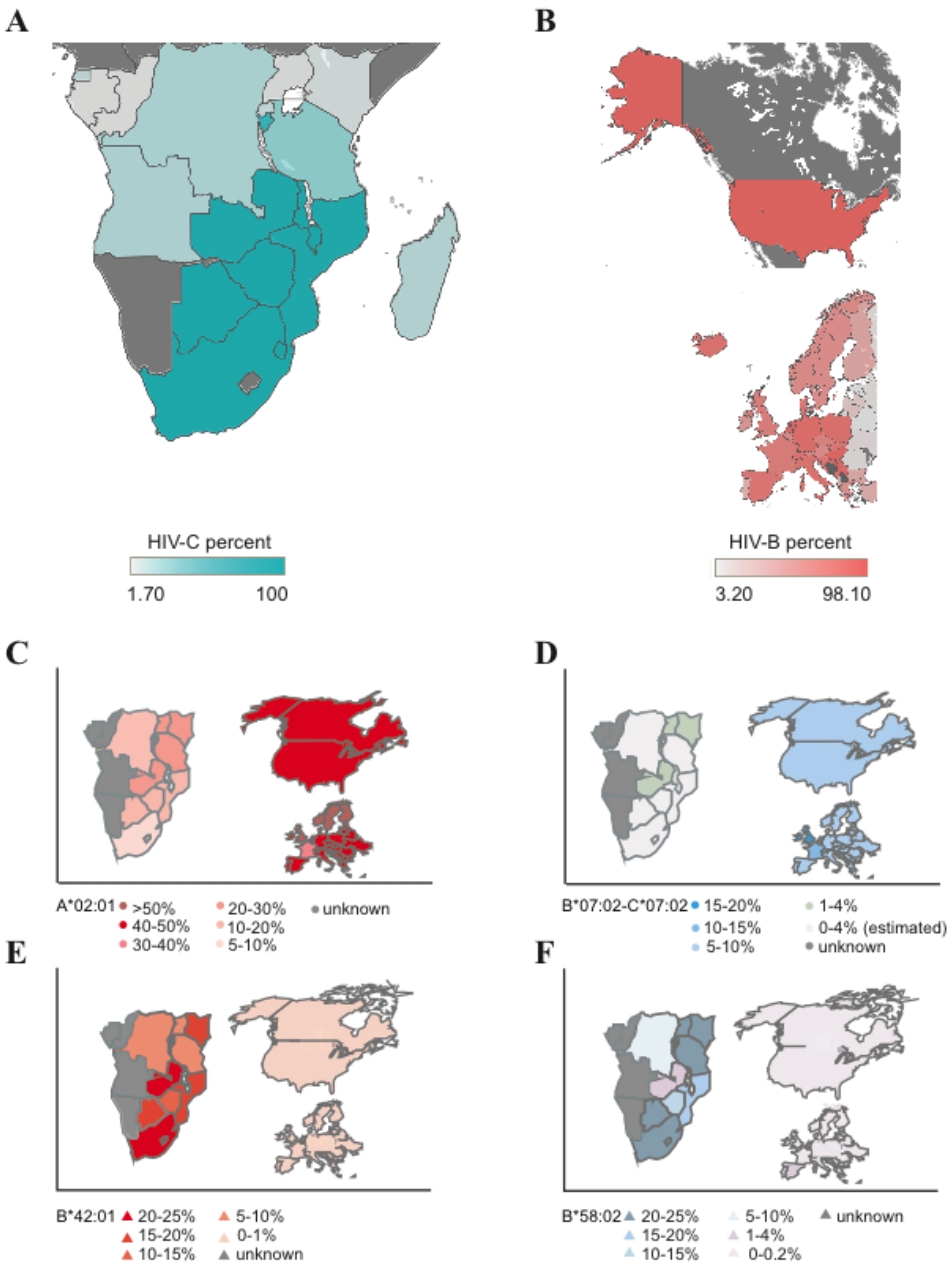



**Fig. S5: HIV-C, HIV-B and selected HLA frequencies in Southern Africa, the US, and Europe**

**(A)**. The percentage of HIV-C in countries in Southern African. Dark grey signifies either no data or that the region is outside the geographical region of interest.

**(B)**. The percentage of HIV-B in the US and Europe. The percentage of non-HIV-B sequences in Europe is higher than in the US due to a higher proportion of non-B HIV-1-infected immigrants and refugees primarily from Africa. Furthermore, the epidemic in some European countries was in part, or mostly, founded by non-B subtypes; for example, the percentage of HIV-B is low in Russia and former Soviet Union countries due to the introduction of HIV-A from the DRC and HIV-G in Portugal because of the introduction of HIV-G from Cape Verde, a former Portuguese colony (Beloukas et al. 2016; Diez-Fuertes et al. 2015).

**(C-F)**. The HLA frequencies of key HLA variants in PC1 and PC2. Note the low frequency of HLA-B*58:02 in Zambia where the epidemic has lasted longer than in South Africa, and where HLA B*42:01, but not HLA B*58:02, is associated with facilitated HIV-1 transmission. The HLA B*07:02-C*07:02 haplotype was acquired from Neanderthals and is found primarily in Eurasians (Abi-Rached et al. 2011). HLA B*07:02 and HLA C*07:02 can be found individually in other combinations in some African ethnic groups (Gonzalez-Galarza et al. 2015).

**Figure S6**

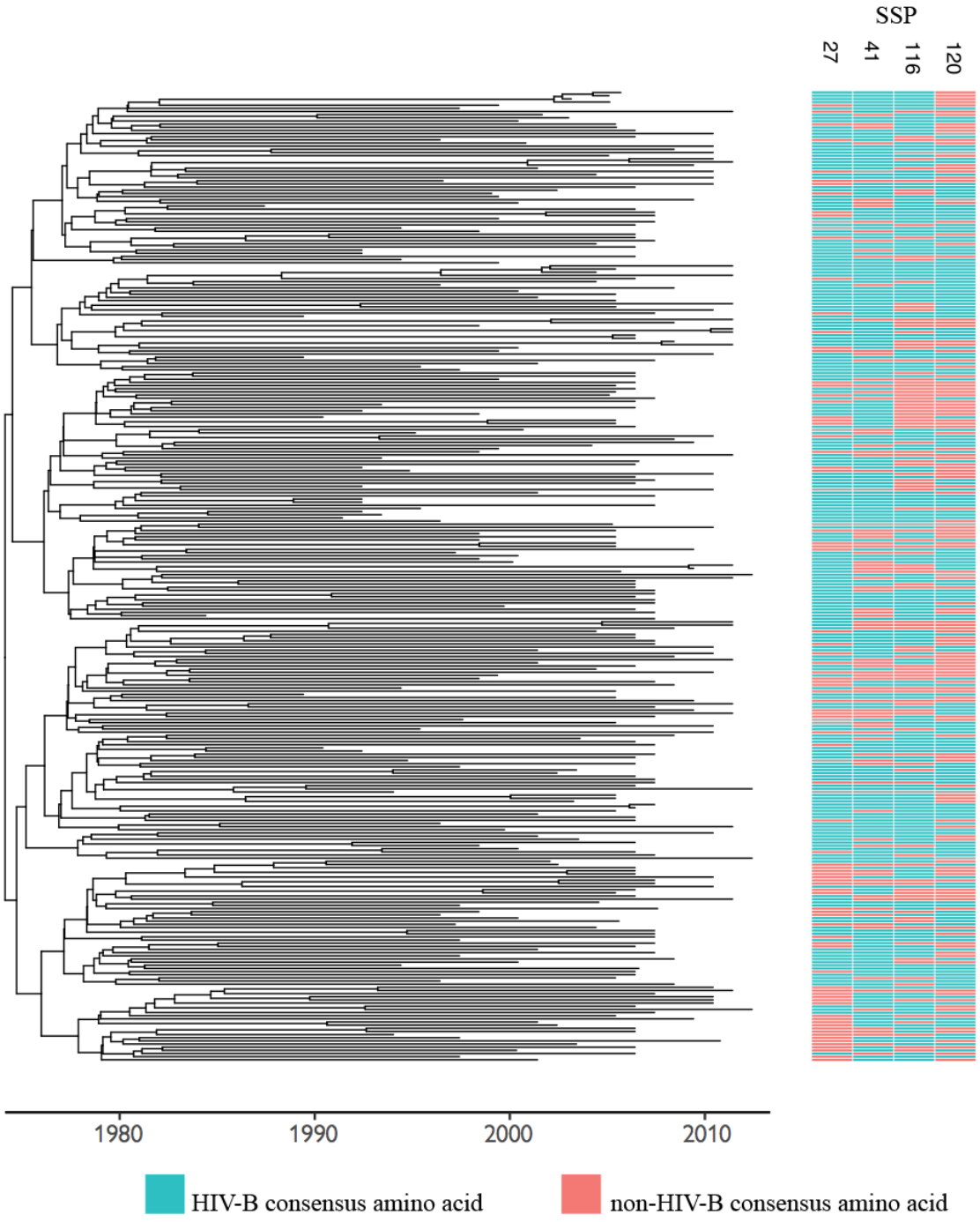



**Fig. S6: HIV-B phylogeny**

The HIV-B BEAST Maximum Clade Credibility phylogeny annotated with the subtype-specific positions (SSP) 27, 41, 116, and 120 (see **Fig. S1**). Every leaf-node on the phylogeny represents a patient, and the four lines next to the leaf node show whether the majority rule amino acid in each

of the four eligible SSPs was identical to the HIV-B consensus amino acid. The amino acid patterns shown here follow the phylogeny (i.e., continuous patches of red are more vertical than horizontal, and are therefore shared between closely related sequences rather than shared between multiple positions on the same patient), and demonstrate the necessity of using a multiple response random effect model where each amino acid is allowed to evolve independently. If the color pattern had been more horizontal, that would have meant a single patient's subtype-specific amino acids tended to change as a block, in which case a single patient parameter (random effect model) would have been more appropriate than the multiple random effect phylogenetic model used in this study.

**Figure S7**

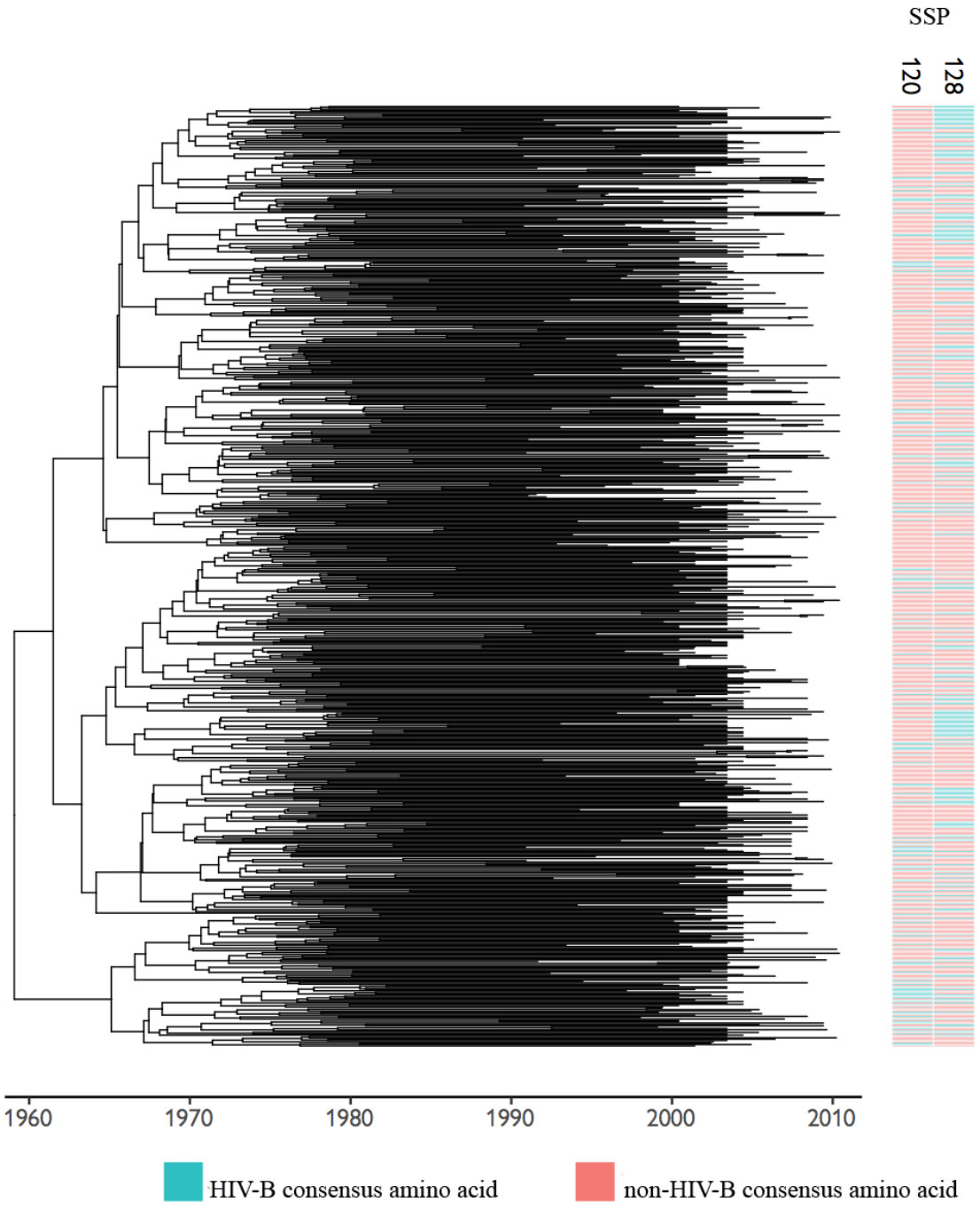



**Fig. S7: HIV-C phylogeny**

The HIV-C BEAST Maximum Clade Credibility phylogeny annotated with the subtype-specific positions (SSP) 120 and 128. Patients, where the majority rule amino acid is the same as the HIV-B (see fig. S1), are colored blue; all other amino acids are colored red.

**Figure S8**

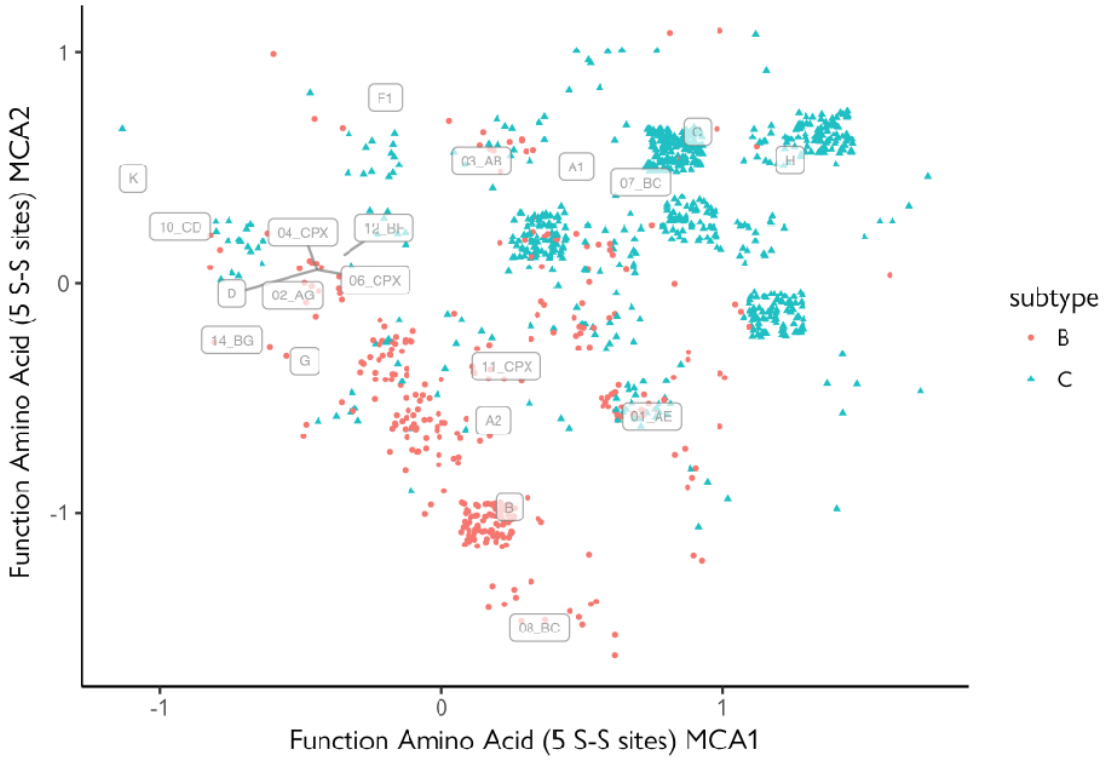



**Fig. S8: MCA of each patient's HIV-1 subtype-specific amino acids**

The MCA converts each patient's HIV subtype-specific position amino acid profile into two dimensions. This MCA was trained using the subtype consensus sequences (labels), the MCA was then used to project the patient's HIV subtype-specific amino acids onto the two-dimensional space. Points were jittered to prevent them obscuring one another. While many patients' subtype-specific amino acids are identical to the subtype consensus (see large clusters around the HIV-B and HIV-C labeled consensus sequences), other HIV-B and HIV-C sequences are identical in the five Gag subtype-specific sites studied here (e.g., around the labeled HIV-1 circulating recombinant (CRF) 01_AE consensus).

**Figure S9**

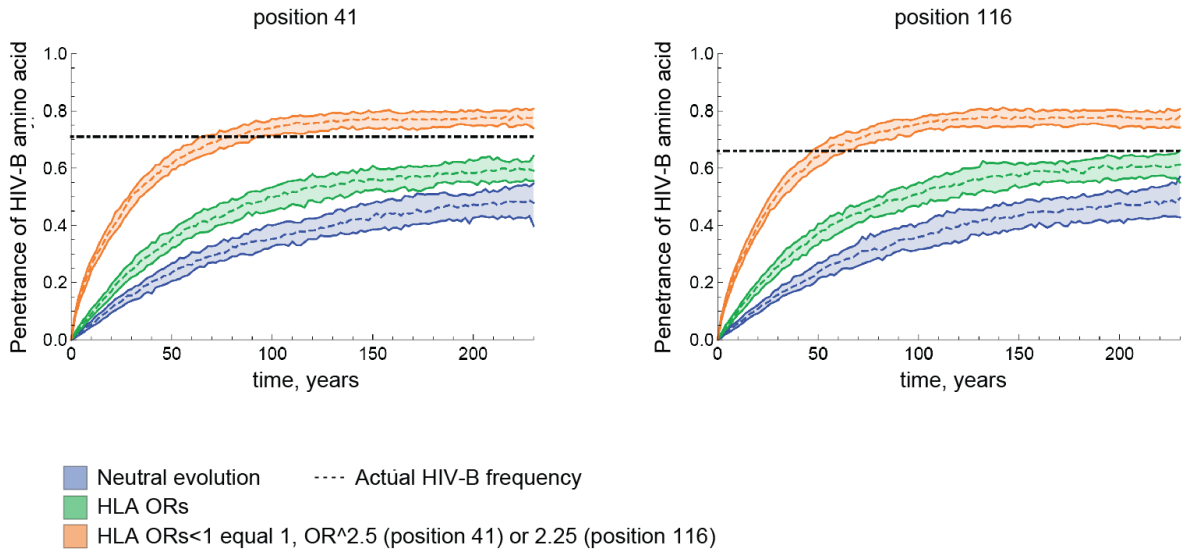



**Fig. S9: Modeling of the evolution of HIV-C SSAA using HIV-B-related HLA ORs**

We used an agent-based model (**1.3**) to estimate the effect of neutral and HLA-mediated selection pressures, respectively, on the evolution of p24Gag positions 41, and 116 on a fictive HIV-1 with HIV-C-like subtype-specific amino acids (SSAAs). For position 41, raising the odds to 2.9 or 3.0 made little difference. These results suggest that other, unidentified, HLA variants might influence the evolution of these positions and/or that structural co-evolutionary factors might be in play.

**Figure S10**

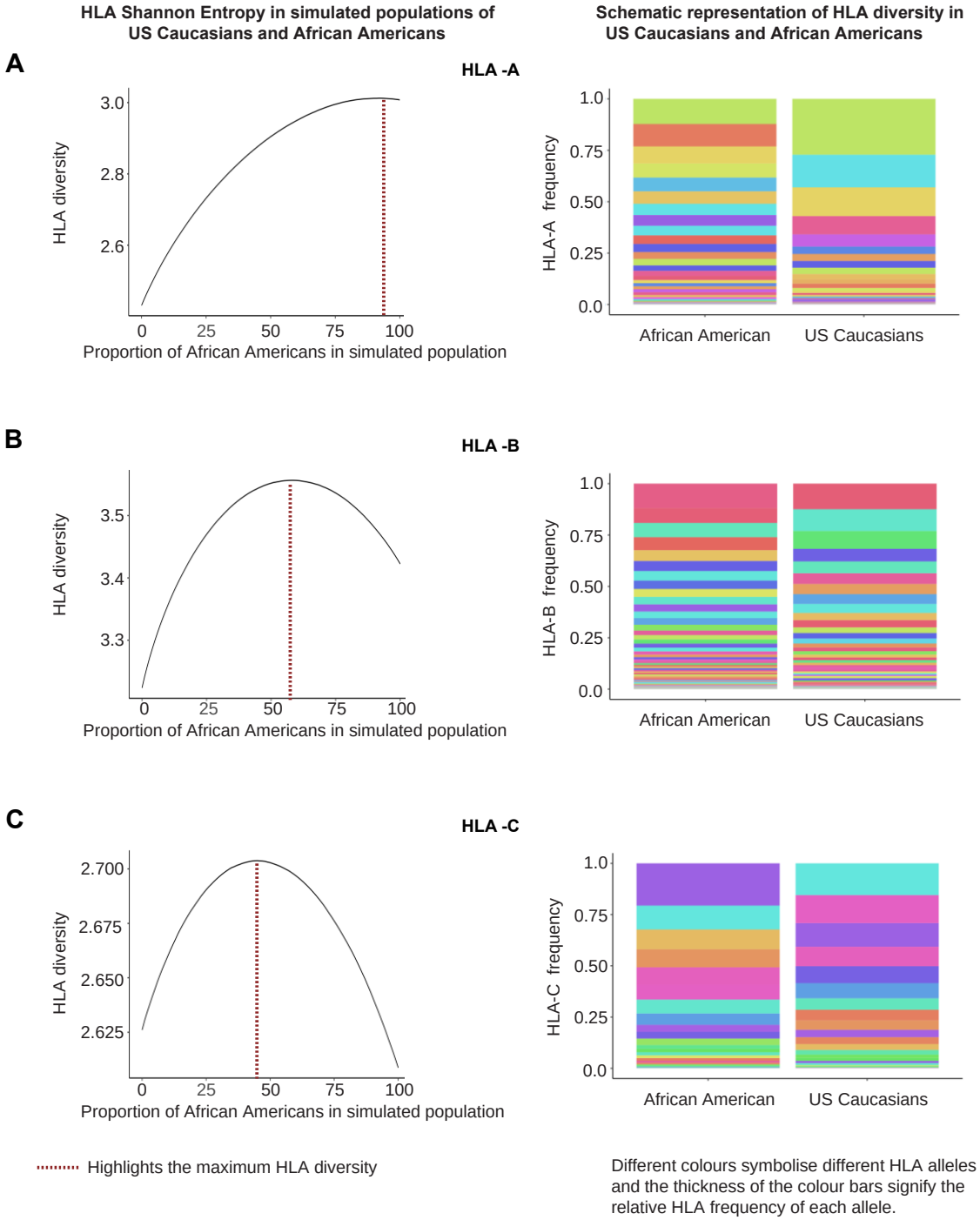



**Fig. S10: Simulated mixtures of African Americans and US Caucasians (left) and HLA diversity in African American and US Caucasian populations (right)**

**(A)**. HLA A diversity in simulated populations with different proportions of African Americans and US Caucasians, and the HLA A variant frequencies in African Americans and US Caucasians (modified from (Gragert et al. 2013)).

**(B, C)**. As in **A** for **B** (HLA B), and **C** (HLA C).

**Figure S11**

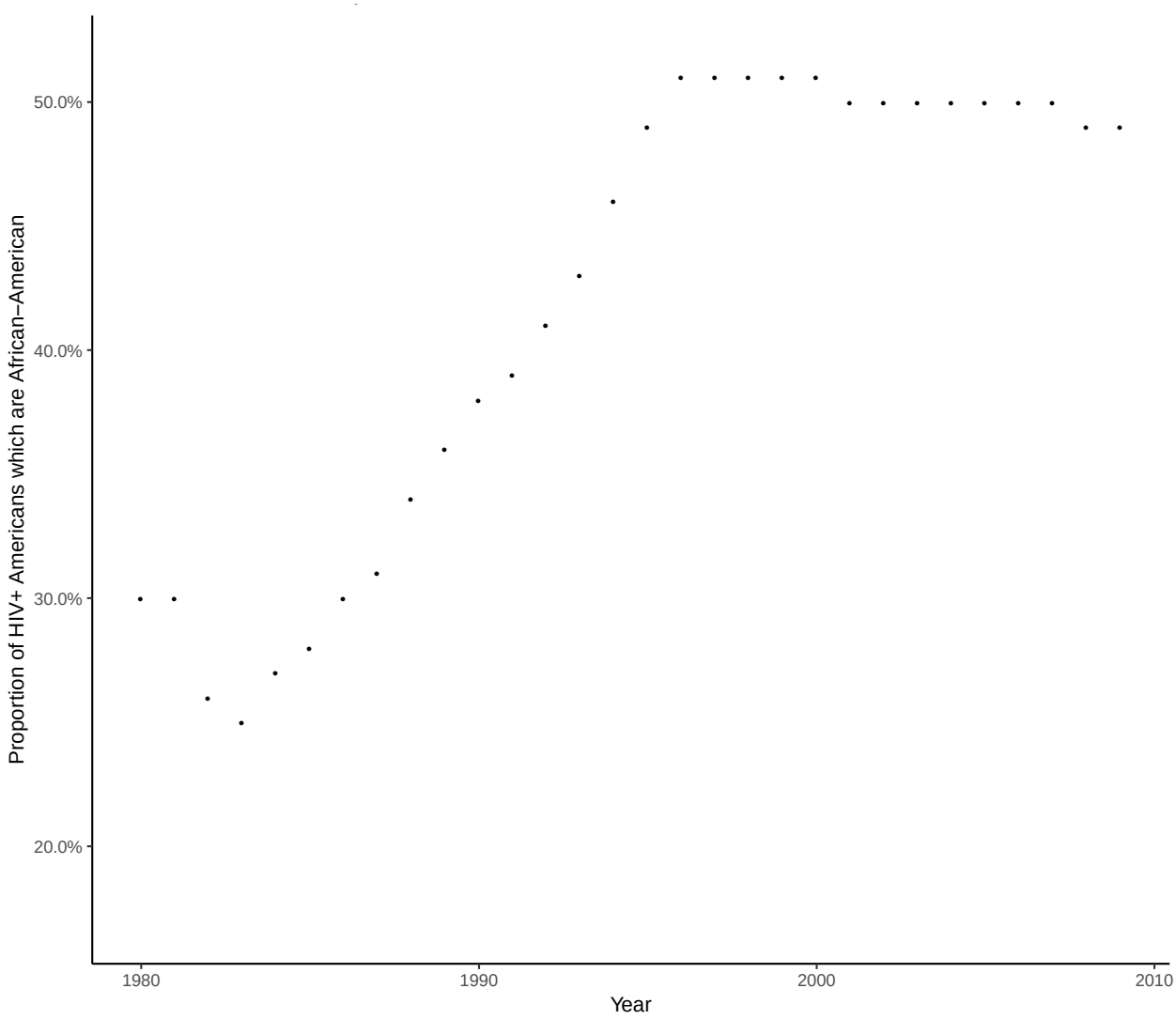



**Fig. S11**: **Proportion of African Americans in the US HIV-infected population over time**

The ethnic demographic proportion of African Americans in the US HIV-infected population (data from (Hall et al. 2008) and (Prejean et al. 2011)) as calculated in the simple population model with the proportion of African Americans shown as points. Samples drawn from the posterior distribution of the change-point model are shown as lines. Note the increased uncertainty around 1980 due to data limitations.

**Figure S12**

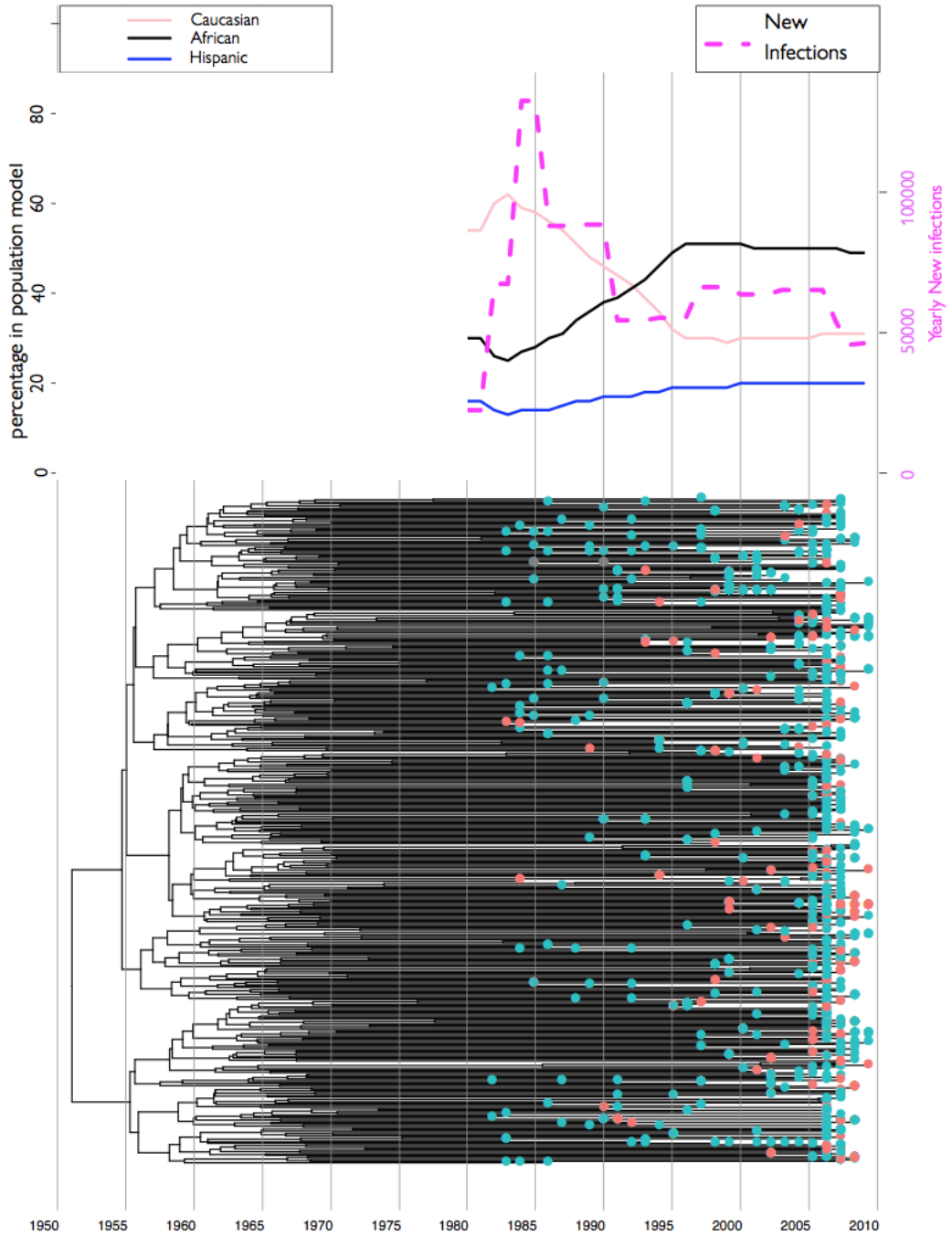



**Fig. S12: Phylogeny of the American HIV-B epidemic combined with demographic data**

The Maximum Clade Credibility tree, annotated with individual sequence data for position 27 (bottom, blue if a sequence is identical to the HIV-B consensus amino acid; red otherwise). In this case all non-B consensus amino acids were similar to the amino acid found in the HIV-C

consensus sequence due to the limited variation possible at this position (**Table S8**). The phylogeny was generated using one sequence per patient, but position 27 amino acid data from multiple sequences sampled from the same patient have been superimposed onto the phylogeny. These annotations illustrate the gradual increase of non-HIV-B consensus amino acids on position 27 as the African-American proportion of HIV-infected individuals in the United States increases (black line, top figure). No demographic data are available before 1980.

**Table S1: African populations with HLA A, B, and C data**

Table showing the only African populations with publicly available 'gold standard' HLA data (i.e., all HLA types were reported as 4-digits HLA variants)(Gonzalez-Galarza et al. 2015).

| Number | Population or ethno-linguistic group | Country | Number of people in cohort |
|---|---|---|---|
| **1** | Bandiagara | Mali | 138 |
| **2** | Ga-Adangbe | Ghana | 131 |
| **3** | Kenya* (likely Masai) | Kenya | 144 |
| **4** | Luo | Kenya | 265 |
| **5** | Nandi | Kenya | 240 |
| **6** | South Africa Black** | South Africa (Free State) | 200 |
| **7** | Zulu | South Africa (KwaZulu Natal) | 606 |
| **8** | Kampala (pop 2)*** | Uganda | 175 |

* The population named 'Kenya' has no ethnic group specifications but the location coordinates (obtained from the database (Gonzalez-Galarza et al. 2015)) are situated within the Masai ethnic group's territory, and the population is described as 'rural.' These coordinates differ from those of the Luo and Nandi ethnic groups, and for this reason, we assumed all three groups were distinct from each other. The data have been presented in a workshop but have not been published. The Luo people in western Kenya are closely related to Luo people in Northern Uganda, and Northern Tanzania, and are part of a larger group of ethno-linguistically-related Luo people in South Sudan and South-western Ethiopia. Their language (Dholuo) belongs to the Western Nilotic branch of the Nilo-Saharan language family. The Nandi people live in Nandi Country (previously Rift Valley Province of Kenya) and speak Nandi, which belongs to the Southern Nilotic language group. The Masai people live in Northern, Central, and Southern Kenya and in Northern Tanzania. Their language (Maasai or Maa) belongs to the Eastern Nilotic language group.

** The location coordinates of the South Africa Black HLA data is situated within the Free State province of South Africa, a region where most people speak Sotho and belong to the Sotho-

Tswana people, however, in the paper by Paximadis, M et al., the HLA data is reported to represent “a cross-section of the black and Caucasian subgroups of the South African population (Paximadis et al. 2012).” The South Africa Black and Caucasian HLA data are recorded separately in the database. The number of South Africa Black ethnic groups is ~10, but the majority of the Zulu population lives in the KwaZulu Natal and Gauteng provinces (http://www.sahistory.org.za/).

*** The HLA data from Uganda does not belong to a distinct population; “All the individuals were Black Africans from Kampala, but no other ethnic affiliation was recorded” (from (Kijak et al. 2009)).”

**Table S2: The number of p24 Gag sequences per country and the number of HIV-1 subtypes, CRFs and URFs per country and the country code abbreviation key**

| ISO alpha-2 two-letter Country code | Country name | p24 sequences / country |
|---|---|---|
| **BW** | Botswana | 51 |
| **CD** | Democratic Republic of the Congo | 22 |
| **CI** | Ivory Coast | 6 |
| **CM** | Cameroon | 318 |
| **ET** | Ethiopia | 35 |
| **GA** | Gabon | 6 |
| **GH** | Ghana | 13 |
| **KE** | Kenya | 633 |
| **NE** | Niger | 72 |
| **NG** | Nigeria | 19 |
| **SN** | Senegal | 27 |
| **TD** | Chad | 18 |
| **TZ** | United Republic of Tanzania | 38 |
| **ZA** | South Africa | 935 |
| **ZM** | Zambia | 12 |

| ISO alpha-2 two-letter Country code | Country name | HIV-1 subtypes, CRFs, and URFs (group M) / country |
|---|---|---|
| **AO** | Angola | 21 |
| **BF** | Burkina Faso | 13 |
| **BI** | Burundi | 8 |
| **BJ** | Benin | 6 |
| **BW** | Botswana | 7 |
| **CD** | Democratic Republic of the Congo | 46 |
| **CF** | Central African Republic | 30 |
| **CG** | Congo | 20 |
| **CI** | Ivory Coast | 22 |
| **CM** | Cameroon | 78 |
| **DJ** | Djibouti | 6 |
| **ET** | Ethiopia | 7 |
| **GA** | Gabon | 33 |
| **GH** | Ghana | 18 |
| **GM** | Republic of the Gambia | 10 |
| **GN** | Guinea | 1 |

| | | |
|---|---|---|
| **GQ** | Equatorial Guinea | 16 |
| **GW** | Guinea Bissau | 7 |
| **KE** | Kenya | 35 |
| **LR** | Liberia | 2 |
| **MG** | Madagascar | 5 |
| **ML** | Mali | 10 |
| **MW** | Malawi | 8 |
| **MZ** | Mozambique | 8 |
| **NE** | Niger | 13 |
| **NG** | Nigeria | 23 |
| **RW** | Rwanda | 10 |
| **SD** | Sudan | 5 |
| **SN** | Senegal | 35 |
| **SO** | Somalia | 2 |
| **TD** | Chad | 19 |
| **TG** | Togo | 21 |
| **TZ** | United Republic of Tanzania | 19 |
| **UG** | Uganda | 33 |
| **ZA** | South Africa | 30 |
| **ZM** | Zambia | 14 |
| **ZW** | Zimbabwe | 7 |

Data were obtained from the HIV data base (Foley et al. 2018).

# Table S3: Complete key to the African ethnicities in Fig. 2C


JIBU
JIDA
JIE
JIJI
JIMI
JIMINI
JIRU
JITA
JIU
JONGA
JOOLA
JOPADHOLA
JUKUN
JUKUN KONA
JUKUN WASE
JUKUN WURKUM
JUNGO
JUR
JUR MODO
JWIRA-PEPESA
KAAMBA
KABA
KABA DEMI
KABA NAA
KABA SO
KABONGA
KABRE
KABYE
KABYLE
KADARA
KADARU
KAFFA
KAGOMA
KAGORO
KAGURU
KAJAKSE
KAKO
KAKONDA
KALAMSE
KALANGA
KALIKO
KALUNDWE
KAM
KAMANGA
KAMANTAN
KAMBA
KAMBARI
KAMBATA
KAMIR
KAMKAM
KAMO
KAMUKU
KAMWE
KANAKURU
KANDAWIRE
KANDERMA
KANEMBU
KANGO
KANTANA
KANU
KANUFI
KANURI
KANYOK
KAONDE
KAONDE, SANGA
KARA
KARABORO
KARANGA
KARE
KAREKARE
KASEM
KASENA
KASONGI
KASONGO
KASONKE
KATAB
KATYA
KAWENDI
KAYAMBA
KAYLA
KEDERU
KEENGE
KEL
KEL AHAGGAR
KEL AYR ASBEN
KEL GRES
KEL IFFORA
KELE
KENGA
KENYANG
KENYI
KERA
KERARISH
KEREWE
KETE
KETE EAST
KETE LULUA
KETE NORTH
KETE SOUTH
KEYO
KHARGA
KHUMALO
KIBET
KIGA
KIKUYU
KIM
KIMBU
KIMBUNDU
KINGA
KIONG
KIPSIKIS
KIR-BALAR
KISAMA
KISSI
KITA
KIYAK
KLAOH
KO
KOBIANA
KOFFA
KOFYAR
KOHUMONO
KOKE
KOKO
KOL BIKELE
KOLBILA
KOM
KOMA
KOME
KONGO DINGA
KONJI
KONJO
KONKOMBA
KONO
KONONGO
KONSO
KONYANKE
KOONI
KORO
KOROP
KOTA
KOTE
KOTOKO
KOTOPO
KOUYA
KPELLE
KPESSI
KRAN, NGERE
KRESH
KRIM
KUANG
KUDU
KUFRA
KUKA
KUKELE
KUKU
KUKUYA
KULANGO
KULERE
KULUNG
KUMU
KUNAMA
KUNDA
KUNDU
KUNTA
KUNUZ
KUNYI
KUPTO
KURAMA
KURANKO
KURFEI
KURI
KURUMFE
KURYA
KUSAAL
KUSHI
KUSU
KUTEP
KUTIN
KUTU
KUTURMI
KUYERI
KUYU
KWAAMI
KWAFI
KWAKUM
KWAME
KWANDI
KWANGARI
KWANGWA
KWANJA
KWANKA
KWAYA
KWERE
KWESE
LAALI
LABWOR
LADO
LAGBA
LAGUAT
LAGWAN
LAKA
LALA
LALA ROBA
LALIA
LAMA
LAMBA
LAMBYA
LAMJA
LANDUMA
LANGA
LANGO
LARU
LEELAU
LEGA
LEGBO
LEKA
LELE
LEMBWE
LEMORO
LENDU
LENGI
LENGOLA
LENJE
LESE
LIA
LIBOLO
LIGBI
LIGI
LIJILI
LIKO
LIKWALA
LIMA
LIMO
LO
LOBI
LOBI LORHON
LOGO
LOGORIK
LOI
LOKALO
LOKELE
LOKO
LOKORO
LOKOYA
LOMA
LOMBI
LOMOTWA
LOMWE
LONDO
LONGARIM
LONGTO
LOPA
LOSAKANYI
LOTSU-PIRI
LOVALE
LOVEDU
LOZI
LUANO
LUBA
LUBA EASTERN
LUBA KASAYI
LUBA SHANKADI
LUBA UPEMBA
LUBA, LUNDA
LUBA, NDEMBU
LUBA, NDEMBU, LUNDA
LUBILA
LUCHAZI
LUGBARE
LUGURU
LULA
LULUBA
LULUWA
LUMBU
LUNDA
LUNDA, CHOKWE
LUNDA, KETE
LUNDA, NDEMBU
LUNGU
LUNTU
LURI
LUSHANGI
LUUNDA
LUYIA
LWALU
LWENA
LWER
LWIMBI
LYANGALILE
LYELE
MA
MAAKA
MAASAI
MAAZA
MABA
MABENDI
MABEZA
MABIHA
MABINZA
MACHINGA
MADA
MADI
MAGHARBA
MAGUZAWA
MAHAFALY
MAHONGWE
MAK
MAKAA
MAKERE
MAKOMA
MAKONDE
MAKUA
MAKWA
MALINKE
MALWAL
MAMBAI
MAMBILA
MAMBWE
MAMPRUSI
MAMVU
MANALA
MANDA
MANDARI
MANDINKA
MANDYAK
MANGAS
MANGAYAT
MANGBELE
MANGBETU
MANGBUTU
MANINKA
MANKANYA
MANO
MANYANGA
MANYIKA
MANZA
MAO SOUTH
MAPOPOI
MARARIT
MARBA
MARFA
MARGHI
MARGHI SOUTH
MARKA
MASAKIN
MASALIT
MASANA
MASHASHA
MASHI
MASMAJE
MASONGO
MASSA
MASSALAT
MATENGO
MATUMBI
MAU
MAURI
MAYOGO
MBA
MBAAMBA
MBAATI
MBAGANI
MBAI BEDIONDO
MBALA
MBANGWE
MBANZA
MBATA
MBEKO
MBELIME
MBELO
MBEMBE
MBEMBRE
MBENE
MBERE
MBESA
MBINSA
MBO
M'BOKO
MBOLE
MBOMOTABA
MBONGO
MBOSHI
MBUGWE
MBUKUSHU
MBULI
MBULU
MBUM
MBUNDA
MBUNGA
MBUUN
MBWELA
MBWELA KOLWE
MBWERA
MDUNDULU
MEDJE
MEDOGO
MENDE
MENING
MERINA
MERU
MESME
MFINU
MFUNU
MIDOB
MIGAAMA
MILTU
MIMI
MINIANKA
MINUNGO
MINUNGU
MITSOGHO
MITTU
MITUKU
MIYA
MOBA
MOBANGO
MOBENGE
MOKOLE
MOLO
MONDOMBE
MONDZOMBO
MONGO
MONO
MONTOL
MOROCCAN
MORU
MOSSI
MOTEMBO
MOYE
MPAMA
MPANGU
MPE
MPEZENI'S NGUNI
MPONGMPONG
MPOTO
MPUT
MPUTU
MPYEMO
MSER
MUBI
MUCHIKONGO
MUKULU
MUMUYE
MUN
MUNDANG
MUNDU
MUNGA
MURLE
MUSGU
MUSSEY
MUSSUMBE
MVELE
MWAGHAVUL
MWENYI
MYANG
MYENE
NABAN
NAFANA
NAGUMI
NALU
NAMBA
NAMJI
NANCERE
NANDE
NANDI
NANDU-TARI
NANDYA
NATENI
NATIORO
NAWDM
NBWERA
NDAAKA
NDALI
NDAM
NDAMBA
NDAU
NDEBELE
NDEMBU
NDEMLI
NDENDEULE
NDENGEREKO
NDENGESE
NDIBU
NDO
NDOGO
NDOLO
NDOOLA
NDULU
NDUNGA
NDZABI
NDZUNDZA
NEFUSA
NEYO
NGABA
NGALA
NGALANGI
NGAM
NGAM GIR BOR
NGAMBAI
NGAMO
NGANDA
NGANDO
NGANDU
NGANDYERA
NGANGELA
N'GARE
NGASA SHAKA
NGATA
NGBAGA
NGBAKA
NGBANDI
NGEENDE
NGENGE
NGENGELE
NGENZA
NGGWAHYI
NGHWELE
NGINDO
NGIZIM
NGOLA
NGOMBE
NGONGI
NGONGO
NGOUNGWONI
NGUL
NGULU
NGUMBA
NGUMBO
NGUNDI
NGUNGULU
NGUNI
NGURIMI
NGWABA
NHAM
NIABWA
NIELLIM
NIMBARI
NINZAM
NJAJGULGULE
NJAMUS
NKANGALA
NKANSI
NKANU
NKHUMBI
NKOLE
NKOMI
NKOYA
NKUKOLI
NKUTSHU
NKWE
NOBIIN


NOLE
NOWOLO
NSAMBA
NSAPO
NSENGA
NSONGO
NTANDU
NTCHAM
NTOMBA
NUBA
NUER
NUMANA
NUNA
NUNGU
NUNU
NUNUMA
NUPE
NWENSHI
NWERA
NYAKYUSA
NYALA
NYALI
NYAMWANGA
NYAMWEZI
NYANEKA-HUMBE
NYANGA
NYANGATOM
NYANGBARA
NYANJA
NYARAFOLO
NYATURU
NYEMBA
NYENGO
NYIHA
NYILAMBA
NYIMANG
NYINDU
NYONG
NYORO
NYULI
NZAKARA
NZANYI
NZIMA
OBOLO
OBULKOM
ODUT
OKAK
OKO-ENI-OSAYEN
OKPE
OKPE-AKUKU
OKPELA
OMBO
OMETO
OMONO
OOLI
OPA
OPAMERI
OPUUO
ORANA
ORING
OROMO
ORON
OTANG
OTUHU
OUASSA
PAI
PALLAKA
PAMBIA
PANA
PANDE
PANI
PAPEL
PARE
PARE, SAA
PATU
PAYE
PENDE
PERE
PERO
PEVE
PHOKA
PIMBWE
PINDA
PINDI
PITI
PIYA
PODJULU
PODZO
POGORO
POKOMO
POMBO
POMO
PONGWE
POTO
PUGULI
PUKU
PUNU
PYAANG
QUILENGUE
RASHAD
REGEIBAT
RENDILL
RIF
RIYAH
RONGA
RUARHA
RUBASA
RUFIJI
RUNGA
RUNYANKOLE
RURI
RUSHA
RUUND
RUWENG
SAAB
SAADI
SABA
SAFWA
SAGALA
SAGARA
SAHEL
SAHO
SAKA
SAKALAVA
SAKATA
SALA
SALA MPASU
SAMBA
SAMBA LEKO
SAMBO
SAMBU
SAMBURU
SAN
SANGA
SANGO
SANGU
SANUSI
SAPAUT
SAR
SARA
SARA GULA
SARA GULAY
SARO
SARUA
SASARU
SAYA
SEBA
SEEKU
SEKE
SENA
SENGA
SENGELE
SERE
SERER
SESE
SH
SHAGAWU
SHAIKIA
SHALL
SHAMBAA
SHANJO
SHATT
SHEBELLE
SHERBRO
SHI
SHIKI
SHILA
SHILLUK
SHINJI
SHIRAWA
SHLUH
SHONA
SHOO
SHUBI
SHUKRIA
SHUWA
SIDAMO
SIDI
SIHANAKA
SIMAA
SINASHA
SINGA
SINYAR
SIRTICAN
SISAALA
SISYA
SIWA
SIZAKI
SO
SOKORO
SOLA
SOLI
SOLONGO
SOLU
SOMRAI
SOMYEWE
SONGHAI
SONGO
SONGOLA
SONGYE
SONINKE
SONJO
SOONDE
SOSSO
SOTHO
SOUTHERN BANGANTU
SUAFA
SUBA
SUBIYA
SUGA
SUK
SUKU
SUKUMA
SUKWA
SUMA
SUMBWA
SUNDI
SUNGOR
SURI
SURUBU
SUSU
SWAKA
SWAZI
SYEMU
TABWA
TAFIRE
TAGWANA
TAHOU
TAJAKANT
TAJUASO
TALE
TALENSI
TAMA
TAMAZIGHT
TAMBAHUAKA
TAMBAS
TAMBERAMA
TAMBO
TAMEZRET
TAMPOLENSE
TANALA
TANDA
TANGA
TANGALE
TANGBAGO
TANKARA
TAPSHIN
TAROK
TASUMSA
TATOG
TAWANA
TAZARAWA
TCHWABO
TEDA
TEFASI
TEGE
TEITA
TEM
TEMNE
TENDA
TEPETH
TERA
TERE
TESO
TETELA
TEVUNDRU
TIBA
TIBEA
TIE
TIENE
TIGRAY
TIGRINYA
TIKAR
TIKUU
TIO
TITU
TIV
TOBANGA
TOGBO
TOKA
TOMA
TONGA
TONGA-INHAMBANE
TONGWE
TOPOKE
TOPOSA
TORAM
TORO
TOROBE
TOTELA
TOUGOURT
TOUYO
TOW
TRARZA
TRIBOUE
TRIPOLITANIAN
TSAAM
TSHIAKA
TSIENIMBALALA
TSIMIHETY
TSONG
TSONGA
TUAT
TUKEN
TUKULOR
TULA
TULAMA
TUMBUKA
TUMBWE
TUMTUM
TUNDJUR
TUNGU
TUNISIAN
TUPURI
TURA
TURKA
TURKANA
TURUMBU
TUSIAN
TUTSI
TYEMBARA
TYENGA
UBANG
UHAMI-IYAYU
UJOGOMA
UKAAN
UKPE
UKPET
UKWUANI
ULED NAIL
UMON
UNGA
UREGU
URHOBO
UTUGWANG
UZEKWE
VAGALA
VAI
VEKHEE
VENDA
VERE
VEZO
VIDUNDA
VILI
VIN
VINZA
VIRA
VIYE
VUMA
VUTE
WAAMA
WADA
WAJA
WALA
WALLAGA
WAMBU
WANDYA
WARGLA
WARJI
WASA
WASI ALAWA
WE
WENYA
WILE
WILLE
WINIAMA
WINJI
WOLLO
WOLOF
WOM
WONGO
WOYO
XHOSA
YAA
YAELIMA
YAGOUTE
YAHI
YAKA
YAKA, SUKU
YAKOMA
YALIWASA
YAMAIE
YAMANDUNDU
YAMONGO
YANA
YANZI
YAO
YARSE
YASA
YASAMA
YEKE
YELA
YESKWA
YEYE
YIWOM
YOKO
YOMBE
YORUBA
YOWA
YUKUTARE
YUNGUR
ZAGHAWA
ZANAKI
ZANDE
ZANDE ABANDIA
ZANDE AVUNGARA
ZARAMO
ZARI
ZARMA
ZEEM
ZEKARA
ZELA
ZENAGA
ZIBAN
ZIGUA
ZILMANU
ZIMBA
ZINZU
ZOMBO
ZULA

**Table S4: Complete key over HIV-1 subtypes shown in Fig. 2D**

HIV-B and HIV-C are indicated by stars.

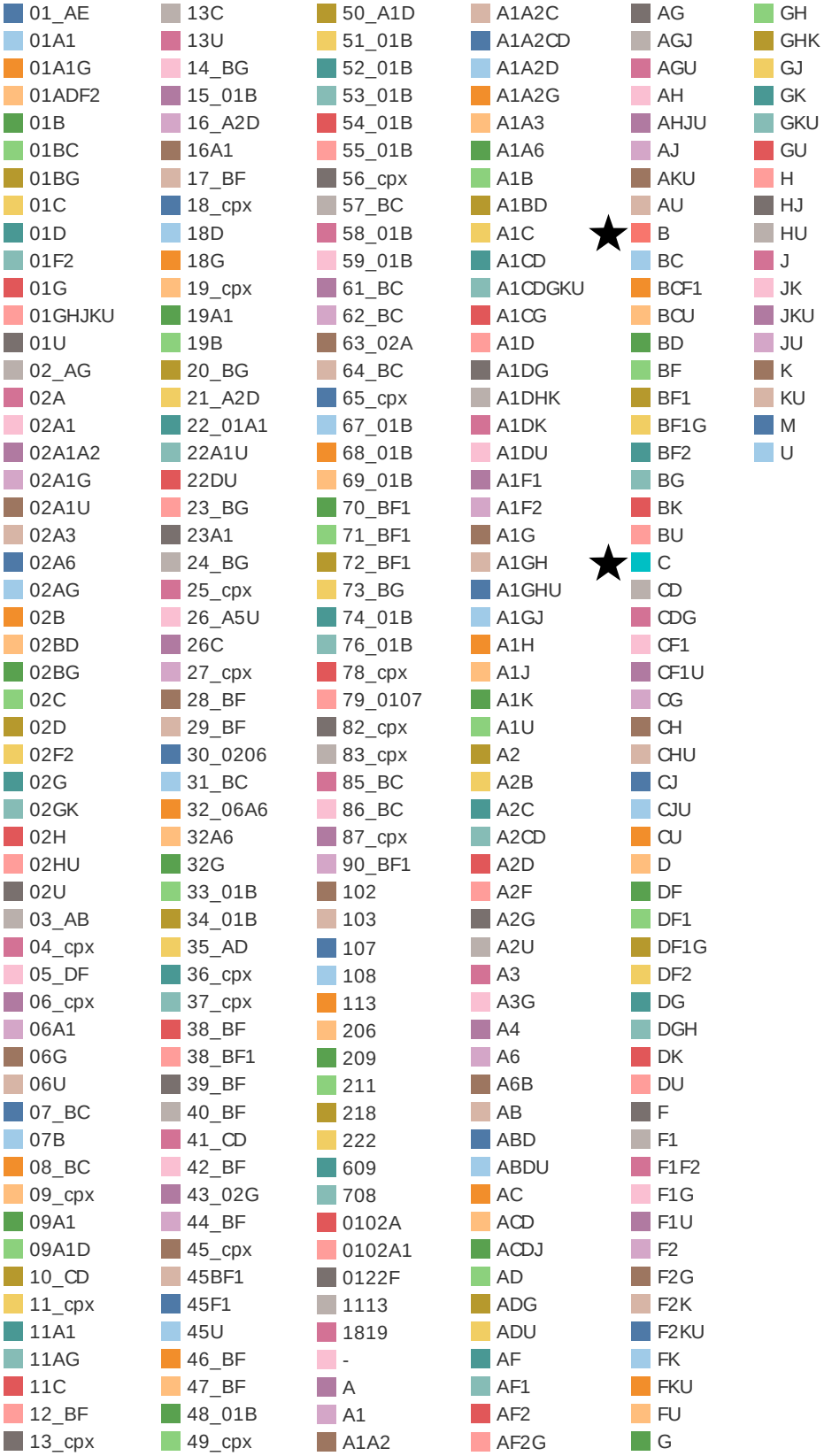


**Table S5: Complete key over African languages shown in Fig. 2E**

Bantu languages are indicated by stars.

**Table S6: HLA class I allele frequencies in worldwide populations**

| | ZA Blacks | ZA Zulu | US African | US Caribbean | Thailand | US Chinese | US Caucasian | UK central |
|---|---|---|---|---|---|---|---|---|
| **A*02:01** | 0.083 | 0 | 0.123 | 0.111 | 0.255 | 0.0946 | 0.2755 | 0.2697 |
| **B*07:02** | 0.046 | 0.07 | 0.073 | 0.073 | 0.038 | 0.0079 | 0.1306 | 0.1391 |
| **C*07:02** | 0.058 | 0.08 | 0.073 | 0.067 | 0 | 0.1945 | 0.1415 | 0.155 |
| **A*30:01** | 0.101 | 0.1 | 0.068 | 0.072 | 0.013 | 0.0274 | 0.013 | 0.011 |
| **B*42:01** | 0.089 | 0.11 | 0.053 | 0.053 | 0 | 0.0001 | 0.0003 | 0.0019 |
| **C*17:01** | 0.111 | 0.05 | 0.068 | 0.068 | 0 | 0.0005 | 0.0088 | 0.004 |
| **B*58:02** | 0.094 | 0.11 | 0.042 | 0.032 | | 0 | 0.0001 | 0 |
| **C*06:02** | 0.149 | 0.15 | 0.087 | 0.072 | 0 | 0.0447 | 0.0932 | 0.11 |
| **Population** | ZA Blacks | ZA Zulu | US African | US Caribbean | Thailand | US Chinese | US Caucasian | UK central |
| **Sum of HLA variants common in Caucasian** | 0.187 | 0.15 | 0.269 | 0.251 | 0.293 | 0.297 | 0.5476 | 0.5638 |
| **Sum of HLA variants common in Africans** | 0.544 | 0.52 | 0.318 | 0.297 | 0.013 | 0.0727 | 0.1154 | 0.1269 |

The individual HLA alleles frequencies of HLA variants that contributed most to the first two principal components (PCs) were also found in common African and Caucasian haplotypes. Here we show the individual frequencies of each HLA variant and the sum of the variants that can form haplotypes in Africans and Caucasians. The data derive from (Gonzalez-Galarza et al. 2015), and the South Africa (ZA) Zulu data derive from the Females Rising through Education, Support, and Health (FRESH) project.

**Table S7: Overview of the amino acid variation at SSP in HIV-1 subtype consensus sequences**

| HIV subtype consensus | Position 27 amino acid | Position 41 | Position 116 | Position 120 | Position 128 |
|---|---|---|---|---|---|
| HIV-A1 | I | S | G | G | D |
| HIV-A2 | V | T | G | S | E |
| HIV-B | V | S | G | N | E |
| HIV-C | I | T | A | S | D |
| HIV-D | I | S | G | S | E |
| HIV-F1* | I | S | Q | S | D |
| HIV-G | V | S | R | S | E |
| HIV-H* | V | S | A | G | D |
| HIV-K* | I | S | T | S | E |

(From www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html, (Foley et al. 2018))

*HIV-F1 predominates in South America, especially in Brazil, Uruguay, and Argentina, HIV-H is found in Central Africa, Eastern Europe, and Central Asia, and HIV-K is found in Central Africa and Pakistan (Khan et al. 2018) at low frequency.

The subtype distribution of HIV-F1, HIV-H and HIV-K outside of Africa – and the presence of, e.g., HIV-C in India and China - is due to founder effects. The evolution of these subtypes over time in non-African countries is unknown due to limited sampling and the brevity of the ongoing HIV epidemic (Buonaguro et al. 2007; Hemelaar et al. 2011; Osmanov et al. 2002). Overall, HIV-F (0.59%), HIV-H (0.17%), HIV-J (0.14%), and HIV-K (0.04%) together cause fewer than 1% of HIV-1 infections worldwide (Buonaguro et al. 2007; Hemelaar et al. 2011). No consensus sequences for HIV-J p24Gag are available from the Los Alamos HIV sequence database (Foley et al. 2018).

**Table S8: LANL HIV database patient IDs of patients with imputed four digit HLAs**

**B*27 → B*27:05:**

402, 3393, 3394, 10542, 11222, 13225, 15792, 19252, 19542, 19878, 22972, 22973, 22974, 22975, 22982, 22984, 22986, 23077, 27220, 34369, 36112, 36113, 49996

**B*35 → B*35:01:**

24029, 30977, 35886, 36113, 49994, 51971

**B*57 → B*57:01:**

10487, 10488, 10489, 10490, 10491, 10492, 10493, 10494, 10495, 10496, 10497, 10498, 10499, 10500, 10501, 10502, 10503, 10504, 10505, 10506, 10507, 10508, 10509, 10510, 11222, 13224, 15396, 15397, 19539

## References


Abi-Rached, L., et al. (2011), 'The shaping of modern human immune systems by multiregional admixture with archaic humans', *Science,* 334 (6052), 89-94.

Beloukas, A., et al. (2016), 'Molecular epidemiology of HIV-1 infection in Europe: An overview', *Infect Genet Evol,* 46, 180-89.

Buonaguro, L., Tornesello, M. L., and Buonaguro, F. M. (2007), 'Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications', *J Virol,* 81 (19), 10209-19.

Diez-Fuertes, F., Cabello, M., and Thomson, M. M. (2015), 'Bayesian phylogeographic analyses clarify the origin of the HIV-1 subtype A variant circulating in former Soviet Union's countries', *Infect Genet Evol,* 33, 197-205.

Draenert, R., et al. (2004), 'Immune selection for altered antigen processing leads to cytotoxic T lymphocyte escape in chronic HIV-1 infection', *J Exp Med,* 199 (7), 905-15.

Faroudi, M., et al. (2003), 'Lytic versus stimulatory synapse in cytotoxic T lymphocyte/target cell interaction: manifestation of a dual activation threshold', *Proc Natl Acad Sci U S A,* 100 (24), 14145-50.

Foley, Brian, et al. (2018), *HIV Sequence Compendium 2018* (Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, New Mexico 87545 U.S.A.LA-UR-18-25673).

Gonzalez-Galarza, F. F., et al. (2015), 'Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations', *Nucleic Acids Res,* 43 (Database issue), D784-8.

Goulder, P. J. and Watkins, D. I. (2008), 'Impact of MHC class I diversity on immune control of immunodeficiency virus replication', *Nat Rev Immunol,* 8 (8), 619-30.

Goulder, P. J. and Walker, B. D. (2012), 'HIV and HLA class I: an evolving relationship', *Immunity,* 37 (3), 426-40.

Gragert, L., et al. (2013), 'Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry', *Hum Immunol,* 74 (10), 1313-20.

Hall, H. I., et al. (2008), 'Estimation of HIV incidence in the United States', *JAMA,* 300 (5), 520-9.

Hemelaar, J., et al. (2011), 'Global trends in molecular epidemiology of HIV-1 during 2000-2007', *AIDS,* 25 (5), 679-89.

Khan, S., et al. (2018), 'HIV-1 genetic diversity, geographical linkages and antiretroviral drug resistance among individuals from Pakistan', *Arch Virol,* 163 (1), 33-40.

Kijak, G. H., et al. (2009), 'HLA class I allele and haplotype diversity in Ugandans supports the presence of a major east African genetic cluster', *Tissue Antigens,* 73 (3), 262-9.

Lucchiari-Hartz, M., et al. (2003), 'Differential proteasomal processing of hydrophobic and hydrophilic protein regions: contribution to cytotoxic T lymphocyte epitope clustering in HIV-1-Nef', *Proc Natl Acad Sci U S A,* 100 (13), 7755-60.

Osmanov, S., et al. (2002), 'Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000', *J Acquir Immune Defic Syndr,* 29 (2), 184-90.

Paximadis, M., et al. (2012), 'Human leukocyte antigen class I (A, B, C) and II (DRB1) diversity in the black and Caucasian South African population', *Hum Immunol,* 73 (1), 80-92.

Prejean, J., et al. (2011), 'Estimated HIV incidence in the United States, 2006-2009', *PLoS One,* 6 (8), e17502.

Tenzer, S., et al. (2014), 'HIV-1 adaptation to antigen processing results in population-level immune evasion and affects subtype diversification', *Cell Rep*, 7 (2), 448-63.
Tenzer, S., et al. (2009), 'Antigen processing influences HIV-specific cytotoxic T lymphocyte immunodominance', *Nat Immunol*, 10 (6), 636-46.

# Conclusion

The ever increasing amount of biological data, both freely available and in private databases, offer the possibility of answering previously intractable questions and having direct impacts on human health and wellbeing. However, processing the information presents its own unique and often frustrating set of challenges. Attempts to create data format standards more often than not leads to an additional standard being added to mix, rather than the hoped for consolidation. Machine learning and Artificial Intelligence are making some headway in database clean up, but the bulk of this work still falls to humans.

Elsewhere, human involvement also remains critical. Computers are very good at rapidly doing exactly what they have been programmed to do, but results still require interpretation by humans. In commercial and clinical settings, machine learning is making some headway into this as well, but this utilises vast datasets with well-defined and pre-delineated results. In research, where results are ideally by definition novel, humans are still required, and the formulation of sensible and useful research questions still remains a principally human endeavour. Being able to notice outliers as biologically interesting and follow up on them, rather than simply discarding them as noise, is one of the key reasons why we need biologically trained computational scientists.

Despite humanity's desire to neatly classify life along strictly linear branches of descent,

I have shown in Chapter One that attempts to make such phylogenies of virophages and polintons can be misleading. While some genes have (mostly) shared evolutionary histories, the fact that no single gene is shared by all members indicates that there is no such thing as a 'core' set to be examined. There is also likely to be extensive horizontal gene transfer HGT present in the virophage/polinton/polinton like elements (PLE) world, and taken together these indicate that single genes trees will likely be more informative. The extreme timescales involved have also erased much of the phylogenetic signal further compounding the problem. Finally, shedding light on a long standing question of how trans-species horizontal gene transfer of homing endonuclease genes (HEGs) occurs, I also found evidence suggesting that the HEGs are using viruses to hitchhike between genomes of disparate organisms.

When modern humans left Africa, they encountered archaic hominids such as Neanderthals and Denisovans that had been living in Eurasia for hundreds of thousands of years. The archaic hominids had substantial amounts of time to coevolve with and granularly adapt to the local pathogens across the length and breadth of Eurasia. As I showed in Chapter Two, introgression of archaic hominid HLA genes into our genomes has left an indelible mark still observable today. This is clinically relevant as modern humans have migrated at unprecedented levels since the age of exploration. Modern crises have displaced millions and we are already seeing small amounts of climate refugees displaced by rising sea levels and desertification. The next century is likely to see levels of human movement that are frankly, beyond comprehension. This movement

has shifted, and will continue to shift, pathogens along with us. As old pathogens move into new areas, non-preadapted populations may experience higher levels of morbidity and mortality than populations that have coevolved with these pathogens, e.g. the recent outbreak of Zika in the Americas. Furthermore, other human activities like deforestation and encroachment into natural habitats can result in the outbreak of novel pathogens, e.g. SARS-CoV-2. In even without coevolution, because of what we have adapted to before there can be differences in the way populations respond to these novel pathogens, e.g. recent studies have shown that a major genetic risk factor for respiratory failure after SARS-CoV-2 infection and a genomic region associated with protection from SARS-CoV-2 are both inherited from Neanderthals (Zeberg and Pääbo, 2021, 2020). Differences in the efficiency in which different HLA types sort and separate dispersed populations also suggests additional functions for these genes that may not be as immediately apparent, e.g. HLA C's involvement in human placentation.

Chapter Two showed hosts adapting to pathogens, while Chapter Three showed that this is a two-way street with pathogens also adapting to their hosts. In Chapter Three, I showed that the dominant paradigm (that different HIV subtypes are principally artefacts of selectively neutral processes such as founder effects) is unlikely to be the whole story. This has profound implications for HIV vaccine research, which has thus far been unsuccessful in delivering a broadly effective vaccine despite nearly 40 years of effort. It also has implications for other pathogen vaccines, particularly for other highly antigenically variable pathogens such as influenza. These findings also highlight the

importance of extensive metadata associated with pathogen sequences. Often pathogen sequences will be devoid of any metadata, and this is a lost opportunity for drawing connections via data mining. While it may be politically charged and uncomfortable, gathering information on host ethnicity, gender, etc. is of paramount importance for maximising the long term utility of sequence information.

As I have shown, host/pathogen coevolution occurs at many levels in biological systems, and examining it can open up new avenues of research. Many of these may be basic research without immediate obvious clinical benefits, but even aside from the goal of expanding the frontiers of human knowledge, these always have the potential to pay dividends down the line. James Clerk Maxwell's equations of electromagnetism underpin our current understanding of all forms of radiation, and make the entirety of modern telecommunications (and therefore debatably, modern life) possible. Despite this, his theories were somewhat slow to catch on:

> Meantime, it took physicists some decades to grasp the full significance of Maxwell's discovery, so bold was the leap that his genius forced upon the conceptions of his fellow-workers. Only after Hertz had demonstrated experimentally the existence of Maxwell's electromagnetic waves, did resistance to the new theory break down. (Einstein, 1940)

Similarly, albeit in all probability to a far lesser extent, the clear demonstration that linguistic diversity is correlated with genetic diversity in Sub-Saharan Africa may not have always had clear clinical implications at the time. Nevertheless, building upon such work, and the work of countless others, I hope to contribute in my own small way towards ending the HIV pandemic.

> "Bernard of Chartres used to compare us to dwarfs perched on the shoulders of giants. He pointed out that we see more and farther than our predecessors, not because we have keener vision or greater height, but because we are lifted up and borne aloft on their gigantic stature." (John of Salisbury, 1159; McGarry, 2009)

However, we not only see further because we are standing on the shoulders of giants, but also because we collectively are standing on a great pyramid of knowledge built of innumerable smaller blocks, and I am pleased to lay my own tiny block on this ever growing collective effort.

# Bibliography


Abebe, A., Pollakis, G., Fontanet, A.L., Fisseha, B., Tegbaru, B., Kliphuis, A., Tesfaye, G., Negassa, H., Cornelissen, M., Goudsmit, J., Rinke de Wit, T.F., 2000. Identification of a genetic subcluster of HIV type 1 subtype C (C') widespread in Ethiopia. AIDS Res. Hum. Retroviruses 16, 1909–1914. https://doi.org/10.1089/08892220050195865

Abecasis, A.B., Lemey, P., Vidal, N., de Oliveira, T., Peeters, M., Camacho, R., Shapiro, B., Rambaut, A., Vandamme, A.-M., 2007. Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form. J. Virol. 81, 8543–8551. https://doi.org/10.1128/JVI.00463-07

Abergel, C., Legendre, M., Claverie, J.-M., 2015. The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. FEMS Microbiology Reviews 39, 779–796. https://doi.org/10.1093/femsre/fuv037

Abi-Rached, L., Jobin, M.J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., Babrzadeh, F., Gharizadeh, B., Luo, M., Plummer, F.A., et al. (2011). The shaping of modern human immune systems by multiregional admixture with archaic humans. Science 334, 89-94.

Abram, M.E., Ferris, A.L., Shao, W., Alvord, W.G., Hughes, S.H., 2010. Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. J. Virol. 84, 9864–9878. https://doi.org/10.1128/JVI.00915-10

Adl, S.M., Simpson, A.G.B., Lane, C.E., Lukeš, J., Bass, D., Bowser, S.S., Brown, M.W., Burki, F., Dunthorn, M., Hampl, V., Heiss, A., Hoppenrath, M., Lara, E., le Gall, L., Lynn, D.H., McManus, H., Mitchell, E.A.D., Mozley-Stanridge, S.E., Parfrey, L.W., Pawlowski, J., Rueckert, S., Shadwick, L., Schoch, C.L., Smirnov, A., Spiegel, F.W., 2012. The Revised Classification of Eukaryotes. Journal of Eukaryotic Microbiology 59, 429–514. https://doi.org/10.1111/j.1550-7408.2012.00644.x

Adland, E., Hill, M., Lavandier, N., Csala, A., Edwards, A., Chen, F., Radkowski, M., Kowalska, J.D., Paraskevis, D., Hatzakis, A., Valenzuela-Ponce, H., Pfafferott, K., Williams, I., Pellegrino, P., Borrow, P., Mori, M., Rockstroh, J., Prado, J.G., Mothe, B., Dalmau, J., Martinez-Picado, J., Tudor-Williams, G., Frater, J., Stryhn, A., Buus, S., Reyes Teran, G., Mallal, S., John, M., Buchbinder, S., Kirk, G., Martin, J., Michael, N., Fellay, J., Deeks, S., Walker, B., Avila-Rios, S., Cole, D., Brander, C., Carrington, M., Goulder, P., 2017. Differential Immunodominance Hierarchy of CD8+ T-Cell Responses in HLA-B*27:05- and -B*27:02-Mediated Control of HIV-1 Infection. Journal of Virology JVI.01685-17. https://doi.org/10.1128/JVI.01685-17

Agoni, L., Golden, a., Guha, C., and Lenz, J. (2012). Neanderthal and Denisovan retroviruses. Current Biology 22, 437-438.

Aherfi, S., Colson, P., La Scola, B., Raoult, D., 2016. Giant Viruses of Amoebas: An

Update. Frontiers in Microbiology 7. https://doi.org/10.3389/fmicb.2016.00349
Ahmad, T., Neville, M., Marshall, S.E., Armuzzi, A., Mulcahy-Hawes, K., Crawshaw, J., Sato, H., Ling, K.-L., Barnardo, M., Goldthorpe, S., Walton, R., Bunce, M., Jewell, D.P., Welsh, K.I., 2003. Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC. Human Molecular Genetics 12, 647–656. https://doi.org/10.1093/hmg/ddg066
Aiewsakun, P., Katzourakis, A., 2016. Time-Dependent Rate Phenomenon in Viruses. Journal of Virology 90, 7184–7195. https://doi.org/10.1128/jvi.00593-16
Alberts, B. (Ed.), 2002. Molecular biology of the cell, 4th ed. ed. Garland Science, New York.
Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., Wacziarg, R., 2003. Fractionalization. J. Econ. Growth 8, 155–94.
Alizon, M., Wain-Hobson, S., Montagnier, L., Sonigo, P., 1986. Genetic variability of the AIDS virus: nucleotide sequence analysis of two isolates from African patients. Cell 46, 63–74. https://doi.org/10.1016/0092-8674(86)90860-3
Allen Dodds, J., Cole, A., 1980. Microscopy and biology ofUronema gigas, a filamentous eucaryotic green alga, and its associated tailed virus-like particle. Virology 100, 156–165. https://doi.org/10.1016/0042-6822(80)90561-9
Aman, S., Borggren, M., Fomsgaard, A., Kebe, K., Jaye, A., 2015. Optimization of full-length sequencing of HIV type 1 from Senegal. Unpublished direct submission to Genbank.
Amogne, W., Bontell, I., Grossmann, S., Aderaye, G., Lindquist, L., Sönnerborg, A., Neogi, U., 2016. Phylogenetic Analysis of Ethiopian HIV-1 Subtype C Near Full-Length Genomes Reveals High Intrasubtype Diversity and a Strong Geographical Cluster. AIDS Res. Hum. Retroviruses 32, 471–474. https://doi.org/10.1089/aid.2015.0380
Anderson, J.P., Rodrigo, A.G., Learn, G.H., Madan, A., Delahunty, C., Coon, M., Girard, M., Osmanov, S., Hood, L., Mullins, J.I., 2000. Testing the hypothesis of a recombinant origin of human immunodeficiency virus type 1 subtype E. J. Virol. 74, 10752–10765. https://doi.org/10.1128/jvi.74.22.10752-10765.2000
Antiretroviral Therapy Cohort Collaboration, 2008. Life expectancy of individuals on combination antiretroviral therapy in high-income countries: a collaborative analysis of 14 cohort studies. Lancet 372, 293–299. https://doi.org/10.1016/S0140-6736(08)61113-7
Archer, J., Robertson, D.L., 2007. Understanding the diversification of HIV-1 groups M and O: AIDS 21, 1693–1700. https://doi.org/10.1097/QAD.0b013e32825eabd0
Ariën, K.K., Abraha, A., Quiñones-Mateu, M.E., Kestens, L., Vanham, G., Arts, E.J., 2005. The replicative fitness of primary human immunodeficiency virus type 1 (HIV-1) group M, HIV-1 group O, and HIV-2 isolates. J Virol 79, 8979–8990. https://doi.org/10.1128/JVI.79.14.8979-8990.2005
Arrildt, K.T., Joseph, S.B., Swanstrom, R., 2012. The HIV-1 env protein: a coat of many colors. Curr. HIV/AIDS Rep. 9, 52–63. https://doi.org/10.1007/s11904-

011-0107-3
Arroyo, M.A., Hoelscher, M., Sanders-Buell, E., Herbinger, K.-H., Samky, E., Maboko, L., Hoffmann, O., Robb, M.R., Birx, D.L., McCutchan, F.E., 2004. HIV type 1 subtypes among blood donors in the Mbeya region of southwest Tanzania. AIDS Res. Hum. Retroviruses 20, 895–901. https://doi.org/10.1089/0889222041725235
Arroyo, M.A., Hoelscher, M., Sateren, W., Samky, E., Maboko, L., Hoffmann, O., Kijak, G., Robb, M., Birx, D.L., McCutchan, F.E., 2005. HIV-1 diversity and prevalence differ between urban and rural areas in the Mbeya region of Tanzania. AIDS 19, 1517–1524. https://doi.org/10.1097/01.aids.0000183515.14642.76
Baalwa, J., Wang, S., Parrish, N.F., Decker, J.M., Keele, B.F., Learn, G.H., Yue, L., Ruzagira, E., Ssemwanga, D., Kamali, A., Amornkul, P.N., Price, M.A., Kappes, J.C., Karita, E., Kaleebu, P., Sanders, E., Gilmour, J., Allen, S., Hunter, E., Montefiori, D.C., Haynes, B.F., Cormier, E., Hahn, B.H., Shaw, G.M., 2013. Molecular identification, cloning and characterization of transmitted/founder HIV-1 subtype A, D and A/D infectious molecular clones. Virology 436, 33–48. https://doi.org/10.1016/j.virol.2012.10.009
Baesi, K., Moallemi, S., Farrokhi, M., Alinaghi, S.A.S., Truong, H.-H.M., 2014. Subtype classification of Iranian HIV-1 sequences registered in the HIV databases, 2006-2013. PLoS ONE 9, e105098. https://doi.org/10.1371/journal.pone.0105098
Baeten, J.M., Chohan, B., Lavreys, L., Chohan, V., McClelland, R.S., Certain, L., Mandaliya, K., Jaoko, W., Overbaugh, J., 2007. HIV-1 subtype D infection is associated with faster disease progression than subtype A in spite of similar plasma HIV-1 loads. J Infect Dis 195, 1177–1180. https://doi.org/10.1086/512682
Baker, C.A.R., Bousheri, S., Ssewanyana, I., Jones, N.G., K'Aluoch, O., Baliruno, D., Ssali, F., Cao, H., 2007. HIV Subtypes Distribution and Implication for Antiretroviral Treatment in a Ugandan Population. J Int Assoc Physicians AIDS Care (Chic) 6, 260–263. https://doi.org/10.1177/1545109707303938
Baldrich-Rubio, E., Anagonou, S., Stirrups, K., Lafia, E., Candotti, D., Lee, H., Allain, J.P., 2001. A complex human immunodeficiency virus type 1 A/G/J recombinant virus isolated from a seronegative patient with AIDS from Benin, West Africa. J. Gen. Virol. 82, 1095–1106. https://doi.org/10.1099/0022-1317-82-5-1095
Baltimore, D., 1971. Expression of animal virus genomes. Bacteriol Rev 35, 235–241.
Bao, W., Kojima, K.K., Kohany, O., 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6. https://doi.org/10.1186/s13100-015-0041-9
Bapteste, E., Susko, E., Leigh, J., MacLeod, D., Charlebois, R.L., Doolittle, W.F., 2005. Do orthologous gene phylogenies really support tree-thinking? BMC Evol Biol 5, 33. https://doi.org/10.1186/1471-2148-5-33

Barré-Sinoussi, F., Chermann, J.C., Rey, F., Nugeyre, M.T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vézinet-Brun, F., Rouzioux, C., Rozenbaum, W., Montagnier, L., 1983. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science 220, 868–871. https://doi.org/10.1126/science.6189183

Bártolo, I., Calado, R., Borrego, P., Leitner, T., Taveira, N., 2016. Rare HIV-1 Subtype J Genomes and a New H/U/CRF02_AG Recombinant Genome Suggests an Ancient Origin of HIV-1 in Angola. AIDS Res. Hum. Retroviruses 32, 822–828. https://doi.org/10.1089/AID.2016.0084

Barton, N.H., 1979. Gene flow past a cline. Heredity 43, 333–339. https://doi.org/10.1038/hdy.1979.86

Barugahare, B., Baker, C., K'Aluoch, O., Donovan, R., Elrefaei, M., Eggena, M., Jones, N., Mutalya, S., Kityo, C., Mugyenyi, P., Cao, H., 2005. Human immunodeficiency virus-specific responses in adult Ugandans: patterns of cross-clade recognition. J. Virol. 79, 4132–4139. https://doi.org/10.1128/JVI.79.7.4132-4139.2005

Basu, V.P., Song, M., Gao, L., Rigby, S.T., Hanson, M.N., Bambara, R.A., 2008. Strand transfer events during HIV-1 reverse transcription. Virus Res. 134, 19–38. https://doi.org/10.1016/j.virusres.2007.12.017

Beijerinck, M.W., 1898. Über ein contagium vivum fluidum als Ursache der Fleckenkrankheit der Tabaksblätter. Verhandelingen Kon. Akademie van Wetenschappen, Amsterdam, Deel VI.

Bellas, C.M., Anesio, A.M., Barker, G., 2015. Analysis of virus genomes from glacial environments reveals novel virus groups with unusual host interactions. Frontiers in Microbiology 6. https://doi.org/10.3389/fmicb.2015.00656

Benamar, S., Reteno, D.G.I., Bandaly, V., Labas, N., Raoult, D., La Scola, B., 2016. Faustoviruses: Comparative Genomics of New Megavirales Family Members. Frontiers in Microbiology 7. https://doi.org/10.3389/fmicb.2016.00003

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2009. GenBank. Nucleic Acids Res. 37, D26-31. https://doi.org/10.1093/nar/gkn723

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L., 2008. GenBank. Nucleic Acids Res. 36, D25-30. https://doi.org/10.1093/nar/gkm929

Berg, M.G., Yamaguchi, J., Alessandri-Gradt, E., Tell, R.W., Plantier, J.-C., Brennan, C.A., 2016. A Pan-HIV Strategy for Complete Genome Sequencing. J. Clin. Microbiol. 54, 868–882. https://doi.org/10.1128/JCM.02479-15

Bergström, T.F., Josefsson, A., Erlich, H.A., Gyllensten, U., 1998. Recent origin of HLA-DRB1 alleles and implications for human evolution. Nat Genet 18, 237–242. https://doi.org/10.1038/ng0398-237

Berry, V., Gascuel, O., 1996. On the Interpretation of Bootstrap Trees: Appropriate Threshold of Clade Selection and Induced Gain. Molecular Biology and Evolution 13, 999–1011. https://doi.org/10.1093/molbev/13.7.999

Biek, R., Pybus, O.G., Lloyd-Smith, J.O., Didelot, X., 2015. Measurably evolving pathogens in the genomic era. Trends in Ecology & Evolution 30, 306–313.

https://doi.org/10.1016/j.tree.2015.03.009
Billings, E., Sanders-Buell, E., Bose, M., Bradfield, A., Lei, E., Kijak, G.H., Arroyo, M.A., Kibaya, R.M., Scott, P.T., Wasunna, M.K., Sawe, F.K., Shaffer, D.N., Birx, D.L., McCutchan, F.E., Michael, N.L., Robb, M.L., Kim, J.H., Tovanabutra, S., 2015. The Number and Complexity of Pure and Recombinant HIV-1 Strains Observed within Incident Infections during the HIV and Malaria Cohort Study Conducted in Kericho, Kenya, from 2003 to 2006. PLoS ONE 10, e0135124. https://doi.org/10.1371/journal.pone.0135124
Billings, E., Sanders-Buell, E., Bose, M., Kijak, G.H., Bradfield, A., Crossler, J., Arroyo, M.A., Maboko, L., Hoffmann, O., Geis, S., Birx, D.L., Kim, J.H., Michael, N.L., Robb, M.L., Hoelscher, M., Tovanabutra, S., 2017. HIV-1 Genetic Diversity Among Incident Infections in Mbeya, Tanzania. AIDS Res. Hum. Retroviruses 33, 373–381. https://doi.org/10.1089/AID.2016.0111
Birditt, B.A., Rousseau, C.M., Korber, B.T., Goulder, P., Brander, C., Kiepiela, P., Walker, B.D., Mullins, J.I., 2005. HIV HLA epitope mapping from Durban, South Africa. Unpublished direct submission to Genbank.
Bjorkman, P.J., Parham, P., 1990. Structure, function, and diversity of class I major histocompatibility complex molecules. Annu Rev Biochem 59, 253–288. https://doi.org/10.1146/annurev.bi.59.070190.001345
Bjorkman, P.J., Saper, M.A., Samraoui, B., Bennett, W.S., Strominger, J.L., Wiley, D.C., 1987a. The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. Nature 329, 512–518. https://doi.org/10.1038/329512a0
Bjorkman, P.J., Saper, M.A., Samraoui, B., Bennett, W.S., Strominger, J.L., Wiley, D.C., 1987b. Structure of the human class I histocompatibility antigen, HLA-A2. Nature 329, 506–512. https://doi.org/10.1038/329506a0
Blanc, G., Duncan, G., Agarkova, I., Borodovsky, M., Gurnon, J., Kuo, A., Lindquist, E., Lucas, S., Pangilinan, J., Polle, J., Salamov, A., Terry, A., Yamada, T., Dunigan, D.D., Grigoriev, I.V., Claverie, J.-M., Van Etten, J.L., 2010. The Chlorella variabilis NC64A Genome Reveals Adaptation to Photosymbiosis, Coevolution with Viruses, and Cryptic Sex. The Plant Cell 22, 2943–2955. https://doi.org/10.1105/tpc.110.076406
Blanc, G., Gallot-Lavallée, L., Maumus, F., 2015. Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses. Proceedings of the National Academy of Sciences 112, E5318–E5326. https://doi.org/10.1073/pnas.1506469112
Bodelle, P., Vallari, A., Coffey, R., McArthur, C.P., Beyeme, M., Devare, S.G., Schochetman, G., Brennan, C.A., 2004. Identification and genomic sequence of an HIV type 1 group N isolate from Cameroon. AIDS Res. Hum. Retroviruses 20, 902–908. https://doi.org/10.1089/0889222041725262
Bodmer, W.F. (1972). Evolutionary significance of the HL-A system. Nature 237, 139-145 passim.
Bonelli, M., Shih, H.-Y., Hirahara, K., Singelton, K., Laurence, A., Poholek, A., Hand,

T., Mikami, Y., Vahedi, G., Kanno, Y., O'Shea, J.J., 2014. Helper T cell plasticity: impact of extrinsic and intrinsic signals on transcriptomes and epigenomes. Curr Top Microbiol Immunol 381, 279–326. https://doi.org/10.1007/82_2014_371

Borcard, D., Legendre, P., 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. Ecological Modelling 153, 51–68. https://doi.org/10.1016/S0304-3800(01)00501-4

Borcard, D., Legendre, P., Avois-Jacquet, C., Tuomisto, H., 2004. Dissecting the spatial structure of ecological data at multiple scales. Ecology 85, 1826–1832. https://doi.org/10.1890/03-3111

Bostoen, K., Clist, B., Doumenge, C., Grollemund, R., Hombert, J.-M., Muluwa, J.K., Maley, J., 2015. Middle to Late Holocene Paleoclimatic Change and the Early Bantu Expansion in the Rain Forests of Western Central Africa. Curr. Anthropol. 56, 354–384. https://doi.org/10.1086/681436

Boutwell, C.L., Rowley, C.F., Essex, M., 2009. Reduced viral replication capacity of human immunodeficiency virus type 1 subtype C caused by cytotoxic-T-lymphocyte escape mutations in HLA-B57 epitopes of capsid protein. J. Virol. 83, 2460–2468. https://doi.org/10.1128/JVI.01970-08

Box, G.E.P., 1979. Robustness in the Strategy of Scientific Model Building, in: Robustness in Statistics. Elsevier, pp. 201–236. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

Bradwell, K., Combe, M., Domingo-Calap, P., Sanjuán, R., 2013. Correlation Between Mutation Rate and Genome Size in Riboviruses: Mutation Rate of Bacteriophage $Q\beta$ . Genetics 195, 243–251. https://doi.org/10.1534/genetics.113.154963

Brady, A., Salzberg, S.L., 2009. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nature Methods 6, 673–676. https://doi.org/10.1038/nmeth.1358

Brave, A., Ljungberg, K., Wahren, B., 2002. HIV-1 92UG031 p17p24gag. Unpublished direct submission to Genbank.

Bredell, H., Martin, D.P., Van Harmelen, J., Varsani, A., Sheppard, H.W., Donovan, R., Gray, C.M., HIVNET028 Study Team, Williamson, C., 2007. HIV type 1 subtype C gag and nef diversity in Southern Africa. AIDS Res. Hum. Retroviruses 23, 477–481. https://doi.org/10.1089/aid.2006.0232

Brennan, C.A., Bodelle, P., Coffey, R., Devare, S.G., Golden, A., Hackett, J., Harris, B., Holzmayer, V., Luk, K.-C., Schochetman, G., Swanson, P., Yamaguchi, J., Vallari, A., Ndembi, N., Ngansop, C., Makamche, F., Mbanya, D., Gürtler, L.G., Zekeng, L., Kaptué, L., 2008. The prevalence of diverse HIV-1 strains was stable in Cameroonian blood donors from 1996 to 2004. J. Acquir. Immune Defic. Syndr. 49, 432–439. https://doi.org/10.1097/QAI.0b013e31818a6561

Brown, B.K., Darden, J.M., Tovanabutra, S., Oblander, T., Frost, J., Sanders-Buell, E., de Souza, M.S., Birx, D.L., McCutchan, F.E., Polonis, V.R., 2005. Biologic and

genetic characterization of a panel of 60 human immunodeficiency virus type 1 isolates, representing clades A, B, C, D, CRF01_AE, and CRF02_AG, for the development and assessment of candidate vaccines. J. Virol. 79, 6089–6101. https://doi.org/10.1128/JVI.79.10.6089-6101.2005

Brown, J.H., Jardetzky, T.S., Gorga, J.C., Stern, L.J., Urban, R.G., Strominger, J.L., Wiley, D.C., 1993. Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. Nature 364, 33–39. https://doi.org/10.1038/364033a0

Brum, J.R., Ignacio-Espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., Chaffron, S., Cruaud, C., de Vargas, C., Gasol, J.M., Gorsky, G., Gregory, A.C., Guidi, L., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Poulos, B.T., Schwenck, S.M., Speich, S., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Tara Oceans Coordinators, Bork, P., Bowler, C., Sunagawa, S., Wincker, P., Karsenti, E., Sullivan, M.B., 2015. Ocean plankton. Patterns and ecological drivers of ocean viral communities. Science 348, 1261498. https://doi.org/10.1126/science.1261498

Bruno, W.J., Socci, N.D., Halpern, A.L., 2000. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. Mol Biol Evol 17, 189–197. https://doi.org/10.1093/oxfordjournals.molbev.a026231

Buhler, S., Nunes, J.M., Nicoloso, G., Tiercy, J.M., and Sanchez-Mazas, A. (2012). The heterogeneous HLA genetic makeup of the Swiss population. PloS one 7, e41400.

Buonaguro, L., Tornesello, M.L., Buonaguro, F.M., 2007. Human Immunodeficiency Virus Type 1 Subtype Distribution in the Worldwide Epidemic: Pathogenetic and Therapeutic Implications. Journal of Virology 81, 10209–10219. https://doi.org/10.1128/JVI.00872-07

Burchell, A.N., Raboud, J., Donelle, J., Loutfy, M.R., Rourke, S.B., Rogers, T., Rosenes, R., Liddy, C., Kendall, C.E., 2019. Cause-specific mortality among HIV-infected people in Ontario, 1995-2014: a population-based retrospective cohort study. CMAJ Open 7, E1–E7. https://doi.org/10.9778/cmajo.20180159

Bürger, R., 2017. Two-locus clines on the real line with a step environment. Theoretical Population Biology 117, 1–22. https://doi.org/10.1016/j.tpb.2017.08.002

Burke, D., 1997. Recombination in HIV: An Important Viral Evolutionary Strategy. Emerging Infectious Diseases 3, 253–259. https://doi.org/10.3201/eid0303.970301

Burki, F., Shalchian-Tabrizi, K., Minge, M., Skjæveland, Å., Nikolaev, S.I., Jakobsen, K.S., Pawlowski, J., 2007. Phylogenomics Reshuffles the Eukaryotic Supergroups. PLoS ONE 2, e790. https://doi.org/10.1371/journal.pone.0000790

Burnie, J., Guzzo, C., 2019. The Incorporation of Host Proteins into the External HIV-1 Envelope. Viruses 11. https://doi.org/10.3390/v11010085

Burns, C.C., Gleason, L.M., Mozaffarian, A., Giachetti, C., Carr, J.K., Overbaugh, J.,

2002. Sequence variability of the integrase protein from a diverse collection of HIV type 1 isolates representing several subtypes. AIDS Res. Hum. Retroviruses 18, 1031–1041. https://doi.org/10.1089/08892220260235399

Camin, J.H., Sokal, R.R., 1965. A Method for Deducing Branching Sequences in Phylogeny. Evolution 19, 311. https://doi.org/10.2307/2406441

Campbell, M.S., Mullins, J.I., Hughes, J.P., Celum, C., Wong, K.G., Raugi, D.N., Sorensen, S., Stoddard, J.N., Zhao, H., Deng, W., Kahle, E., Panteleeff, D., Baeten, J.M., McCutchan, F.E., Albert, J., Leitner, T., Wald, A., Corey, L., Lingappa, J.R., Partners in Prevention HSV/HIV Transmission Study Team, 2011. Viral linkage in HIV-1 seroconverters and their partners in an HIV-1 prevention clinical trial. PLoS ONE 6, e16986. https://doi.org/10.1371/journal.pone.0016986

Campbell, R.D., Trowsdale, J., 1993. Map of the human MHC. Immunology Today 14, 349–352. https://doi.org/10.1016/0167-5699(93)90234-C

Campbell-Yesufu, O.T., Gandhi, R.T., 2011. Update on human immunodeficiency virus (HIV)-2 infection. Clin. Infect. Dis. 52, 780–787. https://doi.org/10.1093/cid/ciq248

Candotti, D., Adu-Sarkodie, Y., Davies, F., Baldrich-Rubio, E., Stirrups, K., Lee, H., Allain, J.P., 2000. AIDS in an HIV-seronegative Ghanaian woman with intersubtype A/G recombinant HIV-1 infection. J. Med. Virol. 62, 1–8. https://doi.org/10.1002/1096-9071(200009)62:1<1::aid-jmv1>3.0.co;2-3

Canty, A., Ripley, B.D., 2020. boot: Bootstrap R (S-Plus) Functions.

Cao, K., Moormann, A.M., Lyke, K.E., Masaberg, C., Sumba, O.P., Doumbo, O.K., Koech, D., Lancaster, A., Nelson, M., Meyer, D., Single, R., Hartzman, R.J., Plowe, C.V., Kazura, J., Mann, D.L., Sztein, M.B., Thomson, G., Fernández-Viña, M.A., 2004. Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci. Tissue Antigens 63, 293–325.

Cappy, P., De Oliveira, F., Gueudin, M., Alessandri-Gradt, E., Plantier, J.-C., 2016. A Multiplex PCR Approach for Detecting Dual Infections and Recombinants Involving Major HIV Variants. J. Clin. Microbiol. 54, 1282–1288. https://doi.org/10.1128/JCM.03222-15

Carlson, J.M., Du, V.Y., Pfeifer, N., Bansal, A., Tan, V.Y.F., Power, K., Brumme, C.J., Kreimer, A., DeZiel, C.E., Fusi, N., Schaefer, M., Brockman, M.A., Gilmour, J., Price, M.A., Kilembe, W., Haubrich, R., John, M., Mallal, S., Shapiro, R., Frater, J., Harrigan, P.R., Ndung'u, T., Allen, S., Heckerman, D., Sidney, J., Allen, T.M., Goulder, P.J.R., Brumme, Z.L., Hunter, E., Goepfert, P.A., 2016. Impact of pre-adapted HIV transmission. Nat. Med. 22, 606–613. https://doi.org/10.1038/nm.4100

Carlson, J.M., Schaefer, M., Monaco, D.C., Batorsky, R., Claiborne, D.T., Prince, J., Deymier, M.J., Ende, Z.S., Klatt, N.R., DeZiel, C.E., Lin, T.-H., Peng, J., Seese, A.M., Shapiro, R., Frater, J., Ndung'u, T., Tang, J., Goepfert, P., Gilmour, J., Price, M.A., Kilembe, W., Heckerman, D., Goulder, P.J.R., Allen, T.M., Allen,

S., Hunter, E., 2014. HIV transmission. Selection bias at the heterosexual HIV-1 transmission bottleneck. Science 345, 1254031. https://doi.org/10.1126/science.1254031

Carr, J.K., Kim, B., Sanders-Buell, E., Salminen, M.O., Alaeus, A., Albert, J.A., Birx, D.L., McCutchan, F.E., 1998. HIV-1 isolate VI354 from Gabon, complete genome. Unpublished direct submission to Genbank.

Carr, J.K., Lar, P., Bemis, K., Jelpe, J., Ayuba, L., Zella, D., Blattner, W., Kanki, P., Abimiku, A.G., 2005. HIV Subtype and Drug Resistance Patterns among Drug Naive Persons in Jos, Nigeria. Unpublished direct submission to Genbank.

Carr, J.K., Salminen, M.O., Albert, J., Sanders-Buell, E., Gotte, D., Birx, D.L., McCutchan, F.E., 1998. Full genome sequences of human immunodeficiency virus type 1 subtypes G and A/G intersubtype recombinants. Virology 247, 22–31. https://doi.org/10.1006/viro.1998.9211

Carr, J.K., Torimiro, J.N., Wolfe, N.D., Eitel, M.N., Kim, B., Sanders-Buell, E., Jagodzinski, L.L., Gotte, D., Burke, D.S., Birx, D.L., McCutchan, F.E., 2001. The AG recombinant IbNG and novel strains of group M HIV-1 are common in Cameroon. Virology 286, 168–181. https://doi.org/10.1006/viro.2001.0976

Carr, J.K., Torimiro, J.N., Wolfe, N.D., Tamoufe, U., Mpoudi-Ngole, E., Kim, B., Sanders-Buell, E., Eyzaguirre, L.M., Jagodzinski, L.L., Gotte, D., Birx, D.L., McCutchan, F.E., Burke, D.S., 2006. High Genetic Complexity of HIV-1 in low-prevalence, rural regions of Cameroon. Unpublished direct submission to Genbank.

Carr, J.K., Wolfe, N.D., Torimiro, J.N., Tamoufe, U., Mpoudi-Ngole, E., Eyzaguirre, L., Birx, D.L., McCutchan, F.E., Burke, D.S., 2010. HIV-1 recombinants with multiple parental strains in low-prevalence, remote regions of Cameroon: evolutionary relics? Retrovirology 7, 39. https://doi.org/10.1186/1742-4690-7-39

Carr, J.K., Zhao, J., Tang, S., Viswanath, R., Awazi, B., Hewlett,I., 2009. Identification of a novel CRF22_01A1 recombinant form in Cameroon. Unpublished direct submission to Genbank.

Carter, B.J., Fife, K.H., de la Maza, L.M., Berns, K.I., 1976. Genome localization of adeno-associated virus RNA. J. Virol. 19, 1044–1053.

Carter, B.J., Khoury, G., Rose, J.A., 1972. Adenovirus-associated virus multiplication. IX. Extent of transcription of the viral genome in vivo. J. Virol. 10, 1118–1125.

Carter, B.J., Rose, J.A., 1972. Adenovirus-associated virus multiplication. 8. Analysis of in vivo transcription induced by complete or partial helper viruses. J. Virol. 10, 9–16.

Casado, G., Thomson, M.M., Sierra, M., Nájera, R., 2005. Identification of a novel HIV-1 circulating ADG intersubtype recombinant form (CRF19_cpx) in Cuba. J. Acquir. Immune Defic. Syndr. 40, 532–537. https://doi.org/10.1097/01.qai.0000186363.27587.c0

Cashdan, E. (2014). Biogeography of human infectious diseases: a global historical

analysis. PloS one 9, e106752.
Cavalli-Sforza, L.L., Edwards, A.W., 1967. Phylogenetic analysis. Models and estimation procedures. Am J Hum Genet 19, 233–257.
Cech, T.R., 2012. The RNA worlds in context. Cold Spring Harb Perspect Biol 4, a006742. https://doi.org/10.1101/cshperspect.a006742
Center for Geographic Analysis, 2013. Ethnicity Felix 2001 [WWW Document]. URL https://worldmap.harvard.edu/africamap/
Cham, F., Heyndrickx, L., Janssens, W., Van der Auwera, G., Vereecken, K., De Houwer, K., Coppens, S., Whittle, H., van der Groen, G., 2000. Study of HIV type 1 gag/env variability in The Gambia, using a multiplex DNA polymerase chain reaction. AIDS Res. Hum. Retroviruses 16, 1915–1919. https://doi.org/10.1089/08892220050195874
Chang, C.-H., Kist, N.C., Stuart Chester, T.L., Sreenu, V.B., Herman, M., Luo, M., Lunn, D., Bell, J., Plummer, F.A., Ball, T.B., Katzourakis, A., Iversen, A.K.N., 2015. HIV-infected sex workers with beneficial HLA-variants are potential hubs for selection of HIV-1 recombinants that may affect disease progression. Sci Rep 5, 11253. https://doi.org/10.1038/srep11253
Chapman, S.J., Hill, A.V.S., 2012. Human genetic susceptibility to infectious disease. Nat. Rev. Genet. 13, 175–188. https://doi.org/10.1038/nrg3114
Charon, J., Grigg, M.J., Eden, J.-S., Piera, K.A., William, T., Rose, K., Davenport, M.P., Anstey, N.M., Holmes, E.C., 2019. Matryoshka RNA virus 1: a novel RNA virus associated with parasites in human malaria. bioRxiv. https://doi.org/10.1101/624403
Charurat, M., Nasidi, A., Delaney, K., Saidu, A., Croxton, T., Mondal, P., Aliyu, G.G., Constantine, N., Abimiku, A., Carr, J.K., Vertefeuille, J., Blattner, W., 2012. Characterization of acute HIV-1 infection in high-risk Nigerian populations. J. Infect. Dis. 205, 1239–1247. https://doi.org/10.1093/infdis/jis103
Chen, J., Foroozesh, M., and Qin, Z. (2019). Transactivation of human endogenous retroviruses by tumor viruses and their functions in virus-associated malignancies. Oncogenesis 8, 6.
Chen, R.X.F., 2018. A Brief Introduction to Shannon's Information Theory. ArXiv161209316v2 Cs Math.
Cherkasova, E., Scrivani, C., Doh, S., Weisman, Q., Takahashi, Y., Harashima, N., Yokoyama, H., Srinivasan, R., Linehan, W.M., Lerman, M.I., et al. (2016). Detection of an Immunogenic HERV-E Envelope with Selective Expression in Clear Cell Kidney Cancer. Cancer Res 76, 2177-2185.
Cherkasova, E., Weisman, Q., and Childs, R.W. (2013). Endogenous retroviruses as targets for antitumor immunity in renal cell cancer and other tumors. Front Oncol 3, 243.
Chicz, R.M., Urban, R.G., Lane, W.S., Gorga, J.C., Stern, L.J., Vignali, D.A.A., Strominger, J.L., 1992. Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in

size. Nature 358, 764–768. https://doi.org/10.1038/358764a0
Choge, I., Cilliers, T., Walker, P., Taylor, N., Phoswa, M., Meyers, T., Viljoen, J., Violari, A., Gray, G., Moore, P.L., Papathanosopoulos, M., Morris, L., 2006. Genotypic and phenotypic characterization of viral isolates from HIV-1 subtype C-infected children with slow and rapid disease progression. AIDS Res. Hum. Retroviruses 22, 458–465. https://doi.org/10.1089/aid.2006.22.458
Choi, D.J., Dube, S., Spicer, T.P., Slade, H.B., Jensen, F.C., Poiesz, B.J., 1997. HIV type 1 isolate Z321, the strain used to make a therapeutic HIV type 1 immunogen, is intersubtype recombinant. AIDS Res. Hum. Retroviruses 13, 357–361. https://doi.org/10.1089/aid.1997.13.357
Choo, S.Y., 2007. The HLA system: genetics, immunology, clinical testing, and clinical implications. Yonsei Med J 48, 11–23. https://doi.org/10.3349/ymj.2007.48.1.11
Chopera, D.R., Mann, J.K., Mwimanzi, P., Omarjee, S., Kuang, X.T., Ndabambi, N., Goodier, S., Martin, E., Naranbhai, V., Karim, S.A., Karim, Q.A., Brumme, Z.L., Ndung'u, T., Williamson, C., Brockman, M.A., CAPRISA 004 TRAPS Team, 2013. No evidence for selection of HIV-1 with enhanced gag-protease or Nef function among breakthrough infections in the CAPRISA 004 tenofovir microbicide trial. PLoS ONE 8, e71758. https://doi.org/10.1371/journal.pone.0071758
Chopera, D.R., Mlotshwa, M., Woodman, Z., Mlisana, K., de Assis Rosa, D., Martin, D.P., Abdool Karim, S., Gray, C.M., Williamson, C., CAPRISA 002 Study Team, 2011. Virological and immunological factors associated with HIV-1 differential disease progression in HLA-B 58:01-positive individuals. J. Virol. 85, 7070–7080. https://doi.org/10.1128/JVI.02543-10
Chopera, D.R., Woodman, Z., Mlisana, K., Mlotshwa, M., Martin, D.P., Seoighe, C., Treurnicht, F., de Rosa, D.A., Hide, W., Karim, S.A., Gray, C.M., Williamson, C., CAPRISA 002 Study Team, 2008. Transmission of HIV-1 CTL escape variants provides HLA-mismatched recipients with a survival advantage. PLoS Pathog. 4, e1000033. https://doi.org/10.1371/journal.ppat.1000033
Chuong, E.B., Elde, N.C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science 351, 1083-1087.
Cilliers, T., Nhlapo, J., Coetzer, M., Orlovic, D., Ketas, T., Olson, W.C., Moore, J.P., Trkola, A., Morris, L., 2003. The CCR5 and CXCR4 coreceptors are both used by human immunodeficiency virus type 1 primary isolates from subtype C. J. Virol. 77, 4449–4456. https://doi.org/10.1128/jvi.77.7.4449-4456.2003
Cilliers, T., Patience, T., Pillay, C., Papathanasopoulos, M., Morris, L., 2004. Sensitivity of HIV type 1 subtype C isolates to the entry inhibitor T-20. AIDS Res. Hum. Retroviruses 20, 477–482. https://doi.org/10.1089/088922204323087714
Claiborne, D.T., Prince, J.L., Scully, E., Macharia, G., Micci, L., Lawson, B., Kopycinski, J., Deymier, M.J., Vanderford, T.H., Nganou-Makamdop, K.,

Ende, Z., Brooks, K., Tang, J., Yu, T., Lakhi, S., Kilembe, W., Silvestri, G., Douek, D., Goepfert, P.A., Price, M.A., Allen, S.A., Paiardini, M., Altfeld, M., Gilmour, J., Hunter, E., 2015. Replicative fitness of transmitted HIV-1 drives acute immune activation, proviral load in memory CD4+ T cells, and disease progression. Proc. Natl. Acad. Sci. U.S.A. 112, E1480-1489. https://doi.org/10.1073/pnas.1421607112

Clarke, K.R., and WarWick, R.M. (2001). Change in marine communities: an approach to statistical analysis and interpretation (Plymouth, United Kingdom).

Claverie, J.-M., Abergel, C., 2013. Open Questions About Giant Viruses. Advances in Virus Research 25–56. https://doi.org/10.1016/b978-0-12-408116-1.00002-1

Coetzer, M., Cilliers, T., Papathanasopoulos, M., Ramjee, G., Karim, S.A., Williamson, C., Morris, L., 2007. Longitudinal analysis of HIV type 1 subtype C envelope sequences from South Africa. AIDS Res. Hum. Retroviruses 23, 316–321. https://doi.org/10.1089/aid.2006.0207

Colson, P., De Lamballerie, X., Yutin, N., Asgari, S., Bigot, Y., Bideshi, D.K., Cheng, X.-W., Federici, B.A., Van Etten, J.L., Koonin, E.V., La Scola, B., Raoult, D., 2013. "Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. Archives of Virology 158, 2517–2521. https://doi.org/10.1007/s00705-013-1768-6

Colson, P., Pagnier, I., Yoosuf, N., Fournous, G., La Scola, B., Raoult, D., 2012. "Marseilleviridae", a new family of giant viruses infecting amoebae. Archives of Virology 158, 915–920. https://doi.org/10.1007/s00705-012-1537-y

Congo Basin Art History Research Center, Meur, C., Felix, M.L., Tribal Arts S.P.R.L, 2001. Peoples of Africa: ethno-linguistic map.

Cornelissen, M., van Den Burg, R., Zorgdrager, F., Goudsmit, J., 2000. Spread of distinct human immunodeficiency virus type 1 AG recombinant lineages in Africa. J. Gen. Virol. 81, 515–523. https://doi.org/10.1099/0022-1317-81-2-515

Correa, A.M.S., Ainsworth, T.D., Rosales, S.M., Thurber, A.R., Butler, C.R., Vega Thurber, R.L., 2016. Viral Outbreak in Corals Associated with an In Situ Bleaching Event: Atypical Herpes-Like Viruses and a New Megavirus Infecting Symbiodinium. Frontiers in Microbiology 7. https://doi.org/10.3389/fmicb.2016.00127

Correia, B.E., Bates, J.T., Loomis, R.J., Baneyx, G., Carrico, C., Jardine, J.G., Rupert, P., Correnti, C., Kalyuzhniy, O., Vittal, V., Connell, M.J., Stevens, E., Schroeter, A., Chen, M., Macpherson, S., Serra, A.M., Adachi, Y., Holmes, M.A., Li, Y., Klevit, R.E., Graham, B.S., Wyatt, R.T., Baker, D., Strong, R.K., Crowe, J.E., Johnson, P.R., Schief, W.R., 2014. Proof of principle for epitope-focused vaccine design. Nature 507, 201–206. https://doi.org/10.1038/nature12966

Craig, N.L., 1995. Unity in Transposition Reactions. Science 270, 253–253. https://doi.org/10.1126/science.270.5234.253

Crawford, H., Matthews, P.C., Schaefer, M., Carlson, J.M., Leslie, A., Kilembe, W.,

Allen, S., Ndung'u, T., Heckerman, D., Hunter, E., Goulder, P.J.R., 2011. The hypervariable HIV-1 capsid protein residues comprise HLA-driven CD8+ T-cell escape mutations and covarying HLA-independent polymorphisms. J. Virol. 85, 1384–1390. https://doi.org/10.1128/JVI.01879-10

Crawford, L.V., Follett, E.A.C., Burdon, M.G., McGeoch, D.J., 1969. The DNA of a Minute Virus of Mice. Journal of General Virology 4, 37–46. https://doi.org/10.1099/0022-1317-4-1-37

Cresswell, P., 1994. Assembly, Transport, and Function of MHC Class II Molecules. Annual Review of Immunology 12, 259–291. https://doi.org/10.1146/annurev.iy.12.040194.001355

Cribari-Neto, F., Zeileis, A., 2010. Beta Regression in R. J. Stat. Soft. 34. https://doi.org/10.18637/jss.v034.i02

Cromer, D., Schlub, T.E., Smyth, R.P., Grimm, A.J., Chopra, A., Mallal, S., Davenport, M.P., Mak, J., 2016. HIV-1 Mutation and Recombination Rates Are Different in Macrophages and T-cells. Viruses 8, 118. https://doi.org/10.3390/v8040118

Crotty, S., 2011. Follicular Helper CD4 T Cells (T FH ). Annual Review of Immunology 29, 621–663. https://doi.org/10.1146/annurev-immunol-031210-101400

Crux, N.B., Elahi, S., 2017. Human Leukocyte Antigen (HLA) and Immune Regulation: How Do Classical and Non-Classical HLA Alleles Modulate Immune Response to Human Immunodeficiency Virus and Hepatitis C Virus Infections? Frontiers in Immunology 8. https://doi.org/10.3389/fimmu.2017.00832

Currier, J.R., Harris, M.E., Cox, J.H., McCutchan, F.E., Birx, D.L., Maayan, S., Ferrari, G., 2005. Immunodominance and cross-reactivity of B5703-restricted CD8 T lymphocytes from HIV type 1 subtype C-infected Ethiopians. AIDS Res. Hum. Retroviruses 21, 239–245. https://doi.org/10.1089/aid.2005.21.239

Currier, J.R., Visawapoka, U., Tovanabutra, S., Mason, C.J., Birx, D.L., McCutchan, F.E., Cox, J.H., 2006. CTL epitope distribution patterns in the Gag and Nef proteins of HIV-1 from subtype A infected subjects in Kenya: use of multiple peptide sets increases the detectable breadth of the CTL response. BMC Immunol. 7, 8. https://doi.org/10.1186/1471-2172-7-8

D'Arrigo, R., Ciccozzi, M., Gori, C., Montieri, S., Aquaro, S., Bellagamba, R., Boumis, E., Di Perri, G., Pizzi, D., Antinori, A., Rezza, G., Perno, C.F., 2007. gp41 sequence variability in HIV type 1 non-B subtypes infected patients undergoing enfuvirtide pressure. AIDS Res. Hum. Retroviruses 23, 1296–1302. https://doi.org/10.1089/aid.2007.0095

Daar, A.S., Fuggle, S.V., Fabre, J.W., Ting, A., Morris, P.J., 1984. The detailed distribution of HLA-A, B, C antigens in normal human organs. Transplantation 38, 287–292. https://doi.org/10.1097/00007890-198409000-00018

Dahirel, V., Shekhar, K., Pereyra, F., Miura, T., Artyomov, M., Talsania, S., Allen,

T.M., Altfeld, M., Carrington, M., Irvine, D.J., Walker, B.D., Chakraborty, A.K., 2011. Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. Proc. Natl. Acad. Sci. U. S. A. 108, 11530–11535. https://doi.org/10.1073/pnas.1105315108

Darwin, C., 1809-1882, 1859. On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life. London : John Murray, 1859.

Dausset, J., 1958. [Iso-leuko-antibodies]. Acta Haematol. 20, 156–166. https://doi.org/10.1159/000205478

Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and Their Applications. Cambridge University Press, Cambridge.

Dayhoff, M.O., Eck, R.V., Chang, M.A., Sochard, M.R., 1965. Atlas of Protein Sequence and Structure 1965. National Biomedical Research Foundation.

De Mareuil, J., Salaun, D., Chermann, J.C., Hirsch, I., 1995. Fusogenic determinants of highly cytopathic subtype D Zairian isolate HIV-1 NDK. Virology 209, 649–653. https://doi.org/10.1006/viro.1995.1298

De Oliveira, F., Mourez, T., Vessiere, A., Ngoupo, P.-A., Alessandri-Gradt, E., Simon, F., Rousset, D., Plantier, J.-C., 2017. Multiple HIV-1/M + HIV-1/O dual infections and new HIV-1/MO inter-group recombinant forms detected in Cameroon. Retrovirology 14, 1. https://doi.org/10.1186/s12977-016-0324-3

De Oliveira,F., Cappy,P., Lemee,V., 2016. HIV-1/MO recombinants detection in France: impact on diagnosis and therapeutic management. Unpublished direct submission to Genbank.

de Silva, T.I., Turner, R., Hué, S., Trikha, R., van Tienen, C., Onyango, C., Jaye, A., Foley, B., Whittle, H., Rowland-Jones, S.L., Cotten, M., 2010. HIV-1 subtype distribution in the Gambia and the significant presence of CRF49_cpx, a novel circulating recombinant form. Retrovirology 7, 82. https://doi.org/10.1186/1742-4690-7-82

Delaroque, N., Maier, I., Knippers, R., Müller, D.G., 1999. Persistent virus integration into the genome of its algal host, Ectocarpus siliculosus (Phaeophyceae). J. Gen. Virol. 80 ( Pt 6), 1367–1370. https://doi.org/10.1099/0022-1317-80-6-1367

Dendrou, C.A., Petersen, J., Rossjohn, J., Fugger, L., 2018. HLA variation and disease. Nature Reviews Immunology 18, 325–339. https://doi.org/10.1038/nri.2017.143

Denison, M.R., Graham, R.L., Donaldson, E.F., Eckerle, L.D., Baric, R.S., 2011. Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. RNA Biol 8, 270–279. https://doi.org/10.4161/rna.8.2.15013

Desnues, C., La Scola, B., Yutin, N., Fournous, G., Robert, C., Azza, S., Jardot, P., Monteil, S., Campocasso, A., Koonin, E.V., Raoult, D., 2012. Provirophages and transpovirons as the diverse mobilome of giant viruses. Proceedings of the National Academy of Sciences 109, 18078–18083. https://doi.org/10.1073/pnas.1208835109

Desnues, C., Raoult, D., 2012. Virophages question the existence of satellites. Nature

Reviews Microbiology 10, 234–234. https://doi.org/10.1038/nrmicro2676-c3
Deymier, M.J., Ende, Z., Fenton-May, A.E., Dilernia, D.A., Kilembe, W., Allen, S.A., Borrow, P., Hunter, E., 2015. Heterosexual Transmission of Subtype C HIV-1 Selects Consensus-Like Variants without Increased Replicative Capacity or Interferon-$\alpha$ Resistance. PLoS Pathog. 11, e1005154. https://doi.org/10.1371/journal.ppat.1005154
Diemer, G.S., Stedman, K.M., 2012. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. Biology Direct 7, 13. https://doi.org/10.1186/1745-6150-7-13
Diniz-Filho, J.A.F., Soares, T.N., Lima, J.S., Dobrovolski, R., Landeiro, V.L., de Campos Telles, M.P., Rangel, T.F., Bini, L.M., 2013. Mantel test in population genetics. Genet Mol Biol 36, 475–485. https://doi.org/10.1590/S1415-47572013000400002
Diop-Ndiaye, H., Toure-Kane, C., Leye, N., Ngom-Gueye, N.F., Montavon, C., Peeters, M., Mboup, S., 2010. Antiretroviral drug resistance mutations in antiretroviral-naive patients from Senegal. AIDS Res. Hum. Retroviruses 26, 1133–1138. https://doi.org/10.1089/aid.2009.0295
Dobzhansky, T., 1964. Biology, molecular and organismic. Am Zool 4, 443–452. https://doi.org/10.1093/icb/4.4.443
Doherty, P.C., and Zinkernagel, R.M. (1975). Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. Nature 256, 50-52.
Dohrmann, M., Wörheide, G., 2017. Dating early animal evolution using phylogenomic data. Scientific Reports 7. https://doi.org/10.1038/s41598-017-03791-w
Douma, J.C., Weedon, J.T., 2019. Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. Methods Ecol Evol 10, 1412–1430. https://doi.org/10.1111/2041-210X.13234
Dowling, W.E., Kim, B., Mason, C.J., Wasunna, K.M., Alam, U., Elson, L., Birx, D.L., Robb, M.L., McCutchan, F.E., Carr, J.K., 2002. Forty-one near full-length HIV-1 sequences from Kenya reveal an epidemic of subtype A and A-containing recombinants. AIDS 16, 1809–1820. https://doi.org/10.1097/00002030-200209060-00015
Downey, R.F., Sullivan, F.J., Wang-Johanning, F., Ambs, S., Giles, F.J., and Glynn, S.A. (2015). Human endogenous retrovirus K and cancer: Innocent bystander or tumorigenic accomplice? Int J Cancer 137, 1249-1257.
Drake, J.W., 1991. A constant rate of spontaneous mutation in DNA-based microbes. Proceedings of the National Academy of Sciences 88, 7160–7164. https://doi.org/10.1073/pnas.88.16.7160
Dray, S., Bauman, D., Blanchet, G., Borcard, D., Clappe, S., Guenard, G., Jombart, T., Larocque, G., Legendre, P., Madi, N., Wagner, H.H., 2021. adespatial: Multivariate Multiscale Spatial Analysis.
Dray, S., Legendre, P., Peres-Neto, P.R., 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM).

Ecological Modelling 196, 483–493.
https://doi.org/10.1016/j.ecolmodel.2006.02.015
Duffy, S., Shackelton, L.A., Holmes, E.C., 2008. Rates of evolutionary change in viruses: patterns and determinants. Nature Reviews Genetics 9, 267–276.
https://doi.org/10.1038/nrg2323
Duke-Sylvester, S.M., Biek, R., Real, L.A., 2013. Molecular evolutionary signatures reveal the role of host ecological dynamics in viral disease emergence and spread. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 368, 20120194.
https://doi.org/10.1098/rstb.2012.0194
Dupuy, C., Periquet, G., Serbielle, C., Bézier, A., Louis, F., Drezen, J.-M., 2011. Transfer of a chromosomal Maverick to endogenous bracovirus in a parasitoid wasp. Genetica 139, 489–496. https://doi.org/10.1007/s10709-011-9569-x
Edwards, A.W., Cavalli-Sforza, L.L., 1965. A method for cluster analysis. Biometrics 21, 362–375.
Edwards, R.A., Rohwer, F., 2005. Viral metagenomics. Nature Reviews Microbiology 3, 504–510. https://doi.org/10.1038/nrmicro1163
Edwards, S.V., 2009. Is a new and general theory of molecular systematics emerging? Evolution 63, 1–19. https://doi.org/10.1111/j.1558-5646.2008.00549.x
Efron, B., Halloran, E., Holmes, S., 1996. Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci U S A 93, 13429–13434.
https://doi.org/10.1073/pnas.93.23.13429
Einstein, A., 1940. Considerations Concerning The Fundaments Of Theoretical PhysicS. Science 91, 487–492. https://doi.org/10.1126/science.91.2369.487
Elde, N.C., Child, S.J., Eickbush, M.T., Kitzman, J.O., Rogers, K.S., Shendure, J., Geballe, A.P., Malik, H.S., 2012. Poxviruses Deploy Genomic Accordions to Adapt Rapidly against Host Antiviral Defenses. Cell 150, 831–841.
https://doi.org/10.1016/j.cell.2012.05.049
Enard, D., and Petrov, D.A. (2018). Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. Cell 175, 360-371 e313.
Enard, D., Cai, L., Gwennap, C., and Petrov, D.A. (2016). Viruses are a dominant driver of protein adaptation in mammals. Elife 5.
Engelbrecht, S., de Villiers, T., Sampson, C.C., zur Megede, J., Barnett, S.W., van Rensburg, E.J., 2001. Genetic analysis of the complete gag and env genes of HIV type 1 subtype C primary isolates from South Africa. AIDS Res. Hum. Retroviruses 17, 1533–1547. https://doi.org/10.1089/08892220152644241
Epperson, B.K., 2005. Estimating dispersal from short distance spatial autocorrelation. Heredity 95, 7–15. https://doi.org/10.1038/sj.hdy.6800680
Epperson, B.K., Li, T., 1996. Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics. Proceedings of the National Academy of Sciences 93, 10528–10532.
https://doi.org/10.1073/pnas.93.19.10528
Esbjörnsson, J., Månsson, F., Kvist, A., Isberg, P.-E., Nowroozalizadeh, S., Biague,

A.J., da Silva, Z.J., Jansson, M., Fenyö, E.M., Norrgren, H., Medstrand, P., 2012. Inhibition of HIV-1 disease progression by contemporaneous HIV-2 infection. N. Engl. J. Med. 367, 224–232. https://doi.org/10.1056/NEJMoa1113244

Esparza, J., 2013. A brief history of the global effort to develop a preventive HIV vaccine. Vaccine 31, 3502–3518. https://doi.org/10.1016/j.vaccine.2013.05.018

Excoffier, L., Harding, R.M., Sokal, R.R., Pellegrini, B., Sanchez-Mazas, A., 1991. Spatial differentiation of RH and GM haplotype frequencies in Sub-Saharan Africa and its relation to linguistic affinities. Hum. Biol. 63, 273–307.

Excoffier, L., Pellegrini, B., Sanchez-Mazas, A., Simon, C., Langaney, A., 1987. Genetics and history of sub-Saharan Africa. Am. J. Phys. Anthropol. 30, 151–194. https://doi.org/10.1002/ajpa.1330300510

Eyerich, S., Eyerich, K., Pennino, D., Carbone, T., Nasorri, F., Pallotta, S., Cianfarani, F., Odorisio, T., Traidl-Hoffmann, C., Behrendt, H., Durham, S.R., Schmidt-Weber, C.B., Cavani, A., 2009. Th22 cells represent a distinct human T cell subset involved in epidermal immunity and remodeling. J Clin Invest 119, 3573–3585. https://doi.org/10.1172/JCI40202

Falk, K., Rötzschke, O., Stevanović, S., Jung, G., Rammensee, H.-G., 1991. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature 351, 290–296. https://doi.org/10.1038/351290a0

Fang, G., Kuiken, C., Weiser, B., Rowland-Jones, S., Plummer, F., Chen, C.-H., Kaul, R., Anzala, A.O., Bwayo, J., Kimani, J., Philpott, S.M., Kitchen, C., Sinsheimer, J.S., Gaschen, B., Lang, D., Shi, B., Kemal, K.S., Rostron, T., Brunner, C., Beddows, S., Sattenau, Q., Paxinos, E., Oyugi, J., Burger, H., 2004a. Long-term survivors in Nairobi: complete HIV-1 RNA sequences and immunogenetic associations. J. Infect. Dis. 190, 697–701. https://doi.org/10.1086/421504

Fang, G., Weiser, B., Kuiken, C., Philpott, S.M., Rowland-Jones, S., Plummer, F., Kimani, J., Shi, B., Kaul, R., Bwayo, J., Anzala, O., Burger, H., 2004b. Recombination following superinfection by HIV-1. AIDS 18, 153–159. https://doi.org/10.1097/00002030-200401230-00003

Faria, N.R., Rambaut, A., Suchard, M.A., Baele, G., Bedford, T., Ward, M.J., Tatem, A.J., Sousa, J.D., Arinaminpathy, N., Pépin, J., Posada, D., Peeters, M., Pybus, O.G., Lemey, P., 2014. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. Science 346, 56–61. https://doi.org/10.1126/science.1256739

Faria, N.R., Vidal, N., Lourenco, J., Raghwani, J., Sigaloff, K.C.E., Tatem, A.J., van de Vijver, D.A.M., Pineda-Peña, A.-C., Rose, R., Wallis, C.L., Ahuka-Mundeke, S., Muyembe-Tamfum, J.-J., Muwonga, J., Suchard, M.A., Rinke de Wit, T.F., Hamers, R.L., Ndembi, N., Baele, G., Peeters, M., Pybus, O.G., Lemey, P., Dellicour, S., 2019. Distinct rates and patterns of spread of the major HIV-1 subtypes in Central and East Africa. PLOS Pathogens 15, e1007976. https://doi.org/10.1371/journal.ppat.1007976

Farris, J.S., Kluge, A.G., Eckardt, M.J., 1970. A Numerical Approach to Phylogenetic Systematics. Systematic Zoology 19, 172. https://doi.org/10.2307/2412452

Fauci, A.S., Folkers, G.K., Marston, H.D., 2014a. Ending the global HIV/AIDS pandemic: the critical role of an HIV vaccine. Clin. Infect. Dis. 59 Suppl 2, S80-84. https://doi.org/10.1093/cid/ciu420

Fauci, A.S., Marston, H.D., Folkers, G.K., 2014b. An HIV cure: feasibility, discovery, and implementation. JAMA 312, 335–336. https://doi.org/10.1001/jama.2014.4754

Fearon, J.D., 2003. Ethnic and Cultural Diversity by Country. J. Econ. Growth 8, 195–222.

Fehren-Schmitz, L., Jarman, C.L., Harkins, K.M., Kayser, M., Popp, B.N., and Skoglund, P. (2017). Genetic Ancestry of Rapanui before and after European Contact. Curr Biol 27, 3209-3215 e3206.

Felsenstein, J., 1978. Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. Systematic Zoology 27, 401. https://doi.org/10.2307/2412923

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution 17, 368–376. https://doi.org/10.1007/BF01734359

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791. https://doi.org/10.1111/j.1558-5646.1985.tb00420.x

Felsenstein, J., 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Mass.

Felsenstein, J., Kishino, H., 1993. Is there Something Wrong with the Bootstrap on Phylogenies? A Reply to Hillis and Bull. Systematic Biology 42, 193–200. https://doi.org/10.1093/sysbio/42.2.193

Felsenstein, J., Yokoyama, S., 1976. The evolutionary advantage of recombination. II. Individual selection for recombination. Genetics 83, 845–859.

Ferdinandy, B., Mones, E., Vicsek, T., Müller, V., 2015. HIV Competition Dynamics over Sexual Networks: First Comer Advantage Conserves Founder Effects. PLOS Comput. Biol. 11, e1004093. https://doi.org/10.1371/journal.pcbi.1004093

Ferrari, S., Cribari-Neto, F., 2004. Beta Regression for Modelling Rates and Proportions. Journal of Applied Statistics 31, 799–815. https://doi.org/10.1080/0266476042000214501

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G., Ysebaert, M., 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature 260, 500–507. https://doi.org/10.1038/260500a0

Filée, J., 2015. Genomic comparison of closely related Giant Viruses supports an accordion-like model of evolution. Front Microbiol 6, 593. https://doi.org/10.3389/fmicb.2015.00593

Firquet, S., Beaujard, S., Lobert, P.-E., Sané, F., Caloone, D., Izard, D., Hober, D., 2015. Survival of Enveloped and Non-Enveloped Viruses on Inanimate Surfaces. Microbes Environ 30, 140–144. https://doi.org/10.1264/jsme2.ME14145

Fischer, M.G., 2011. Sputnik and Mavirus: more than just satellite viruses. Nature Reviews Microbiology 10, 78–78. https://doi.org/10.1038/nrmicro2676-c1

Fischer, M.G., Allen, M.J., Wilson, W.H., Suttle, C.A., 2010. Giant virus with a remarkable complement of genes infects marine zooplankton. Proceedings of the National Academy of Sciences 107, 19508–19513. https://doi.org/10.1073/pnas.1007615107

Fischer, M.G., Hackl, T., 2016. Host genome integration and giant virus-induced reactivation of the virophage mavirus. Nature 540, 288–291. https://doi.org/10.1038/nature20593

Fischer, M.G., Suttle, C.A., 2011. A Virophage at the Origin of Large DNA Transposons. Science 332, 231–234. https://doi.org/10.1126/science.1199412

Fischetti, L., Opare-Sem, O., Candotti, D., Lee, H., Allain, J.-P., 2004. Higher viral load may explain the dominance of CRF02_AG in the molecular epidemiology of HIV in Ghana. AIDS 18, 1208–1210. https://doi.org/10.1097/00002030-200405210-00017

Fisher, R.A., 1912. On an absolute criterion for fitting frequency curves. Messenger of Mathematics 41, 155–160.

Fisher, R.A., 1922. On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 222, 309–368. https://doi.org/10.1098/rsta.1922.0009

Fitch, W.M., 1986a. The estimate of total nucleotide substitutions from pairwise differences is biased. Philosophical Transactions of the Royal Society of London. B, Biological Sciences 312, 317–324. https://doi.org/10.1098/rstb.1986.0010

Fitch, W.M., 1986b. An estimation of the number of invariable sites is necessary for the accurate estimation of the number of nucleotide substitutions since a common ancestor. Prog Clin Biol Res 218, 149–159.

Fitch, W.M., Margoliash, E., 1967. Construction of phylogenetic trees. Science 155, 279–284. https://doi.org/10.1126/science.155.3760.279

Fox, J., Weisberg, S., 2019. An R Companion to Applied Regression, Third. ed. Sage, Thousand Oaks CA.

Frahm, N., Nickle, D.C., Linde, C.H., Cohen, D.E., Zuñiga, R., Lucchetti, A., Roach, T., Walker, B.D., Allen, T.M., Korber, B.T., Mullins, J.I., Brander, C., 2008. Increased detection of HIV-specific T cell responses by combination of central sequences with comparable immunogenicity. AIDS 22, 447–456. https://doi.org/10.1097/QAD.0b013e3282f42412

Friedlaender, J.S., Friedlaender, F.R., Reed, F.A., Kidd, K.K., Kidd, J.R., Chambers, G.K., Lea, R.A., Loo, J.H., Koki, G., Hodgson, J.A., et al. (2008). The genetic

structure of Pacific Islanders. PLoS genetics 4, e19.
Fu, Q., Hajdinjak, M., Moldovan, O.T., Constantin, S., Mallick, S., Skoglund, P., Patterson, N., Rohland, N., Lazaridis, I., Nickel, B., et al. (2015). An early modern human from Romania with a recent Neanderthal ancestor. Nature 524, 216-219.
Fuhrman, J.A., 1999. Marine viruses and their biogeochemical and ecological effects. Nature 399, 541–548. https://doi.org/10.1038/21119
Furstenau, T.N., Cartwright, R.A., 2016. The effect of the dispersal kernel on isolation-by-distance in a continuous population. PeerJ 4, e1848. https://doi.org/10.7717/peerj.1848
Futuyma, D.J., 2005. Evolution. Sinauer Associates, Sunderland, Mass.
Gaia, M., Benamar, S., Boughalmi, M., Pagnier, I., Croce, O., Colson, P., Raoult, D., La Scola, B., 2014. Zamilon, a Novel Virophage with Mimiviridae Host Specificity. PLoS ONE 9, e94923. https://doi.org/10.1371/journal.pone.0094923
Gaia, M., Pagnier, I., Campocasso, A., Fournous, G., Raoult, D., La Scola, B., 2013. Broad Spectrum of Mimiviridae Virophage Allows Its Isolation Using a Mimivirus Reporter. PLoS ONE 8, e61912. https://doi.org/10.1371/journal.pone.0061912
Gallo, R.C., Montagnier, L., 2003. The Discovery of HIV as the Cause of AIDS. New England Journal of Medicine 349, 2283–2285. https://doi.org/10.1056/NEJMp038194
Gao, C., Ren, X., Mason, A.S., Liu, H., Xiao, M., Li, J., Fu, D., 2013. Horizontal gene transfer in plants. Functional & Integrative Genomics 14, 23–29. https://doi.org/10.1007/s10142-013-0345-0
Gao, F., 2002. Complete genomes of HIV-1 subtype C viruses in South Africa. Unpublished direct submission to Genbank.
Gao, F., Robertson, D.L., Carruthers, C.D., Li, Y., Bailes, E., Kostrikis, L.G., Salminen, M.O., Bibollet-Ruche, F., Peeters, M., Ho, D.D., Shaw, G.M., Sharp, P.M., Hahn, B.H., 1998a. An isolate of human immunodeficiency virus type 1 originally classified as subtype I represents a complex mosaic comprising three different group M subtypes (A, G, and I). J. Virol. 72, 10234–10241.
Gao, F., Robertson, D.L., Carruthers, C.D., Morrison, S.G., Jian, B., Chen, Y., Barré-Sinoussi, F., Girard, M., Srinivasan, A., Abimiku, A.G., Shaw, G.M., Sharp, P.M., Hahn, B.H., 1998b. A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. J. Virol. 72, 5680–5698.
Gao, F., Robertson, D.L., Morrison, S.G., Hui, H., Craig, S., Decker, J., Fultz, P.N., Girard, M., Shaw, G.M., Hahn, B.H., Sharp, P.M., 1996. The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. J. Virol. 70, 7013–7029.
Gao, F., Trask, S.A., Hui, H., Mamaeva, O., Chen, Y., Theodore, T.S., Foley, B.T., Korber, B.T., Shaw, G.M., Hahn, B.H., 2001a. Molecular characterization of a

highly divergent HIV type 1 isolate obtained early in the AIDS epidemic from the Democratic Republic of Congo. AIDS Res. Hum. Retroviruses 17, 1217–1222. https://doi.org/10.1089/088922201316912826

Gao, F., Vidal, N., Li, Y., Trask, S.A., Chen, Y., Kostrikis, L.G., Ho, D.D., Kim, J., Oh, M.D., Choe, K., Salminen, M., Robertson, D.L., Shaw, G.M., Hahn, B.H., Peeters, M., 2001b. Evidence of two distinct subsubtypes within the HIV-1 subtype A radiation. AIDS Res. Hum. Retroviruses 17, 675–688. https://doi.org/10.1089/088922201750236951

Gao, F., Yue, L., Craig, S., Thornton, C.L., Robertson, D.L., McCutchan, F.E., Bradac, J.A., Sharp, P.M., Hahn, B.H., 1994. Genetic variation of HIV type 1 in four World Health Organization-sponsored vaccine evaluation sites: generation of functional envelope (glycoprotein 160) clones representative of sequence subtypes A, B, C, and E. WHO Network for HIV Isolation and Characterization. AIDS Res. Hum. Retroviruses 10, 1359–1368. https://doi.org/10.1089/aid.1994.10.1359

Garcia-Knight, M.A., Slyker, J., Payne, B.L., Pond, S.L.K., de Silva, T.I., Chohan, B., Khasimwa, B., Mbori-Ngacha, D., John-Stewart, G., Rowland-Jones, S.L., Esbjörnsson, J., 2016. Viral Evolution and Cytotoxic T Cell Restricted Selection in Acute Infant HIV-1 Infection. Sci Rep 6, 29536. https://doi.org/10.1038/srep29536

Garrison, K.E., Jones, R.B., Meiklejohn, D.A., Anwar, N., Ndhlovu, L.C., Chapman, J.M., Erickson, A.L., Agrawal, A., Spotts, G., Hecht, F.M., Rakoff-Nahoum, S., Lenz, J., Ostrowski, M.A., Nixon, D.F., 2007. T cell responses to human endogenous retroviruses in HIV-1 infection. PLoS Pathog 3, e165. https://doi.org/10.1371/journal.ppat.0030165

Gascuel, O., 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Molecular Biology and Evolution 14, 685–695. https://doi.org/10.1093/oxfordjournals.molbev.a025808

Gauss, C.F., 1816. Bestimmung der Genauigkeit der Beobachtungen. Z. Astronom. Verwandte Wiss. 1, 185–196.

Geginat, J., Paroni, M., Maglie, S., Alfen, J.S., Kastirr, I., Gruarin, P., De Simone, M., Pagani, M., Abrignani, S., 2014. Plasticity of Human CD4 T Cell Subsets. Frontiers in Immunology 5. https://doi.org/10.3389/fimmu.2014.00630

Geldmacher, C., Metzler, I.S., Tovanabutra, S., Asher, T.E., Gostick, E., Ambrozak, D.R., Petrovas, C., Schuetz, A., Ngwenyama, N., Kijak, G., Maboko, L., Hoelscher, M., McCutchan, F., Price, D.A., Douek, D.C., Koup, R.A., 2009. Minor viral and host genetic polymorphisms can dramatically impact the biologic outcome of an epitope-specific CD8 T-cell response. Blood 114, 1553–1562. https://doi.org/10.1182/blood-2009-02-206193

Gelfand, A.E., Smith, A.F.M., 1990. Sampling-Based Approaches to Calculating Marginal Densities. Journal of the American Statistical Association 85, 398. https://doi.org/10.2307/2289776

Genbank, 2019. GenBank and WGS Statistics [WWW Document]. GenBank. URL

https://www.ncbi.nlm.nih.gov/genbank/statistics/
Gerhardt, M., Mloka, D., Tovanabutra, S., Sanders-Buell, E., Hoffmann, O., Maboko, L., Mmbando, D., Birx, D.L., McCutchan, F.E., Hoelscher, M., 2005. In-depth, longitudinal analysis of viral quasispecies from an individual triply infected with late-stage human immunodeficiency virus type 1, using a multiple PCR primer approach. J. Virol. 79, 8249–8261. https://doi.org/10.1128/JVI.79.13.8249-8261.2005
Giandhari, J., Basson, A.E., Coovadia, A., Kuhn, L., Abrams, E.J., Strehlau, R., Morris, L., Hunt, G.M., 2015. Genetic Changes in HIV-1 Gag-Protease Associated with Protease Inhibitor-Based Therapy Failure in Pediatric Patients. AIDS Res. Hum. Retroviruses 31, 776–782. https://doi.org/10.1089/AID.2014.0349
Gilbert, M.T.P., Rambaut, A., Wlasiuk, G., Spira, T.J., Pitchenik, A.E., Worobey, M., 2007. The emergence of HIV/AIDS in the Americas and beyond. Proc. Natl. Acad. Sci. 104, 18566–18570. https://doi.org/10.1073/pnas.0705329104
Girard, P.R., 2005. Paradise Lost. Palgrave Macmillan US, New York. https://doi.org/10.1057/9781403980311
Goldman, N., 1993. Statistical tests of models of DNA substitution. Journal of Molecular Evolution 36, 182–198. https://doi.org/10.1007/BF00166252
Gong, C., Zhang, W., Zhou, X., Wang, H., Sun, G., Xiao, J., Pan, Y., Yan, S., Wang, Y., 2016. Novel Virophages Discovered in a Freshwater Lake in China. Frontiers in Microbiology 7. https://doi.org/10.3389/fmicb.2016.00005
Gonzalez, I., Déjean, S., Martin, P., and Baccini, A. (2008). An R Package to Extend Canonical Correlation Analysis. Journal of Statistical Software 23.
Gonzalez-Galarza, F.F., Takeshita, L.Y., Santos, E.J., Kempson, F., Maia, M.H., da Silva, A.L., Teles e Silva, A.L., Ghattaoraya, G.S., Alfirevic, A., Jones, A.R., et al. (2015). Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. Nucleic acids research 43, D784-788.
González-Galarza, F.F., Takeshita, L.Y.C., Santos, E.J.M., Kempson, F., Maia, M.H.T., da Silva, A.L.S., Teles e Silva, A.L., Ghattaoraya, G.S., Alfirevic, A., Jones, A.R., Middleton, D., 2015. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. Nucleic Acids Res. 43, D784-788. https://doi.org/10.1093/nar/gku1166
Goulder, P.J., and Walker, B.D. (2012). HIV and HLA class I: an evolving relationship. Immunity 37, 426-440.
Gounder, K., Oyaro, M., Padayachi, N., Zulu, T.M., de Oliveira, T., Wylie, J., Ndung'u, T., 2017. Complex Subtype Diversity of HIV-1 Among Drug Users in Major Kenyan Cities. AIDS Res. Hum. Retroviruses 33, 500–510. https://doi.org/10.1089/AID.2016.0321
Gounder, K., Padayachi, N., Mann, J.K., Radebe, M., Mokgoro, M., van der Stok, M., Mkhize, L., Mncube, Z., Jaggernath, M., Reddy, T., Walker, B.D., Ndung'u,

T., 2015. High frequency of transmitted HIV-1 Gag HLA class I-driven immune escape variants but minimal immune selection over the first year of clade C infection. PLoS ONE 10, e0119886. https://doi.org/10.1371/journal.pone.0119886

Gowing, M., 1993. Large virus-like particles from vacuoles of phaeodarian radiolarians and from other marine samples. Marine Ecology Progress Series 101, 33–43. https://doi.org/10.3354/meps101033

Grandi, N., and Tramontano, E. (2018). Human Endogenous Retroviruses Are Ancient Acquired Elements Still Shaping Innate Immune Responses. Frontiers in immunology 9, 2039.

Grant, B.J., Rodrigues, A.P.C., ElSawy, K.M., McCammon, J.A., Caves, L.S.D., 2006. Bio3d: an R package for the comparative analysis of protein structures. Bioinforma. Oxf. Engl. 22, 2695–2696. https://doi.org/10.1093/bioinformatics/btl461

Grant, H.E., Hodcroft, E.B., Ssemwanga, D., Kitayimbwa, J.M., Yebra, G., Esquivel Gomez, L.R., Frampton, D., Gall, A., Kellam, P., de Oliveira, T., Bbosa, N., Nsubuga, R.N., Kibengo, F., Kwan, T.H., Lycett, S., Kao, R., Robertson, D.L., Ratmann, O., Fraser, C., Pillay, D., Kaleebu, P., Leigh Brown, A.J., 2020. Pervasive and non-random recombination in near full-length HIV genomes from Uganda. Virus Evol 6, veaa004. https://doi.org/10.1093/ve/veaa004

Gray, R.D., Drummond, A.J., Greenhill, S.J., 2009. Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement. Science 323, 479–483. https://doi.org/10.1126/science.1166858

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., Hansen, N.F., Durand, E.Y., Malaspinas, A.-S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L.F., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., Pääbo, S., 2010. A draft sequence of the Neandertal genome. Science 328, 710–722. https://doi.org/10.1126/science.1188021

Grenfell, B.T., Pybus, O.G., Gog, J.R., Wood, J.L.N., Daly, J.M., Mumford, J.A., Holmes, E.C., 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303, 327–332. https://doi.org/10.1126/science.1090727

Grisson, R.D., Chenine, A.-L., Yeh, L.-Y., He, J., Wood, C., Bhat, G.J., Xu, W., Kankasa, C., Ruprecht, R.M., 2004. Infectious molecular clone of a recently transmitted pediatric human immunodeficiency virus clade C isolate from Africa: evidence of intraclade recombination. J. Virol. 78, 14066–14069. https://doi.org/10.1128/JVI.78.24.14066-14069.2004

Gu, X., Li, W.-H., 1998. Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution. Proceedings of the National Academy of Sciences 95, 5899–5905. https://doi.org/10.1073/pnas.95.11.5899

Guimarães, M.L., Vicente, A.C.P., Otsuki, K., da Silva, R.F.F.C., Francisco, M., da Silva, F.G., Serrano, D., Morgado, M.G., Bello, G., 2009. Close phylogenetic relationship between Angolan and Romanian HIV-1 subtype F1 isolates. Retrovirology 6, 39. https://doi.org/10.1186/1742-4690-6-39

Guirimand, T., Delmotte, S., and Navratil, V. (2015). VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. Nucleic acids research 43, D583-587.

Gupta, R.K., Kohli, A., McCormick, A.L., Towers, G.J., Pillay, D., Parry, C.M., 2010. Full-length HIV-1 Gag determines protease inhibitor susceptibility within in vitro assays. AIDS 24, 1651–1655. https://doi.org/10.1097/qad.0b013e3283398216

Gürtler, L.G., Hauser, P.H., Eberle, J., von Brunn, A., Knapp, S., Zekeng, L., Tsague, J.M., Kaptue, L., 1994. A new subtype of human immunodeficiency virus type 1 (MVP-5180) from Cameroon. J. Virol. 68, 1581–1585.

Haapa-Paananen, S., Wahlberg, N., Savilahti, H., 2014. Phylogenetic analysis of Maverick/Polinton giant transposons across organisms. Molecular Phylogenetics and Evolution 78, 271–274. https://doi.org/10.1016/j.ympev.2014.05.024

Hackett, J., Zekeng, L., Brennan, C.A., Lund, J.K., Vallari, A.S., Hickman, R.K., Gürtler, L., Kaptué, L., Devare, S.G., 1997. Genetic analysis of HIV type 1 group O p24gag sequences from Cameroon and Equatorial Guinea. AIDS Res. Hum. Retroviruses 13, 1155-1158. https://doi.org/10.1089/aid.1997.13.1155

Hackl, T., Martin, R., Barenhoff, K., Duponchel, S., Heider, D., Fischer, M.G., 2019. Four high-quality draft genome assemblies of the marine heterotrophic nanoflagellate Cafeteria roenbergensis. bioRxiv. https://doi.org/10.1101/751586

Hald, A., 1999. On the History of Maximum Likelihood in Relation to Inverse Probability and Least Squares. Statistical Science 14, 214–222.

Haldane, J.B. (1948). The theory of a cline. J Genet 48, 277-284.

Halper-Stromberg, A., Nussenzweig, M.C., 2016. Towards HIV-1 remission: potential roles for broadly neutralizing antibodies. J. Clin. Invest. 126, 415–423. https://doi.org/10.1172/JCI80561

Harris, M.E., Serwadda, D., Sewankambo, N., Kim, B., Kigozi, G., Kiwanuka, N., Phillips, J.B., Wabwire, F., Meehen, M., Lutalo, T., Lane, J.R., Merling, R., Gray, R., Wawer, M., Birx, D.L., Robb, M.L., McCutchan, F.E., 2002. Among 46 near full length HIV type 1 genome sequences from Rakai District, Uganda, subtype D and AD recombinants predominate. AIDS Res. Hum. Retroviruses 18, 1281–1290. https://doi.org/10.1089/088922202320886325

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their

applications. Biometrika 57, 97–109. https://doi.org/10.1093/biomet/57.1.97
Headrick, D.R., 2014. Sleeping sickness epidemics and colonial responses in East and Central Africa, 1900-1940. PLoS Negl. Trop. Dis. 8, e2772. https://doi.org/10.1371/journal.pntd.0002772
Heather, J.M., Chain, B., 2016. The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003
Heipertz, R.A., Ayemoba, O., Sanders-Buell, E., Poltavee, K., Pham, P., Kijak, G.H., Lei, E., Bose, M., Howell, S., O'Sullivan, A.M., Bates, A., Cervenka, T., Kuroiwa, J., Akintunde, A., Ibezim, O., Alabi, A., Okoye, O., Manak, M., Malia, J., Peel, S., Maisaka, M., Singer, D., O'Connell, R.J., Robb, M.L., Kim, J.H., Michael, N.L., Njoku, O., Tovanabutra, S., 2016. Significant contribution of subtype G to HIV-1 genetic complexity in Nigeria identified by a newly developed subtyping assay specific for subtype G and CRF02_AG. Medicine (Baltimore) 95, e4346. https://doi.org/10.1097/MD.0000000000004346
Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Fleminger, I., Kirtley, S., Williams, B., Gouws-Williams, E., Ghys, P.D., WHO–UNAIDS Network for HIV Isolation Characterisation, 2019. Global and regional molecular epidemiology of HIV-1, 1990-2015: a systematic review, global survey, and trend analysis. Lancet Infect Dis 19, 143–155. https://doi.org/10.1016/S1473-3099(18)30647-9
Hemelaar, J., Gouws, E., Ghys, P.D., Osmanov, S., 2006. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. AIDS 20, W13-23. https://doi.org/10.1097/01.aids.0000247564.73009.bc
Hershkovitz, I., Weber, G.W., Quam, R., Duval, M., Grun, R., Kinsley, L., Ayalon, A., Bar-Matthews, M., Valladas, H., Mercier, N., et al. (2018). The earliest modern humans outside Africa. Science 359, 456-459.
Hertz, T., Logan, M.G., Rolland, M., Magaret, C.A., Rademeyer, C., Fiore-Gartland, A., Edlefsen, P.T., DeCamp, A., Ahmed, H., Ngandu, N., Larsen, B.B., Frahm, N., Marais, J., Thebus, R., Geraghty, D., Hural, J., Corey, L., Kublin, J., Gray, G., McElrath, M.J., Mullins, J.I., Gilbert, P.B., Williamson, C., 2016. A study of vaccine-induced immune pressure on breakthrough infections in the Phambili phase 2b HIV-1 vaccine efficacy trial. Vaccine 34, 5792–5801. https://doi.org/10.1016/j.vaccine.2016.09.054
Hill, A.V.S. (1991). HLA associations with malaria in Africa: some implications for MHC evolution. In Molecular evolution of the major histocompatibility complex, J. Klein, and D. Klein, eds. (Heidelberg, Berlin: Springer Berlin Heidelberg), pp. 403-420.
Hillis, D.M., Bull, J.J., 1993. An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. Systematic Biology 42, 182–192. https://doi.org/10.1093/sysbio/42.2.182
HIV type 1 variation in World Health Organization-sponsored vaccine evaluation sites: genetic screening, sequence analysis, and preliminary biological

characterization of selected viral strains. WHO Network for HIV Isolation and Characterization, 1994. . AIDS Res. Hum. Retroviruses 10, 1327–1343. https://doi.org/10.1089/aid.1994.10.1327

Ho, S.Y.W., Jermiin, L.S., 2004. Tracing the Decay of the Historical Signal in Biological Sequence Data. Systematic Biology 53, 623–637. https://doi.org/10.1080/10635150490503035

Hoelscher, M., Kim, B., Maboko, L., Mhalu, F., von Sonnenburg, F., Birx, D.L., McCutchan, F.E., UNAIDS Network for HIV Isolation and Characterization, 2001. High proportion of unrelated HIV-1 intersubtype recombinants in the Mbeya region of southwest Tanzania. AIDS 15, 1461–1470. https://doi.org/10.1097/00002030-200108170-00002

Holley, R.W., Apgar, J., Everett, G.A., Madison, J.T., Marquisee, M., Merrill, S.H., Penswick, J.R., Zamir, A., 1965. Structure of a ribonucleic acid. Science 147, 1462–1465. https://doi.org/10.1126/science.147.3664.1462

Holley, R.W., Apgar, J., Merrill, S.H., Zubkoff, P.L., 1961. Nucleotide and oligonucleotide compositions of the alanine-, valine-, and tyrosine-acceptor "soluble" ribonucleic acids of yeast. Journal of the American Chemical Society 83, 4861–4862. https://doi.org/10.1021/ja01484a040

Holmes, E.C., 2011. The Evolution of Endogenous Viral Elements. Cell Host & Microbe 10, 368–377. https://doi.org/10.1016/j.chom.2011.09.002

Holzmayer, V., Zekeng, L., Kaptué, L., Gürtler, L., Devare, S.G., Hackett, J., 2005. Near-full-length genomic sequence of a human immunodeficiency type 1 subtype G strain from Cameroon. AIDS Res. Hum. Retroviruses 21, 414–419. https://doi.org/10.1089/aid.2005.21.414

Honess, R.W., 1984. Herpes Simplex and "The Herpes Complex": Diverse Observations and A Unifying Hypothesis. Journal of General Virology 65, 2077–2107. https://doi.org/10.1099/0022-1317-65-12-2077

Honeyborne, I., Prendergast, A., Pereyra, F., Leslie, A., Crawford, H., Payne, R., Reddy, S., Bishop, K., Moodley, E., Nair, K., van der Stok, M., McCarthy, N., Rousseau, C.M., Addo, M., Mullins, J.I., Brander, C., Kiepiela, P., Walker, B.D., Goulder, P.J.R., 2007. Control of human immunodeficiency virus type 1 is associated with HLA-B*13 and targeting of multiple gag-specific CD8+ T-cell epitopes. J. Virol. 81, 3667–3672. https://doi.org/10.1128/JVI.02689-06

Hora, B., Chen,Y., Denny,T.N., Hewlett,I., Ragupathy,V., Gao,F., 2013. Characterization of near full-length genome sequences for standard panels of HIV-1 isolates established at the External Quality Assurance Program Oversight Laboratory (EQAPOL). Unpublished direct submission to Genbank.

Hora, B., Keating, S.M., Chen, Y., Sanchez, A.M., Sabino, E., Hunt, G., Ledwaba, J., Hackett, J., Swanson, P., Hewlett, I., Ragupathy, V., Vikram Vemula, S., Zeng, P., Tee, K.-K., Chow, W.Z., Ji, H., Sandstrom, P., Denny, T.N., Busch, M.P., Gao, F., REDS-III and EQAPOL programs, 2016. Genetic Characterization of a Panel of Diverse HIV-1 Isolates at Seven International Sites. PLoS ONE 11,

e0157340. https://doi.org/10.1371/journal.pone.0157340
Horie, M., Honda, T., Suzuki, Y., Kobayashi, Y., Daito, T., Oshida, T., Ikuta, K., Jern, P., Gojobori, T., Coffin, J.M., Tomonaga, K., 2010. Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature 463, 84–87. https://doi.org/10.1038/nature08695
Howard, T.M., Olaleye, D.O., Rasheed, S., 1995. Molecular characterization of the gp120 encoding region of a new subtype A strain of HIV-1 (HIV-1IbNg) from Nigeria. Unpublished direct submission to Genbank.
Howard, T.M., Olaylele, D.O., Rasheed, S., 1994. Sequence analysis of the glycoprotein 120 coding region of a new HIV type 1 subtype A strain (HIV-1IbNg) from Nigeria. AIDS Res. Hum. Retroviruses 10, 1755–1757. https://doi.org/10.1089/aid.1994.10.1755
Howard, T.M., Rasheed, S., 1996. Genomic structure and nucleotide sequence analysis of a new HIV type 1 subtype A strain from Nigeria. AIDS Res. Hum. Retroviruses 12, 1413–1425. https://doi.org/10.1089/aid.1996.12.1413
Hu, D.J., Buvé, A., Baggs, J., van der Groen, G., Dondero, T.J., 1999. What role does HIV-1 subtype play in transmission and pathogenesis? An epidemiological perspective: AIDS 13, 873–881. https://doi.org/10.1097/00002030-199905280-00002
Huang, K.-H.G., Goedhals, D., Carlson, J.M., Brockman, M.A., Mishra, S., Brumme, Z.L., Hickling, S., Tang, C.S.W., Miura, T., Seebregts, C., Heckerman, D., Ndung'u, T., Walker, B., Klenerman, P., Steyn, D., Goulder, P., Phillips, R., Bloemfontein-Oxford Collaborative Group, van Vuuren, C., Frater, J., 2011. Progression to AIDS in South Africa is associated with both reverting and compensatory viral mutations. PLoS ONE 6, e19018. https://doi.org/10.1371/journal.pone.0019018
Huang, K.-H.G., Goedhals, D., Fryer, H., van Vuuren, C., Katzourakis, A., De Oliveira, T., Brown, H., Cassol, S., Seebregts, C., McLean, A., Klenerman, P., Phillips, R., Frater, J., Bloemfontein-Oxford Collaborative Group, 2009. Prevalence of HIV type-1 drug-associated mutations in pre-therapy patients in the Free State, South Africa. Antivir. Ther. (Lond.) 14, 975–984. https://doi.org/10.3851/IMP1416
Hudgens, M.G., Longini, I.M., Vanichseni, S., Hu, D.J., Kitayaporn, D., Mock, P.A., Halloran, M.E., Satten, G.A., Choopanya, K., Mastro, T.D., 2002. Subtype-specific transmission probabilities for human immunodeficiency virus type 1 among injecting drug users in Bangkok, Thailand. Am J Epidemiol 155, 159–168. https://doi.org/10.1093/aje/155.2.159
Huelsenbeck, J.P., Rannala, B., 2004. Frequentist Properties of Bayesian Posterior Probabilities of Phylogenetic Trees Under Simple and Complex Substitution Models. Systematic Biology 53, 904–913. https://doi.org/10.1080/10635150490522629
Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755.

https://doi.org/10.1093/bioinformatics/17.8.754
Huet, T., Dazza, M.C., Brun-Vézinet, F., Roelants, G.E., Wain-Hobson, S., 1989. A highly defective HIV-1 strain isolated from a healthy Gabonese individual presenting an atypical western blot. AIDS 3, 707–715. https://doi.org/10.1097/00002030-198911000-00004
Hughes, A.L., and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335, 167-170.
Hughes, A.L., Yeager, M., 1998. Natural selection at major histocompatibility complex loci of vertebrates. Annual Review of Genetics 32, 415–435. https://doi.org/10.1146/annurev.genet.32.1.415
Hughes, T.J. (1985). The Great Cenozoic ice sheet. Palaeography, Palaeoclimatology, Palaeoecology 50, 9-43.
Hunt, G.M., Papathanasopoulos, M.A., Gray, G.E., Tiemessen, C.T., 2003. Characterisation of near-full length genome sequences of three South African human immunodeficiency virus type 1 subtype C isolates. Virus Genes 26, 49–56.
Iacob, S.A., Iacob, D.G., Jugulete, G., 2017. Improving the Adherence to Antiretroviral Therapy, a Difficult but Essential Task for a Successful HIV Treatment-Clinical Points of View and Practical Considerations. Front. Pharmacol. 8, 831. https://doi.org/10.3389/fphar.2017.00831
Inkscape Project, 2017. Inkscape.
Institute of Medicine (US) Committee on the Assessment of Future Scientific Needs for Live Variola Virus, 1999. Assessment of Future Scientific Needs for Live Variola Virus. National Academies Press (US), Washington (DC).
Ivanovski, D., 1892. Über die Mosaikkrankheit der Tabakspflanze. St Petersb Acad Imp Sci Bul 67–70.
Iwabu, Y., Kinomoto, M., Tatsumi, M., Fujita, H., Shimura, M., Tanaka, Y., Ishizaka, Y., Nolan, D., Mallal, S., Sata, T., Tokunaga, K., 2010. Differential anti-APOBEC3G activity of HIV-1 Vif proteins derived from different subtypes. J. Biol. Chem. 285, 35350–35358. https://doi.org/10.1074/jbc.M110.173286
Iweriebor, B.C., Bessong, P.O., Mavhandu, L.G., Masebe, T.M., Nwobegahay, J., Moyo, S.R., Mphahlele, J.M., 2011. Genetic analysis of the near full-length genome of an HIV type 1 A1/C unique recombinant form from northern South Africa. AIDS Res. Hum. Retroviruses 27, 911–915. https://doi.org/10.1089/aid.2010.0286
Iweriebor, B.C., Masebe, T., Nwobegahay, J., Mphahlele, J.M., Bessong, P.O., 2010. Genetic Characterization of HIV-1 Unique C/CRF11_cpx Recombinant Form from Limpopo Province, South Africa. Unpublished direct submission to Genbank.
Iyer, L.M., Abhiman, S., Aravind, L., 2008. A new family of polymerases related to superfamily A DNA polymerases and T7-like DNA-dependent RNA polymerases. Biology Direct 3, 39. https://doi.org/10.1186/1745-6150-3-39

Iyer, S.S., Bibollet-Ruche, F., Sherrill-Mix, S., Learn, G.H., Plenderleith, L., Smith, A.G., Barbian, H.J., Russell, R.M., Gondim, M.V.P., Bahari, C.Y., Shaw, C.M., Li, Y., Decker, T., Haynes, B.F., Shaw, G.M., Sharp, P.M., Borrow, P., Hahn, B.H., 2017. Resistance to type 1 interferons is a major determinant of HIV-1 transmission fitness. Proc. Natl. Acad. Sci. U.S.A. 114, E590–E599. https://doi.org/10.1073/pnas.1620144114

Jablonski, D., 1994. Extinctions in the fossil record. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 344, 11–17. https://doi.org/10.1098/rstb.1994.0045

Jacobs, G.B., Bock, S., Schuch, A., Moschall, R., Schrom, E.-M., Zahn, J., Reuter, C., Preiser, W., Rethwilm, A., Engelbrecht, S., Kerkau, T., Bodem, J., 2012. Construction of a high titer infectious HIV-1 subtype C proviral clone from South Africa. Viruses 4, 1830–1843. https://doi.org/10.3390/v4091830

Jacobs, G.B., Loxton, A.G., Laten, A., Engelbrecht, S., 2007. Complete genome sequencing of a non-syncytium-inducing HIV type 1 subtype D strain from Cape Town, South Africa. AIDS Res. Hum. Retroviruses 23, 1575–1578. https://doi.org/10.1089/aid.2007.0167

Jacobs, G.S., Hudjashov, G., Saag, L., Kusuma, P., Darusallam, C.C., Lawson, D.J., Mondal, M., Pagani, L., Ricaut, F.X., Stoneking, M., et al. (2019). Multiple Deeply Divergent Denisovan Ancestries in Papuans. Cell 177, 1010-1021 e1032.

Janssens, W., Heyndrickx, L., Fransen, K., Motte, J., Peeters, M., Nkengasong, J.N., Ndumbe, P.M., Delaporte, E., Perret, J.L., Atende, C., 1994. Genetic and phylogenetic analysis of env subtypes G and H in central Africa. AIDS Res. Hum. Retroviruses 10, 877–879. https://doi.org/10.1089/aid.1994.10.877

Janssens, W., Laukkanen, T., Salminen, M.O., Carr, J.K., Van der Auwera, G., Heyndrickx, L., van der Groen, G., McCutchan, F.E., 2000. HIV-1 subtype H near-full length genome reference strains and analysis of subtype-H-containing inter-subtype recombinants. AIDS 14, 1533–1543. https://doi.org/10.1097/00002030-200007280-00009

Jardetzky, T.S., Lane, W.S., Robinson, R.A., Madden, D.R., Wiley, D.C., 1991. Identification of self peptides bound to purified HLA-B27. Nature 353, 326–329. https://doi.org/10.1038/353326a0

Jeffroy, O., Brinkmann, H., Delsuc, F., Philippe, H., 2006. Phylogenomics: the beginning of incongruence? Trends Genet 22, 225–231. https://doi.org/10.1016/j.tig.2006.02.003

Jern, P., and Coffin, J.M. (2008). Effects of retroviruses on host genome function. Annu Rev Genet 42, 709-732.

Jha, A.R., Nixon, D.F., Rosenberg, M.G., Martin, J.N., Deeks, S.G., Hudson, R.R., Garrison, K.E., and Pillai, S.K. (2011). Human endogenous retrovirus K106 (HERV-K106) was infectious after the emergence of anatomically modern humans. PloS one 6, e20234.

Jiang, S., Steward, G., Jellison, R., Chu, W., Choi, S., 2004. Abundance, Distribution,

and Diversity of Viruses in Alkaline, Hypersaline Mono Lake, California. Microbial Ecology 47, 9–17. https://doi.org/10.1007/s00248-003-1023-x

John of Salisbury, 1159. Joh. Salisberiensis Policraticus et Metalogicon.

Jombart, T., Pontier, D., and Dufour, A.B. (2009). Genetic markers in the playground of multivariate analysis. Heredity (Edinb) 102, 330-341.

Jones, R.B., Garrison, K.E., Mujib, S., Mihajlovic, V., Aidarus, N., Hunter, D.V., Martin, E., John, V.M., Zhan, W., Faruk, N.F., et al. (2012). HERV-K-specific T cells eliminate diverse HIV-1/2 and SIV primary isolates. The Journal of clinical investigation 122, 4473-4489.

Jung, A.C., Paauw, D.S., 1998. Diagnosing HIV-Related disease: Using the CD4 count as a guide. Journal of General Internal Medicine 13, 131–136. https://doi.org/10.1046/j.1525-1497.1998.00031.x

Junqueira, D.M., Almeida, S.E. de M., 2016. HIV-1 subtype B: Traces of a pandemic. Virology 495, 173–184. https://doi.org/10.1016/j.virol.2016.05.003

K''Aluoch, O., Donovan, R., Barugahare, B., Mugyenyi, P., Sheppard, H., Cao, H., 2005. High Frequency of Recombinant A and D HIV-1 subtypes in a Population from Uganda. Unpublished direct submission to Genbank.

Kahle, E., Campbell, M., Lingappa, J., Donnell, D., Celum, C., Ondondo, R., Mujugira, A., Fife, K., Mugo, N., Kapiga, S., Mullins, J.I., Baeten, J.M., Partners in Prevention HSVHIV Transmission Study Team, 2014. HIV-1 subtype C is not associated with higher risk of heterosexual HIV-1 transmission: a multinational study among HIV-1 serodiscordant couples. AIDS 28, 235–243. https://doi.org/10.1097/QAD.0000000000000024

Kaleebu, P., French, N., Mahe, C., Yirrell, D., Watera, C., Lyagoba, F., Nakiyingi, J., Rutebemberwa, A., Morgan, D., Weber, J., Gilks, C., Whitworth, J., 2002. Effect of human immunodeficiency virus (HIV) type 1 envelope subtypes A and D on disease progression in a large cohort of HIV-1-positive persons in Uganda. J Infect Dis 185, 1244–1250. https://doi.org/10.1086/340130

Kalish, M.L., Robbins, K.E., Pieniazek, D., Schaefer, A., Nzilambi, N., Quinn, T.C., St Louis, M.E., Youngpairoj, A.S., Phillips, J., Jaffe, H.W., Folks, T.M., 2004. Recombinant viruses and early global HIV-1 epidemic. Emerging Infect. Dis. 10, 1227–1234. https://doi.org/10.3201/eid1007.030904

Kalish, M.L., Wolfe, N.D., Ndongmo, C.B., McNicholl, J., Robbins, K.E., Aidoo, M., Fonjungo, P.N., Alemnji, G., Zeh, C., Djoko, C.F., Mpoudi-Ngole, E., Burke, D.S., Folks, T.M., 2005. Central African hunters exposed to simian immunodeficiency virus. Emerging Infect. Dis. 11, 1928–1930. https://doi.org/10.3201/eid1112.050394

Kandathil, A.J., Kannangai, R., Abraham, O.C., Rupali, P., Pulimood, S.A., Verghese, V.P., Grant, P., Pillay, D., Sridharan, G., 2009. The frequency of HIV-I drug resistance mutations among treatment-naive individuals at a tertiary care centre in south India. Int J STD AIDS 20, 522–526. https://doi.org/10.1258/ijsa.2008.008403

Kane, C.T., Montavon, C., Toure, M.A., Faye, M.A., Ndiaye, A.G., Diallo, A.G.,

Ndoye, I., Liegeois, F., Delaporte, E., Mboup, S., Peeters, M., 2001. Full-length genome sequencing of HIV type 1 group O viruses isolated from a heterosexual transmission cluster in Senegal. AIDS Res. Hum. Retroviruses 17, 1211–1216. https://doi.org/10.1089/088922201316912817

Kantor, R., Katzenstein, D.A., Efron, B., Carvalho, A.P., Wynhoven, B., Cane, P., Clarke, J., Sirivichayakul, S., Soares, M.A., Snoeck, J., Pillay, C., Rudich, H., Rodrigues, R., Holguin, A., Ariyoshi, K., Bouzas, M.B., Cahn, P., Sugiura, W., Soriano, V., Brigido, L.F., Grossman, Z., Morris, L., Vandamme, A.-M., Tanuri, A., Phanuphak, P., Weber, J.N., Pillay, D., Harrigan, P.R., Camacho, R., Schapiro, J.M., Shafer, R.W., 2005. Impact of HIV-1 subtype and antiretroviral therapy on protease and reverse transcriptase genotype: results of a global collaboration. PLoS Med. 2, e112. https://doi.org/10.1371/journal.pmed.0020112

Kapitonov, V.V., Jurka, J., 2001. Rolling-circle transposons in eukaryotes. Proceedings of the National Academy of Sciences 98, 8714–8719. https://doi.org/10.1073/pnas.151269298

Kapitonov, V.V., Jurka, J., 2006. Self-synthesizing DNA transposons in eukaryotes. Proceedings of the National Academy of Sciences 103, 4540–4545. https://doi.org/10.1073/pnas.0600833103

Kapli, P., Yang, Z., Telford, M.J., 2020. Phylogenetic tree building in the genomic age. Nature Reviews Genetics 21, 428–444. https://doi.org/10.1038/s41576-020-0233-0

Kassiotis, G. (2014). Endogenous retroviruses and the development of cancer. Journal of immunology 192, 1343-1349.

Kassiotis, G., and Stoye, J.P. (2016). Immune responses to endogenous retroelements: taking the bad with the good. Nature reviews Immunology 16, 207-219.

Katzourakis, A., Aswad, A., 2014. The origins of giant viruses, virophages and their relatives in host genomes. BMC Biology 12. https://doi.org/10.1186/s12915-014-0051-y

Katzourakis, A., Gifford, R.J., 2010. Endogenous Viral Elements in Animal Genomes. PLoS Genetics 6, e1001191. https://doi.org/10.1371/journal.pgen.1001191

Kawashima, M., Ohashi, J., Nishida, N., Tokunaga, K., 2012. Evolutionary Analysis of Classical HLA Class I and II Genes Suggests That Recent Positive Selection Acted on DPB1*04:01 in Japanese Population. PLoS ONE 7, e46806. https://doi.org/10.1371/journal.pone.0046806

Kawashima, Y., Pfafferott, K., Frater, J., Matthews, P., Payne, R., Addo, M., Gatanaga, H., Fujiwara, M., Hachiya, A., Koizumi, H., Kuse, N., Oka, S., Duda, A., Prendergast, A., Crawford, H., Leslie, A., Brumme, Z., Brumme, C., Allen, T., Brander, C., Kaslow, R., Tang, J., Hunter, E., Allen, S., Mulenga, J., Branch, S., Roach, T., John, M., Mallal, S., Ogwu, A., Shapiro, R., Prado, J.G., Fidler, S., Weber, J., Pybus, O.G., Klenerman, P., Ndung'u, T., Phillips, R., Heckerman, D., Harrigan, P.R., Walker, B.D., Takiguchi, M., Goulder, P., 2009. Adaptation of HIV-1 to human leukocyte antigen class I. Nature 458,

641–645. https://doi.org/10.1038/nature07746

Keele, B.F., Van Heuverswyn, F., Li, Y., Bailes, E., Takehisa, J., Santiago, M.L., Bibollet-Ruche, F., Chen, Y., Wain, L.V., Liegeois, F., Loul, S., Ngole, E.M., Bienvenue, Y., Delaporte, E., Brookfield, J.F.Y., Sharp, P.M., Shaw, G.M., Peeters, M., Hahn, B.H., 2006. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. Science 313, 523–526. https://doi.org/10.1126/science.1126531

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., Lander, E.S., 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423, 241–254. https://doi.org/10.1038/nature01644

Kemal, K.S., Anastos, K., Weiser, B., Ramirez, C.M., Shi, Q., Burger, H., 2013. Molecular epidemiology of HIV type 1 subtypes in Rwanda. AIDS Res. Hum. Retroviruses 29, 957–962. https://doi.org/10.1089/AID.2012.0095

Key, F.M., Teixeira, J.C., de Filippo, C., and Andres, A.M. (2014). Advantageous diversity maintained by balancing selection in humans. Curr Opin Genet Dev 29, 45-51.

Kiepiela, P., Leslie, A.J., Honeyborne, I., Ramduth, D., Thobakgale, C., Chetty, S., Rathnavalu, P., Moore, C., Pfafferott, K.J., Hilton, L., Zimbwa, P., Moore, S., Allen, T., Brander, C., Addo, M.M., Altfeld, M., James, I., Mallal, S., Bunce, M., Barber, L.D., Szinger, J., Day, C., Klenerman, P., Mullins, J., Korber, B., Coovadia, H.M., Walker, B.D., Goulder, P.J.R., 2004. Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. Nature 432, 769–775. https://doi.org/10.1038/nature03113

Kiepiela, P., Ngumbela, K., Thobakgale, C., Ramduth, D., Honeyborne, I., Moodley, E., Reddy, S., de Pierres, C., Mncube, Z., Mkhwanazi, N., Bishop, K., van der Stok, M., Nair, K., Khan, N., Crawford, H., Payne, R., Leslie, A., Prado, J., Prendergast, A., Frater, J., McCarthy, N., Brander, C., Learn, G.H., Nickle, D., Rousseau, C., Coovadia, H., Mullins, J.I., Heckerman, D., Walker, B.D., Goulder, P., 2007. CD8+ T-cell responses to different HIV proteins have discordant associations with viral load. Nat. Med. 13, 46–53. https://doi.org/10.1038/nm1520

Kieschnick, R., McCullough, B.D., 2003. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. Statistical Modelling 3, 193–213. https://doi.org/10.1191/1471082X03st053oa

Kiguoya, M.W., Mann, J.K., Chopera, D., Gounder, K., Lee, G.Q., Hunt, P.W., Martin, J.N., Ball, T.B., Kimani, J., Brumme, Z.L., Brockman, M.A., Ndung'u, T., 2017. Subtype-Specific Differences in Gag-Protease-Driven Replication Capacity Are Consistent with Intersubtype Differences in HIV-1 Disease Progression. J. Virol. 91. https://doi.org/10.1128/JVI.00253-17

Kijak, G.H., Sanders-Buell, E., Chenine, A.-L., Eller, M.A., Goonetilleke, N., Thomas, R., Leviyang, S., Harbolick, E.A., Bose, M., Pham, P., Oropeza, C., Poltavee, K., O'Sullivan, A.M., Billings, E., Merbah, M., Costanzo, M.C., Warren, J.A., Slike, B., Li, H., Peachman, K.K., Fischer, W., Gao, F., Cicala, C., Arthos, J.,

Eller, L.A., O'Connell, R.J., Sinei, S., Maganga, L., Kibuuka, H., Nitayaphan, S., Rao, M., Marovich, M.A., Krebs, S.J., Rolland, M., Korber, B.T., Shaw, G.M., Michael, N.L., Robb, M.L., Tovanabutra, S., Kim, J.H., 2017. Rare HIV-1 transmitted/founder lineages identified by deep viral sequencing contribute to rapid shifts in dominant quasispecies during acute and early infection. PLoS Pathog. 13, e1006510. https://doi.org/10.1371/journal.ppat.1006510

Kijak, G.H., Sanders-Buell, E., Wolfe, N.D., Mpoudi-Ngole, E., Kim, B., Brown, B., Robb, M.L., Birx, D.L., Burke, D.S., Carr, J.K., McCutchan, F.E., 2004. Development and application of a high-throughput HIV type 1 genotyping assay to identify CRF02_AG in West/West Central Africa. AIDS Res. Hum. Retroviruses 20, 521–530. https://doi.org/10.1089/088922204323087778

Kim, J.H., 2019. Multicollinearity and misleading statistical results. Korean J Anesthesiol 72, 558–569. https://doi.org/10.4097/kja.19087

Kist, N.C., Lambert, B., Campbell, S., Katzourakis, A., Lunn, D., Lemey, P., Iversen, A.K.N., 2020. HIV-1 p24Gag adaptation to modern and archaic HLA-allele frequency differences in ethnic groups contributes to viral subtype diversification. Virus Evolution. https://doi.org/10.1093/ve/veaa085

Kiwanuka, N., Laeyendecker, O., Quinn, T.C., Wawer, M.J., Shepherd, J., Robb, M., Kigozi, G., Kagaayi, J., Serwadda, D., Makumbi, F.E., Reynolds, S.J., Gray, R.H., 2009. HIV-1 subtypes and differences in heterosexual HIV transmission among HIV-discordant couples in Rakai, Uganda. AIDS 23, 2479–2484. https://doi.org/10.1097/QAD.0b013e328330cc08

Klatt, N.R., Hawkins, P., Lumm, W., Farmer, P., Bansal, A., Mulenga, J., Yue, L., Allen, S., Derdeyn, C.A., Goepfert, P.A., Hunter, E., 2011. Transmission of Homogenous Gag after HIV Infection is not dependent on pressure from CTL escape. Unpublished direct submission to Genbank.

Klein, F., Halper-Stromberg, A., Horwitz, J.A., Gruell, H., Scheid, J.F., Bournazos, S., Mouquet, H., Spatz, L.A., Diskin, R., Abadir, A., Zang, T., Dorner, M., Billerbeck, E., Labitt, R.N., Gaebler, C., Marcovecchio, P., Incesu, R.-B., Eisenreich, T.R., Bieniasz, P.D., Seaman, M.S., Bjorkman, P.J., Ravetch, J.V., Ploss, A., Nussenzweig, M.C., 2012. HIV therapy by a combination of broadly neutralizing antibodies in humanized mice. Nature 492, 118–122. https://doi.org/10.1038/nature11604

Klein, J. (1987). Origin of major histocompatibility complex polymorphism: the trans-species hypothesis. Human immunology 19, 155-162.

Klotz, L.C., Blanken, R.L., 1981. A practical method for calculating evolutionary trees from sequence data. Journal of Theoretical Biology 91, 261–272. https://doi.org/10.1016/0022-5193(81)90233-2

Ko, I., Chang, H., 2017. Interactive Visualization of Healthcare Data Using Tableau. Healthc. Inform. Res. 23, 349–354. https://doi.org/10.4258/hir.2017.23.4.349

Koofhethile, C.K., Ndhlovu, Z.M., Thobakgale-Tshabalala, C., Prado, J.G., Ismail, N., Mncube, Z., Mkhize, L., van der Stok, M., Yende, N., Walker, B.D., Goulder, P.J.R., Ndung'u, T., 2016. CD8+ T Cell Breadth and Ex Vivo Virus Inhibition

Capacity Distinguish between Viremic Controllers with and without Protective HLA Class I Alleles. J. Virol. 90, 6818–6831. https://doi.org/10.1128/JVI.00276-16

Koonin, E.V., Krupovic, M., Yutin, N., 2015. Evolution of double-stranded DNA viruses of eukaryotes: from bacteriophages to transposons to giant viruses. Annals of the New York Academy of Sciences 1341, 10–24. https://doi.org/10.1111/nyas.12728

Koonin, E.V., Senkevich, T.G., Dolja, V.V., 2006. The ancient Virus World and evolution of cells. Biol. Direct 1, 29. https://doi.org/10.1186/1745-6150-1-29

Koonin, E.V., Wolf, Y.I., 2009. The fundamental units, processes and patterns of evolution, and the tree of life conundrum. Biol. Direct 4, 33. https://doi.org/10.1186/1745-6150-4-33

Korber, B., Gaschen, B., Yusim, K., Thakallapally, R., Kesmir, C., Detours, V., 2001. Evolutionary and immunological implications of contemporary HIV-1 variation. Br Med Bull 58, 19–42. https://doi.org/10.1093/bmb/58.1.19

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B.H., Wolinsky, S., Bhattacharya, T., 2000. Timing the ancestor of the HIV-1 pandemic strains. Science 288, 1789–1796. https://doi.org/10.1126/science.288.5472.1789

Korn, T., Oukka, M., Kuchroo, V., Bettelli, E., 2007. Th17 cells: Effector T cells with inflammatory properties. Seminars in Immunology 19, 362–371. https://doi.org/10.1016/j.smim.2007.10.007

Koulinska, I.N., Msamanga, G., Mwakagile, D., Essex, M., Renjifo, B., 2002. Common genetic arrangements among human immunodeficiency virus type 1 subtype A and D recombinant genomes vertically transmitted in Tanzania. AIDS Res. Hum. Retroviruses 18, 947–956. https://doi.org/10.1089/088922202760265614

Koulinska, I.N., Ndung'u, T., Mwakagile, D., Msamanga, G., Kagoma, C., Fawzi, W., Essex, M., Renjifo, B., 2001. A new human immunodeficiency virus type 1 circulating recombinant form from Tanzania. AIDS Res. Hum. Retroviruses 17, 423–431. https://doi.org/10.1089/088922201750102508

Kromer, J., Hummel, T., Pietrowski, D., Giani, A.S., Sauter, J., Ehninger, G., Schmidt, A.H., Croy, I., 2016. Influence of HLA on human partnership and sexual satisfaction. Sci Rep 6, 32550. https://doi.org/10.1038/srep32550

Krupovic, M., Cvirkaite-Krupovic, V., 2011. Virophages or satellite viruses? Nature Reviews Microbiology 9, 762–763. https://doi.org/10.1038/nrmicro2676

Krupovic, M., Koonin, E.V., 2014. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. Nature Reviews Microbiology 13, 105–115. https://doi.org/10.1038/nrmicro3389

Krupovic, M., Kuhn, J.H., Fischer, M.G., 2015. A classification system for virophages and satellite viruses. Archives of Virology 161, 233–247. https://doi.org/10.1007/s00705-015-2622-9

Kumar, R., Qureshi, H., Deshpande, S., Bhattacharya, J., 2018. Broadly neutralizing antibodies in HIV-1 treatment and prevention. Ther. Adv. Vaccines

Immunother. 6, 61–68. https://doi.org/10.1177/2515135518800689
Kusagawa, S., Takebe, Y., Yang, R., Motomura, K., Ampofo, W., Brandful, J., Koyanagi, Y., Yamamoto, N., Sata, T., Ishikawa, K., Hachiya, A., Oka, S., Nagai, Y., Tatsumi, M., 2000. Isolation and characterization of full-length molecular DNA clones of Ghanaian HIV-1 intersubtype A/G recombinant (CRF02_AG) which is replication-competent in restricted host-range. Unpublished direct submission to Genbank.
Kusagawa, S., Takebe, Y., Yang, R., Motomura, K., Ampofo, W., Brandful, J., Koyanagi, Y., Yamamoto, N., Sata, T., Ishikawa, K., Nagai, Y., Tatsumi, M., 2001. Isolation and characterization of a full-length molecular DNA clone of Ghanaian HIV type 1 intersubtype A/G recombinant CRF02_AG, which is replication competent in a restricted host range. AIDS Res. Hum. Retroviruses 17, 649–655. https://doi.org/10.1089/088922201300119761
Kuznetsov, Y.G., Klose, T., Rossmann, M., McPherson, A., 2013. Morphogenesis of mimivirus and its viral factories: an atomic force microscopy study of infected cells. J Virol 87, 11200–11213. https://doi.org/10.1128/JVI.01372-13
La Scola, B., Audic, S., Robert, C., Jungang, L., de Lamballerie, X., Drancourt, M., Birtles, R., Claverie, J.-M., Raoult, D., 2003. A giant virus in amoebae. Science 299, 2033. https://doi.org/10.1126/science.1081867
La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., Merchat, M., Suzan-Monti, M., Forterre, P., Koonin, E., Raoult, D., 2008. The virophage as a unique parasite of the giant mimivirus. Nature 455, 100–104. https://doi.org/10.1038/nature07218
Lamnidis, T.C., Majander, K., Jeong, C., Salmela, E., Wessman, A., Moiseyev, V., Khartanovich, V., Balanovsky, O., Ongyerth, M., Weihmann, A., et al. (2018). Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe. Nature communications 9, 5018.
Land, A.M., Ball, T.B., Luo, M., Rutherford, J., Sarna, C., Wachihi, C., Kimani, J., Plummer, F.A., 2008. Full-length HIV type 1 proviral sequencing of 10 highly exposed women from Nairobi, Kenya reveals a high proportion of intersubtype recombinants. AIDS Res. Hum. Retroviruses 24, 865–872. https://doi.org/10.1089/aid.2007.0200
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R.,

Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J., Szustakowki, J., International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921. https://doi.org/10.1038/35057062

Lao, O., Lu, T.T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L.A., Comas, D., Holmlund, G., Kouvatsi, A., Macek, M., Mollet, I., Parson, W., Palo, J., Ploski, R., Sajantila, A., Tagliabracci, A., Gether, U., Werge, T., Rivadeneira, F., Hofman, A., Uitterlinden, A.G., Gieger, C., Wichmann, H.-E., Rüther, A., Schreiber, S., Becker, C., Nürnberg, P., Nelson, M.R., Krawczak, M., Kayser, M., 2008. Correlation between Genetic and Geographic Structure in Europe. Current

Biology 18, 1241–1248. https://doi.org/10.1016/j.cub.2008.07.049
Larget, B., Simon, D.L., 1999. Markov Chasin Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees. Molecular Biology and Evolution 16, 750–759. https://doi.org/10.1093/oxfordjournals.molbev.a026160
Larsson, A., 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30, 3276–3278. https://doi.org/10.1093/bioinformatics/btu531
Lau, K.A., Wong, J.J.L., 2013. Current Trends of HIV Recombination Worldwide. Infect Dis Rep 5, e4. https://doi.org/10.4081/idr.2013.s1.e4
Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513, 409-413.
Lederberg, E.M., Lederberg, J., 1953. Genetic Studies of Lysogenicity in Escherichia Coli. Genetics 38, 51–64.
Legendre, M., Bartoli, J., Shmakova, L., Jeudy, S., Labadie, K., Adrait, A., Lescot, M., Poirot, O., Bertaux, L., Bruley, C., Coute, Y., Rivkina, E., Abergel, C., Claverie, J.-M., 2014. Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. Proceedings of the National Academy of Sciences 111, 4274–4279. https://doi.org/10.1073/pnas.1320670111
Legendre, M., Lartigue, A., Bertaux, L., Jeudy, S., Bartoli, J., Lescot, M., Alempic, J.-M., Ramus, C., Bruley, C., Labadie, K., Shmakova, L., Rivkina, E., Couté, Y., Abergel, C., Claverie, J.-M., 2015. In-depth study of Mollivirus sibericum , a new 30,000-y-old giant virus infecting Acanthamoeba. Proceedings of the National Academy of Sciences 112, E5327–E5335. https://doi.org/10.1073/pnas.1510795112
Legendre, P., Legendre, L., 2012. Numerical ecology, Third English edition. ed, Developments in environmental modelling. Elsevier, Amsterdam.
Leitman, E.M., Willberg, C.B., Tsai, M.-H., Chen, H., Buus, S., Chen, F., Riddell, L., Haas, D., Fellay, J., Goedert, J.J., Piechocka-Trocha, A., Walker, B.D., Martin, J., Deeks, S., Wolinsky, S.M., Martinson, J., Martin, M., Qi, Y., Sáez-Cirión, A., Yang, O.O., Matthews, P.C., Carrington, M., Goulder, P.J.R., 2017. HLA-B*14:02-Restricted Env-Specific CD8+ T-Cell Activity Has Highly Potent Antiviral Efficacy Associated with Immune Control of HIV Infection. J. Virol. 91. https://doi.org/10.1128/JVI.00544-17
Lenth, R.V., 2021. emmeans: Estimated Marginal Means, aka Least-Squares Means.
Lever, J., Krzywinski, M., Altman, N., 2017. Principal component analysis. Nature Methods 14, 641–642. https://doi.org/10.1038/nmeth.4346
Levine, B.B., Ojeda, A., Benacerraf, B., 1963. Studies on artificial antigens. III. The genetic control of the immune response to hapten-poly-l-lysine conjugates in guinea pigs. J Exp Med 118, 953–957. https://doi.org/10.1084/jem.118.6.953
Lewis, P.O., Holder, M.T., Holsinger, K.E., 2005. Polytomies and Bayesian

Phylogenetic Inference. Systematic Biology 54, 241–253. https://doi.org/10.1080/10635150590924208

Li, L.M., Grassly, N.C., Fraser, C., 2017. Quantifying Transmission Heterogeneity Using Both Pathogen Phylogenies and Incidence Time Series. Molecular Biology and Evolution 34, 2982–2995. https://doi.org/10.1093/molbev/msx195

Liang, B., Luo, M., Scott-Herridge, J., Semeniuk, C., Mendoza, M., Capina, R., Sheardown, B., Ji, H., Kimani, J., Ball, B.T., Van Domselaar, G., Graham, M., Tyler, S., Jones, S.J.M., Plummer, F.A., 2011. A comparison of parallel pyrosequencing and sanger clone-based sequencing and its impact on the characterization of the genetic diversity of HIV-1. PLoS ONE 6, e26745. https://doi.org/10.1371/journal.pone.0026745

Liégeois, F., Boué, V., Butel, C., Mouinga-Ondémé, A., Sica, J., Zamba, C., Peeters, M., Delaporte, E., Rouet, F., 2013. HIV type-1 group O infection in Gabon: low prevalence rate but circulation of genetically diverse and drug-resistant HIV type-1 group O strains. AIDS Res. Hum. Retroviruses 29, 1085–1090. https://doi.org/10.1089/aid.2012.0375

Lindblad-Toh, K., Garber, M., Or Zuk, Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, Genome Institute at Washington University, Broad Institute Sequencing Platform and Whole Genome Assembly Team, Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L.D., Lowe, C.B., Holloway, A.K., Clamp, M., Gnerre, S., Alföldi, J., Beal, K., Chang, J., Clawson, H., Cuff, J., Di Palma, F., Fitzgerald, S., Flicek, P., Guttman, M., Hubisz, M.J., Jaffe, D.B., Jungreis, I., Kent, W.J., Kostka, D., Lara, M., Martins, A.L., Massingham, T., Moltke, I., Raney, B.J., Rasmussen, M.D., Robinson, J., Stark, A., Vilella, A.J., Wen, J., Xie, X., Zody, M.C., Worley, K.C., Kovar, C.L., Muzny, D.M., Gibbs, R.A., Warren, W.C., Mardis, E.R., Weinstock, G.M., Wilson, R.K., Birney, E., Margulies, E.H., Herrero, J., Green, E.D., Haussler, D., Siepel, A., Goldman, N., Pollard, K.S., Pedersen, J.S., Lander, E.S., Kellis, M., 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478, 476–482. https://doi.org/10.1038/nature10530

Linde, M.E., Colquhoun, D.R., Ubaida Mohien, C., Kole, T., Aquino, V., Cotter, R., Edwards, N., Hildreth, J.E.K., Graham, D.R., 2013. The conserved set of host proteins incorporated into HIV-1 virions suggests a common egress pathway in multiple cell types. J. Proteome Res. 12, 2045–2054. https://doi.org/10.1021/pr300918r

Liu, F., Kong, Y., 2015. zoib: An R Package for Bayesian Inference for Beta Regression and Zero/One Inflated Beta Regression. The R Journal 7, 34. https://doi.org/10.32614/RJ-2015-019

Liu, M.K.P., Hawkins, N., Ritchie, A.J., Ganusov, V.V., Whale, V., Brackenridge, S., Li, H., Pavlicek, J.W., Cai, F., Rose-Abrahams, M., Treurnicht, F., Hraber, P., Riou, C., Gray, C., Ferrari, G., Tanner, R., Ping, L.-H., Anderson, J.A., Swanstrom, R., CHAVI Core B, Cohen, M., Karim, S.S.A., Haynes, B.,

Borrow, P., Perelson, A.S., Shaw, G.M., Hahn, B.H., Williamson, C., Korber, B.T., Gao, F., Self, S., McMichael, A., Goonetilleke, N., 2013. Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. J. Clin. Invest. 123, 380–393. https://doi.org/10.1172/JCI65330

Lockhart, P.J., Howe, C.J., Bryant, D.A., Beanland, T.J., Larkum, A.W., 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. J Mol Evol 34, 153–162. https://doi.org/10.1007/BF00182392

Los Alamos National Laboratory HIV Sequence Database [WWW Document], 2019. URL http://www.hiv.lanl.gov/

Louwagie, J., Delwart, E.L., Mullins, J.I., McCutchan, F.E., Eddy, G., Burke, D.S., 1994. Genetic analysis of HIV-1 isolates from Brazil reveals presence of two distinct genetic subtypes. AIDS Res. Hum. Retroviruses 10, 561–567. https://doi.org/10.1089/aid.1994.10.561

Louwagie, J., McCutchan, F.E., Peeters, M., Brennan, T.P., Sanders-Buell, E., Eddy, G.A., van der Groen, G., Fransen, K., Gershy-Damet, G.M., Deleys, R., 1993. Phylogenetic analysis of gag genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. AIDS 7, 769–780. https://doi.org/10.1097/00002030-199306000-00003

Loxton, A.G., Treurnicht, F., Laten, A., van Rensburg, E.J., Engelbrecht, S., 2005. Sequence analysis of near full-length HIV type 1 subtype D primary strains isolated in Cape Town, South Africa, from 1984 to 1986. AIDS Res. Hum. Retroviruses 21, 410–413. https://doi.org/10.1089/aid.2005.21.410

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., Chen, J., Meng, Y., Wang, J., Lin, Y., Yuan, J., Xie, Z., Ma, J., Liu, W.J., Wang, D., Xu, W., Holmes, E.C., Gao, G.F., Wu, G., Chen, W., Shi, W., Tan, W., 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 395, 565–574. https://doi.org/10.1016/S0140-6736(20)30251-8

Lüdecke, D., 2018. ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. JOSS 3, 772. https://doi.org/10.21105/joss.00772

Luk, K.-C., Berg, M.G., Naccache, S.N., Kabre, B., Federman, S., Mbanya, D., Kaptué, L., Chiu, C.Y., Brennan, C.A., Hackett, J., 2015. Utility of Metagenomic Next-Generation Sequencing for Characterization of HIV and Human Pegivirus Diversity. PLoS ONE 10, e0141723. https://doi.org/10.1371/journal.pone.0141723

Luk, K.-C., Holzmayer, V., Ndembi, N., Swanson, P., Brennan, C.A., Ngansop, C., Mbanya, D., Kaptué, L., Gürtler, L., Devare, S.G., Hackett, J., 2008. Near full-length genome characterization of an HIV type 1 CRF25_cpx strain from Cameroon. AIDS Res. Hum. Retroviruses 24, 1309–1314. https://doi.org/10.1089/aid.2008.0124

Luk, K.-C., Holzmayer, V., Yamaguchi, J., Swanson, P., Brennan, C.A., Ngansop, C., Mbanya, D., Gayum, H., Djuidje, M.N., Ndembi, N., Kamdem, D., Kaptué, L.,

Gürtler, L., Devare, S.G., Hackett, J., 2007. Near full-length genome characterization of three additional HIV type 1 CRF13_cpx strains from Cameroon. AIDS Res. Hum. Retroviruses 23, 297–302. https://doi.org/10.1089/aid.2006.0177

Luppi, M., Marasca, R., Barozzi, P., Ferrari, S., Ceccherini-Nelli, L., Batoni, G., Merelli, E., Torelli, G., 1993. Three cases of human herpesvirus-6 latent infection: Integration of viral genome in peripheral blood mononuclear cell DNA. Journal of Medical Virology 40, 44–52. https://doi.org/10.1002/jmv.1890400110

Mächler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M., Roudier, P., and Gonzalez, J. (2017). Cluster: "Finding Groups in Data": Cluster Analysis Extended Rousseeuw et Al. Version 2.0.5.

Mafessoni, F. (2019). Encounters with archaic hominins. Nat Ecol Evol 3, 14-15.

Magiorkinis, G., Sypsa, V., Magiorkinis, E., Paraskevis, D., Katsoulidou, A., Belshaw, R., Fraser, C., Pybus, O.G., Hatzakis, A., 2013. Integrating phylodynamics and epidemiology to estimate transmission diversity in viral epidemics. PLoS Comput. Biol. 9, e1002876. https://doi.org/10.1371/journal.pcbi.1002876

Mahasha, P.W., Cassol, E., Danaviah, S., Bond, R., Seebregts, C., van der Merwe, S.W., Rossouw, T., de Oliveira, T., 2013. Differential Distribution and Evolution of HIV-1 RNA Variants in the Gastrointestinal Tract of Antiretroviral Naive African AIDS Patients with Diarrhea and (or) Weight Loss. Unpublished direct submission to Genbank.

Mamadou, S., Montavon, C., Ben, A., Djibo, A., Rabiou, S., Mboup, S., Delaporte, E., Peeters, M., 2002. Predominance of CRF02-AG and CRF06-cpx in Niger, West Africa. AIDS Res. Hum. Retroviruses 18, 723–726. https://doi.org/10.1089/088922202760072357

Mamadou, S., Vidal, N., Montavon, C., Ben, A., Djibo, A., Rabiou, S., Soga, G., Delaporte, E., Mboup, S., Peeters, M., 2003. Emergence of complex and diverse CRF02-AG/CRF06-cpx recombinant HIV type 1 strains in Niger, West Africa. AIDS Res. Hum. Retroviruses 19, 77–82. https://doi.org/10.1089/08892220360474005

Manak, M., Sina, S., Anekella, B., Hewlett, I., Sanders-Buell, E., Ragupathy, V., Kim, J., Vermeulen, M., Stramer, S.L., Sabino, E., Grabarczyk, P., Michael, N., Peel, S., Garrett, P., Tovanabutra, S., Busch, M.P., Schito, M., 2012. Pilot studies for development of an HIV subtype panel for surveillance of global diversity. AIDS Res. Hum. Retroviruses 28, 594–606. https://doi.org/10.1089/AID.2011.0271

Manly, B.F.J., and Navarro Alberto, J.A. (2016). Multivariate Statistical Methods, 4 edn (CRC Press).

Marchi, E., Kanapin, A., Byott, M., Magiorkinis, G., and Belshaw, R. (2013). Neanderthal and Denisovan retroviruses in modern humans. Curr Biol 23, R994-R995.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979). Multivariate analysis, Probability

and mathematical statistics. (London, New York).
Markov, P.V., Pybus, O.G., 2015. Evolution and Diversity of the Human Leukocyte Antigen(HLA). Evol Med Public Health 2015, 1. https://doi.org/10.1093/emph/eou033
Marx, P.A., Li, Y., Lerche, N.W., Sutjipto, S., Gettie, A., Yee, J.A., Brotman, B.H., Prince, A.M., Hanson, A., Webster, R.G., 1991. Isolation of a simian immunodeficiency virus related to human immunodeficiency virus type 2 from a west African pet sooty mangabey. J. Virol. 65, 4480–4485.
Matsumura, M., Fremont, D., Peterson, P., Wilson, I., 1992. Emerging principles for the recognition of peptide antigens by MHC class I molecules. Science 257, 927–934. https://doi.org/10.1126/science.1323878
Matthews, P.C., Prendergast, A., Leslie, A., Crawford, H., Payne, R., Rousseau, C., Rolland, M., Honeyborne, I., Carlson, J., Kadie, C., Brander, C., Bishop, K., Mlotshwa, N., Mullins, J.I., Coovadia, H., Ndung'u, T., Walker, B.D., Heckerman, D., Goulder, P.J.R., 2008. Central role of reverting mutations in HLA associations with human immunodeficiency virus set point. J. Virol. 82, 8548–8559. https://doi.org/10.1128/JVI.00580-08
Maumus, F., Epert, A., Nogué, F., Blanc, G., 2014. Plant genomes enclose footprints of past infections by giant virus relatives. Nature Communications 5. https://doi.org/10.1038/ncomms5268
Mayor, H.D., Torikai, K., Melnick, J.L., Mandel, M., 1969. Plus and Minus Single-Stranded DNA Separately Encapsidated in Adeno-Associated Satellite Virions. Science 166, 1280–1282. https://doi.org/10.1126/science.166.3910.1280
McCandlish, D.M., 2011. Visualizing fitness landscapes. Evolution 65, 1544–1558. https://doi.org/10.1111/j.1558-5646.2011.01236.x
McCutchan, F.E., Hoelscher, M., Tovanabutra, S., Piyasirisilp, S., Sanders-Buell, E., Ramos, G., Jagodzinski, L., Polonis, V., Maboko, L., Mmbando, D., Hoffmann, O., Riedner, G., von Sonnenburg, F., Robb, M., Birx, D.L., 2005. In-depth analysis of a heterosexually acquired human immunodeficiency virus type 1 superinfection: evolution, temporal fluctuation, and intercompartment dynamics from the seronegative window period through 30 months postinfection. J. Virol. 79, 11693–11704. https://doi.org/10.1128/JVI.79.18.11693-11704.2005
McCutchan, F.E., Sankale, J.-L., M'Boup, S., Kim, B., Tovanabutra, S., Hamel, D.J., Brodine, S.K., Kanki, P.J., Birx, D.L., 2004. HIV type 1 circulating recombinant form CRF09_cpx from west Africa combines subtypes A, F, G, and may share ancestors with CRF02_AG and Z321. AIDS Res. Hum. Retroviruses 20, 819–826. https://doi.org/10.1089/0889222041725163
McCutchan, F.E., Ungar, B.L., Hegerich, P., Roberts, C.R., Fowler, A.K., Hira, S.K., Perine, P.L., Burke, D.S., 1992. Genetic analysis of HIV-1 isolates from Zambia and an expanded phylogenetic tree for HIV-1. J. Acquir. Immune Defic. Syndr. 5, 441–449.

McGarry, D.D., 2009. The metalogicon of John of Salisbury: a twelfth-century defense of the verbal and logical arts of the trivium, 1st Paul Dry Books ed. ed. Paul Dry Books, Philadelphia, Pa.

McMichael, A., 1993. Natural selection at work on the surface of virus-infected cells. Science 260, 1771–1772. https://doi.org/10.1126/science.8511585

McMichael, A.J., Campbell-Lendrum, D.H., Corvalán, C.F., Ebi, K.L., Githeko, A.K., Scheraga, J.D., Woodward, A. (Eds.), 2003. Climate change and human health: risks and responses. World Health Organization, Geneva.

Meints, R.H., Ivey, R.G., Lee, A.M., Choi, T.-J., 2007. Identification of Two Virus Integration Sites in the Brown Alga Feldmannia Chromosome. Journal of Virology 82, 1407–1413. https://doi.org/10.1128/jvi.01983-07

Meirmans, P.G., 2012. The trouble with isolation by distance: NEWS AND VIEWS: OPINION. Molecular Ecology 21, 2839–2846. https://doi.org/10.1111/j.1365-294X.2012.05578.x

Meloni, S.T., Kim, B., Sankalé, J.-L., Hamel, D.J., Tovanabutra, S., Mboup, S., McCutchan, F.E., Kanki, P.J., 2004. Distinct human immunodeficiency virus type 1 subtype A virus circulating in West Africa: sub-subtype A3. J. Virol. 78, 12438–12445. https://doi.org/10.1128/JVI.78.22.12438-12445.2004

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics 21, 1087–1092. https://doi.org/10.1063/1.1699114

Meyer, D., C Aguiar, V.R., Bitarello, B.D., C Brandt, D.Y., Nunes, K., 2018. A genomic perspective on HLA evolution. Immunogenetics 70, 5–27. https://doi.org/10.1007/s00251-017-1017-3

Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prufer, K., de Filippo, C., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science 338, 222-226.

Mi, S., Lee, X., Li, X., Veldman, G.M., Finnerty, H., Racie, L., LaVallie, E., Tang, X.Y., Edouard, P., Howes, S., Keith, J.C., McCoy, J.M., 2000. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403, 785–789. https://doi.org/10.1038/35001608

Min Jou, W., Haegeman, G., Ysebaert, M., Fiers, W., 1972. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. Nature 237, 82–88. https://doi.org/10.1038/237082a0

Miyamoto, M.M., Fitch, W.M., 1995. Testing the covarion hypothesis of molecular evolution. Mol Biol Evol 12, 503–513. https://doi.org/10.1093/oxfordjournals.molbev.a040224

Mokili, J.L.K., Rogers, M., Carr, J.K., Simmonds, P., Bopopi, J.M., Foley, B.T., Korber, B.T., Birx, D.L., McCutchan, F.E., 2002. Identification of a novel clade of human immunodeficiency virus type 1 in Democratic Republic of Congo. AIDS Res. Hum. Retroviruses 18, 817–823. https://doi.org/10.1089/08892220260139567

Mónaco, D.C., Dilernia, D.A., Fiore-Gartland, A., Yu, T., Prince, J.L., Dennis, K.K.,

Qin, K., Schaefer, M., Claiborne, D.T., Kilembe, W., Tang, J., Price, M.A., Farmer, P., Gilmour, J., Bansal, A., Allen, S., Goepfert, P., Hunter, E., 2016. Balance between transmitted HLA preadapted and nonassociated polymorphisms is a major determinant of HIV-1 disease progression. J. Exp. Med. 213, 2049–2063. https://doi.org/10.1084/jem.20151984

Montavon, C., Bibollet-Ruche, F., Robertson, D., Koumare, B., Mulanga, C., Esu-Williams, E., Toure, C., Mboup, S., Saman, E., Delaporte, E., Peeters, M., 1999. The identification of a complex A/G/I/J recombinant HIV type 1 virus in various West African countries. AIDS Res. Hum. Retroviruses 15, 1707–1712. https://doi.org/10.1089/088922299309757

Montavon, C., Delaporte, E., Peeters, M., 2000. Two new complete genomes of HIV-1 recombinant AGJ BFP90-like circulating in Mali and Senegal. Unpublished direct submission to Genbank.

Montavon, C., Toure-Kane, C., Liegeois, F., Mpoudi, E., Bourgeois, A., Vergne, L., Perret, J.L., Boumah, A., Saman, E., Mboup, S., Delaporte, E., Peeters, M., 2000. Most env and gag subtype A HIV-1 viruses circulating in West and West Central Africa are similar to the prototype AG recombinant virus IBNG. J. Acquir. Immune Defic. Syndr. 23, 363–374. https://doi.org/10.1097/00126334-200004150-00001

Montavon, C., Toure-Kane, C., Nkengasong, J.N., Vergne, L., Hertogs, K., Mboup, S., Delaporte, E., Peeters, M., 2002a. CRF06-cpx: a new circulating recombinant form of HIV-1 in West Africa involving subtypes A, G, K, and J. J. Acquir. Immune Defic. Syndr. 29, 522–530. https://doi.org/10.1097/00126334-200204150-00014

Montavon, C., Vergne, L., Bourgeois, A., Mpoudi-Ngole, E., Malonga-Mouellet, G., Butel, C., Toure-Kane, C., Delaporte, E., Peeters, M., 2002b. Identification of a new circulating recombinant form of HIV type 1, CRF11-cpx, involving subtypes A, G, J, and CRF01-AE, in Central Africa. AIDS Res. Hum. Retroviruses 18, 231–236. https://doi.org/10.1089/08892220252781301

Moreno-Mayar, J.V., Vinner, L., de Barros Damgaard, P., de la Fuente, C., Chan, J., Spence, J.P., Allentoft, M.E., Vimala, T., Racimo, F., Pinotti, T., et al. (2018). Early human dispersals within the Americas. Science 362.

Morissette, G., Flamand, L., 2010. Herpesviruses and Chromosomal Integration. Journal of Virology 84, 12100–12109. https://doi.org/10.1128/jvi.01169-10

Morse, S.S., 1995. Factors in the emergence of infectious diseases. Emerging Infect. Dis. 1, 7–15. https://doi.org/10.3201/eid0101.950102

Mossel, E., Steel, M., 2005. How much can evolved characters tell us about the tree that generated them?, in: Mathematics of Evolution and Phylogeny. Oxford University Press, Oxford.

Muenchhoff, M., Madurai, S., Hempenstall, A.J., Adland, E., Carlqvist, A., Moonsamy, A., Jaggernath, M., Mlotshwa, B., Siboto, E., Ndung'u, T., Goulder, P.J.R., 2014. Evaluation of the NucliSens EasyQ v2.0 assay in comparison with the Roche Amplicor v1.5 and the Roche CAP/CTM HIV-1

Test v2.0 in quantification of C-clade HIV-1 in plasma. PLoS ONE 9, e103983. https://doi.org/10.1371/journal.pone.0103983

Muller, H.J., 1964. The relation of recombination to mutational advance. Mutat. Res. 106, 2–9. https://doi.org/10.1016/0027-5107(64)90047-8

Mullins, C.S., and Linnebacher, M. (2012). Endogenous retrovirus sequences as a novel class of tumor-specific antigens: an example of HERV-H env encoding strong CTL epitopes. Cancer Immunol Immunother 61, 1093-1100.

Murphy, E., Korber, B., Georges-Courbot, M.C., You, B., Pinter, A., Cook, D., Kieny, M.P., Georges, A., Mathiot, C., Barré-Sinoussi, F., 1993. Diversity of V3 region sequences of human immunodeficiency viruses type 1 from the central African Republic. AIDS Res. Hum. Retroviruses 9, 997–1006. https://doi.org/10.1089/aid.1993.9.997

Mushegian, A., Shipunov, A., Elena, S.F., 2016. Changes in the composition of the RNA virome mark evolutionary transitions in green plants. BMC Biology 14. https://doi.org/10.1186/s12915-016-0288-8

Mustonen, V., Lässig, M., 2009. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. Trends Genet. 25, 111–119. https://doi.org/10.1016/j.tig.2009.01.002

Myers, G., 1994. Tenth anniversary perspectives on AIDS. HIV: between past and future. AIDS Res Hum Retroviruses 10, 1317–1324. https://doi.org/10.1089/aid.1994.10.1317

Myers, G., Korber, B., Berzofsky, J.A., Pavlakis, G., Smith, R.F. (Eds.), 1992. Human retroviruses and aids, 1992. A compilation and analysis of nucleic acid and amino acid sequences.

Natsch, A., 2014. A human chemosensory modality to detect peptides in the nose? Proc Biol Sci 281, 20131678. https://doi.org/10.1098/rspb.2013.1678

Ndiaye, H.D., Tchiakpe, E., Vidal, N., Ndiaye, O., Diop, A.K., Peeters, M., Mboup, S., Toure-Kane, C., 2013. HIV type 1 subtype C remains the predominant subtype in men having sex with men in Senegal. AIDS Res. Hum. Retroviruses 29, 1265–1272. https://doi.org/10.1089/aid.2013.0140

Ndung'u, T., Renjifo, B., Novitsky, V.A., McLane, M.F., Gaolekwe, S., Essex, M., 2000. Molecular cloning and biological characterization of full-length HIV-1 subtype C from Botswana. Virology 278, 390–399. https://doi.org/10.1006/viro.2000.0583

Neddermann, P., Gallinari, P., Lettieri, T., Schmid, D., Truong, O., Hsuan, J.J., Wiebauer, K., Jiricny, J., 1996. Cloning and Expression of Human G/T Mismatch-specific Thymine-DNA Glycosylase. Journal of Biological Chemistry 271, 12767–12774. https://doi.org/10.1074/jbc.271.22.12767

Neilson, J.R., John, G.C., Carr, J.K., Lewis, P., Kreiss, J.K., Jackson, S., Nduati, R.W., Mbori-Ngacha, D., Panteleeff, D.D., Bodrug, S., Giachetti, C., Bott, M.A., Richardson, B.A., Bwayo, J., Ndinya-Achola, J., Overbaugh, J., 1999. Subtypes of human immunodeficiency virus type 1 and disease stage among women in Nairobi, Kenya. J. Virol. 73, 4393–4403.

Niama, F.R., Toure-Kane, C., Vidal, N., Obengui, P., Bikandou, B., Ndoundou Nkodia, M.Y., Montavon, C., Diop-Ndiaye, H., Mombouli, J.V., Mokondzimobe, E., Diallo, A.G., Delaporte, E., Parra, H.-J., Peeters, M., Mboup, S., 2006. HIV-1 subtypes and recombinants in the Republic of Congo. Infect. Genet. Evol. 6, 337–343. https://doi.org/10.1016/j.meegid.2005.12.001

Niama, F.R., Vidal, N., Bazepeo, S.E., Mpoudi, E., Toure-Kane, C., Parra, H.J., Delaporte, E., Peeters, M., 2009. CRF45_AKU, a circulating recombinant from Central Africa, is probably the common ancestor of HIV type 1 MAL and HIV type 1 NOGIL. AIDS Res. Hum. Retroviruses 25, 1345–1353. https://doi.org/10.1089/aid.2009.0169

Njai, H.F., Gali, Y., Vanham, G., Clybergh, C., Jennes, W., Vidal, N., Butel, C., Mpoudi-Ngolle, E., Peeters, M., Ariën, K.K., 2006. The predominance of Human Immunodeficiency Virus type 1 (HIV-1) circulating recombinant form 02 (CRF02_AG) in West Central Africa may be related to its replicative fitness. Retrovirology 3, 40. https://doi.org/10.1186/1742-4690-3-40

Novelli, P., Vella, C., Oxford, J., Daniels, R.S., 2002. Construction and characterization of a full-length HIV-1(92UG001) subtype D infectious molecular clone. AIDS Res. Hum. Retroviruses 18, 85–88. https://doi.org/10.1089/088922202753394754

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., Stephens, M., Bustamante, C.D., 2008. Genes mirror geography within Europe. Nature 456, 98–101. https://doi.org/10.1038/nature07331

Novitsky, V., Wang, R., Margolin, L., Baca, J., Kebaabetswe, L., Rossenkhan, R., Bonney, C., Herzig, M., Nkwe, D., Moyo, S., Musonda, R., Woldegabriel, E., van Widenfelt, E., Makhema, J., Lagakos, S., Essex, M., 2009. Timing constraints of in vivo gag mutations during primary HIV-1 subtype C infection. PLoS ONE 4, e7727. https://doi.org/10.1371/journal.pone.0007727

Novitsky, V., Wang, R., Rossenkhan, R., Moyo, S., Essex, M., 2013. Intra-host evolutionary rates in HIV-1C env and gag during primary infection. Infect. Genet. Evol. 19, 361–368. https://doi.org/10.1016/j.meegid.2013.02.023

Novitsky, V.A., Gaolekwe, S., McLane, M.F., Ndung'u, T.P., Foley, B.T., Vannberg, F., Marlink, R., Essex, M., 2000. HIV type 1 A/J recombinant with a pronounced pol gene mosaicism. AIDS Res. Hum. Retroviruses 16, 1015–1020. https://doi.org/10.1089/08892220050058434

Novitsky, V.A., Montano, M.A., McLane, M.F., Renjifo, B., Vannberg, F., Foley, B.T., Ndung'u, T.P., Rahman, M., Makhema, M.J., Marlink, R., Essex, M., 1999. Molecular cloning and phylogenetic analysis of human immunodeficiency virus type 1 subtype C: a set of 23 full-length clones from Botswana. J. Virol. 73, 4427–4432.

Novitsky,V., 2002. HIV-1 subtype C/Denv/C recombinant from Botswana. Unpublished direct submission to Genbank.

Ntale, R.S., Chopera, D.R., Ngandu, N.K., Assis de Rosa, D., Zembe, L., Gamieldien,

H., Mlotshwa, M., Werner, L., Woodman, Z., Mlisana, K., Abdool Karim, S., Gray, C.M., Williamson, C., CAPRISA 002 Study Team, 2012. Temporal association of HLA-B*81:01- and HLA-B*39:10-mediated HIV-1 p24 sequence evolution with disease progression. J. Virol. 86, 12013–12024. https://doi.org/10.1128/JVI.00539-12

Obuku, A.E., Bugembe, D.L., Musinguzi, K., Watera, C., Serwanga, J., Ndembi, N., Levin, J., Kaleebu, P., Pala, P., 2016. Macrophage Inflammatory Protein-1 Beta and Interferon Gamma Responses in Ugandans with HIV-1 Acute/Early Infections. AIDS Res. Hum. Retroviruses 32, 237–246. https://doi.org/10.1089/AID.2015.0157

Of dialects, armies and navies, 2010. . The Economist.

Oh, S., Yoo, D., Liu, W.-T., 2016. Metagenomics Reveals a Novel Virophage Population in a Tibetan Mountain Lake. Microbes and Environments 31, 173–177. https://doi.org/10.1264/jsme2.me16003

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H., 2020. vegan: Community Ecology Package.

Olaleye, D.O., Sheng, Z., Howard, T.M., Rasheed, S., 1996. Isolation and characterization of a new subtype A variant of human immunodeficiency virus type I from Nigeria. Trop. Med. Int. Health 1, 97–106. https://doi.org/10.1046/j.1365-3156.1996.d01-4.x

Op de Coul, E., van der Schoot, A., Goudsmit, J., van den Burg, R., Janssens, W., Heyndrickx, L., van der Groen, G., Cornelissen, M., 2000. Independent introduction of transmissible F/D recombinant HIV-1 from Africa into Belgium and The Netherlands. Virology 270, 267–277. https://doi.org/10.1006/viro.2000.0247

Oram, J.D., Downing, R.G., Roff, M., Clegg, J.C., Serwadda, D., Carswell, J.W., 1990. Nucleotide sequence of a Ugandan HIV-1 provirus reveals genetic diversity from other HIV-1 isolates. AIDS Res. Hum. Retroviruses 6, 1073–1078. https://doi.org/10.1089/aid.1990.6.1073

Ospina, R., Ferrari, S.L.P., 2012. A general class of zero-or-one inflated beta regression models. Computational Statistics & Data Analysis 56, 1609–1623. https://doi.org/10.1016/j.csda.2011.10.005

Owen, J.A., Punt, J., Stranford, S.A., Jones, P.P., Kuby, J., 2013. Kuby immunology, 7. ed., international ed., 1. print. ed. Palgrave Macmillan, New York, NY.

Padgham, M., Sumner, M.D., 2021. geodist: Fast, Dependency-Free Geodesic Distance Calculations.

Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N.N., Kyrpides, N.C., 2016. Uncovering Earth's virome. Nature 536, 425–430. https://doi.org/10.1038/nature19094

Pagani, L., Kivisild, T., Tarekegn, A., Ekong, R., Plaster, C., Gallego Romero, I., Ayub, Q., Mehdi, S.Q., Thomas, M.G., Luiselli, D., Bekele, E., Bradman, N.,

Balding, D.J., Tyler-Smith, C., 2012. Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. Am. J. Hum. Genet. 91, 83–96. https://doi.org/10.1016/j.ajhg.2012.05.015

Palm, A.A., Esbjörnsson, J., Månsson, F., Biague, A., da Silva, Z.J., Norrgren, H., Jansson, M., Medstrand, P., 2015. Cocirculation of several similar but unique HIV-1 recombinant forms in Guinea-Bissau revealed by near full-length genomic sequencing. AIDS Res. Hum. Retroviruses 31, 938–945. https://doi.org/10.1089/AID.2015.0073

Pamer, E., Cresswell, P., 1998. Mechanisms of MHC class I–restricted antigen processing. Annu Rev Immunol 16, 323–358. https://doi.org/10.1146/annurev.immunol.16.1.323

Pandrea, I., Robertson, D.L., Onanga, R., Gao, F., Makuwa, M., Ngari, P., Bedjabaga, I., Roques, P., Simon, F., Apetrei, C., 2002. Analysis of partial pol and env sequences indicates a high prevalence of HIV type 1 recombinant strains circulating in Gabon. AIDS Res. Hum. Retroviruses 18, 1103–1116. https://doi.org/10.1089/088922202320567842

Papathanasopoulos, M.A., Cilliers, T., Morris, L., Mokili, J.L., Dowling, W., Birx, D.L., McCutchan, F.E., 2002. Full-length genome analysis of HIV-1 subtype C utilizing CXCR4 and intersubtype recombinants isolated in South Africa. AIDS Res. Hum. Retroviruses 18, 879–886. https://doi.org/10.1089/08892220260190362

Papathanasopoulos, M.A., Patience, T., Meyers, T.M., McCutchan, F.E., Morris, L., 2003. Full-length genome characterization of HIV type 1 subtype C isolates from two slow-progressing perinatally infected siblings in South Africa. AIDS Res. Hum. Retroviruses 19, 1033–1037. https://doi.org/10.1089/088922203322588396

Paradis, E., Claude, J., Strimmer, K., 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20, 289–290. https://doi.org/10.1093/bioinformatics/btg412

Parham, P. (1988). Function and polymorphism of human leukocyte antigen-A,B,C molecules. Am J Med 85, 2-5.

Parham, P., and Guethlein, L.A. (2018). Genetics of Natural Killer Cells in Human Health, Disease, and Survival. Annu Rev Immunol 36, 519-548.

Parham, P., Arnett, K.L., Adams, E.J., Little, A.M., Tees, K., Barber, L.D., Marsh, S.G., Ohta, T., Markow, T., and Petzl-Erler, M.L. (1997). Episodic evolution and turnover of HLA-B in the indigenous human populations of the Americas. Tissue antigens 50, 219-232.

Parrish, N.F., Gao, F., Li, H., Giorgi, E.E., Barbian, H.J., Parrish, E.H., Zajic, L., Iyer, S.S., Decker, J.M., Kumar, A., Hora, B., Berg, A., Cai, F., Hopper, J., Denny, T.N., Ding, H., Ochsenbauer, C., Kappes, J.C., Galimidi, R.P., West, A.P., Bjorkman, P.J., Wilen, C.B., Doms, R.W., O'Brien, M., Bhardwaj, N., Borrow, P., Haynes, B.F., Muldoon, M., Theiler, J.P., Korber, B., Shaw, G.M., Hahn, B.H., 2013. Phenotypic properties of transmitted founder HIV-1. Proc. Natl.

Acad. Sci. U.S.A. 110, 6626–6633. https://doi.org/10.1073/pnas.1304288110
Parrish, N.F., Wilen, C.B., Banks, L.B., Iyer, S.S., Pfaff, J.M., Salazar-Gonzalez, J.F., Salazar, M.G., Decker, J.M., Parrish, E.H., Berg, A., Hopper, J., Hora, B., Kumar, A., Mahlokozera, T., Yuan, S., Coleman, C., Vermeulen, M., Ding, H., Ochsenbauer, C., Tilton, J.C., Permar, S.R., Kappes, J.C., Betts, M.R., Busch, M.P., Gao, F., Montefiori, D., Haynes, B.F., Shaw, G.M., Hahn, B.H., Doms, R.W., 2012. Transmitted/founder and chronic subtype C HIV-1 use CD4 and CCR5 receptors with equal efficiency and are not inhibited by blocking the integrin $\alpha 4 \beta 7$. PLoS Pathog. 8, e1002686. https://doi.org/10.1371/journal.ppat.1002686
Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G.H., Barreiro, L.B., Froment, A., Heyer, E., Massougbodji, A., Fortes-Lima, C., Migot-Nabias, F., Bellis, G., Dugoujon, J.-M., Pereira, J.B., Fernandes, V., Pereira, L., Van der Veen, L., Mouguiama-Daouda, P., Bustamante, C.D., Hombert, J.-M., Quintana-Murci, L., 2017. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. Science 356, 543–546. https://doi.org/10.1126/science.aal1988
Payne, R., Muenchhoff, M., Mann, J., Roberts, H.E., Matthews, P., Adland, E., Hempenstall, A., Huang, K.-H., Brockman, M., Brumme, Z., Sinclair, M., Miura, T., Frater, J., Essex, M., Shapiro, R., Walker, B.D., Ndung'u, T., McLean, A.R., Carlson, J.M., Goulder, P.J.R., 2014. Impact of HLA-driven HIV adaptation on virulence in populations of high HIV seroprevalence. Proc. Natl. Acad. Sci. U.S.A. 111, E5393-5400. https://doi.org/10.1073/pnas.1413339111
Peeters, M., Liegeois, F., Torimiro, N., Bourgeois, A., Mpoudi, E., Vergne, L., Saman, E., Delaporte, E., Saragosti, S., 1999. Characterization of a highly replicative intergroup M/O human immunodeficiency virus type 1 recombinant isolated from a Cameroonian patient. J. Virol. 73, 7368–7375.
Peeters, M., Sharp, P.M., 2000. Genetic diversity of HIV-1: the moving target. AIDS 14 Suppl 3, S129-140.
Pendry, J.B., 2000. Negative Refraction Makes a Perfect Lens. Physical Review Letters 85, 3966–3969. https://doi.org/10.1103/physrevlett.85.3966
Pepin, J., 2011. The Origins of AIDS. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139005234
Permar, S.R., Salazar, M.G., Gao, F., Cai, F., Learn, G.H., Kalilani, L., Hahn, B.H., Shaw, G.M., Salazar-Gonzalez, J.F., 2013. Clonal amplification and maternal-infant transmission of nevirapine-resistant HIV-1 variants in breast milk following single-dose nevirapine prophylaxis. Retrovirology 10, 88. https://doi.org/10.1186/1742-4690-10-88
Peskun, P.H., 1973. Optimum Monte-Carlo sampling using Markov chains. Biometrika 60, 607–612. https://doi.org/10.1093/biomet/60.3.607
Pessôa, R., Watanabe, J.T., Calabria, P., Alencar, C.S., Loureiro, P., Lopes, M.E., Proetti, A.B., Félix, A.C., Sabino, E.C., Busch, M.P., Sanabani, S.S.,

International Component of the NHLBI Recipient Epidemiology and Donor Evaluation Study-III (REDS-III), 2015. Enhanced detection of viral diversity using partial and near full-length genomes of human immunodeficiency virus Type 1 provirus deep sequencing data from recently infected donors at four blood centers in Brazil. Transfusion 55, 980–990. https://doi.org/10.1111/trf.12936

Peters, H.O., Mendoza, M.G., Capina, R.E., Luo, M., Mao, X., Gubbins, M., Nagelkerke, N.J.D., Macarthur, I., Sheardown, B.B., Kimani, J., Wachihi, C., Thavaneswaran, S., Plummer, F.A., 2008. An integrative bioinformatic approach for studying escape mutations in human immunodeficiency virus type 1 gag in the Pumwani Sex Worker Cohort. J. Virol. 82, 1980–1992. https://doi.org/10.1128/JVI.02742-06

Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T.J., Manuel, M., Wörheide, G., Baurain, D., 2011. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLoS Biology 9, e1000602. https://doi.org/10.1371/journal.pbio.1000602

Philippe, H., Douady, C.J., 2003. Horizontal gene transfer and phylogenetics. Curr Opin Microbiol 6, 498–505. https://doi.org/10.1016/j.mib.2003.09.008

Philippe, N., Legendre, M., Doutre, G., Coute, Y., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L., Bruley, C., Garin, J., Claverie, J.-M., Abergel, C., 2013. Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes. Science 341, 281–286. https://doi.org/10.1126/science.1239181

Pickett-Heaps, J.D., 1972. A possible virus infection in the green alga oedogonium. Journal of Phycology 8, 44–47. https://doi.org/10.1111/j.1529-8817.1972.tb03999.x

Pierini, F., and Lenz, T.L. (2018). Divergent allele advantage at human MHC genes: signatures of past and ongoing selection. Mol Biol Evol.

Plantier, J.-C., Leoz, M., Dickerson, J.E., De Oliveira, F., Cordonnier, F., Lemée, V., Damond, F., Robertson, D.L., Simon, F., 2009. A new human immunodeficiency virus derived from gorillas. Nat. Med. 15, 871–872. https://doi.org/10.1038/nm.2016

Pollakis, G., Abebe, A., Kliphuis, A., De Wit, T.F.R., Fisseha, B., Tegbaru, B., Tesfaye, G., Negassa, H., Mengistu, Y., Fontanet, A.L., Cornelissen, M., Goudsmit, J., 2003. Recombination of HIV type 1C (C'/C") in Ethiopia: possible link of EthHIV-1C' to subtype C sequences from the high-prevalence epidemics in India and Southern Africa. AIDS Res. Hum. Retroviruses 19, 999–1008. https://doi.org/10.1089/088922203322588350

Poss, M., Gosink, J., Thomas, E., Kreiss, J.K., Ndinya-Achola, J., Mandaliya, K., Bwayo, J., Overbaugh, J., 1997. Phylogenetic evaluation of Kenyan HIV type 1 isolates. AIDS Res. Hum. Retroviruses 13, 493–499. https://doi.org/10.1089/aid.1997.13.493

Poss, M., Overbaugh, J., 1999. Variants from the diverse virus population identified at

seroconversion of a clade A human immunodeficiency virus type 1-infected woman have distinct biological properties. J. Virol. 73, 5255–5264.
Poss, M., Rodrigo, A.G., Gosink, J.J., Learn, G.H., de Vange Panteleeff, D., Martin, H.L., Bwayo, J., Kreiss, J.K., Overbaugh, J., 1998. Evolution of envelope sequences from the genital tract and peripheral blood of women infected with clade A human immunodeficiency virus type 1. J. Virol. 72, 8240–8251.
Posth, C., Nakatsuka, N., Lazaridis, I., Skoglund, P., Mallick, S., Lamnidis, T.C., Rohland, N., Nagele, K., Adamski, N., Bertolini, E., et al. (2018). Reconstructing the Deep Population History of Central and South America. Cell 175, 1185-1197 e1122.
Powell, R.L.R., Zhao, J., Konings, F.A.J., Tang, S., Ewane, L., Burda, S., Urbanski, M.M., Saa, D.R., Hewlett, I., Nyambi, P.N., 2007a. Circulating recombinant form (CRF) 37_cpx: an old strain in Cameroon composed of diverse, genetically distant lineages of subtypes A and G. AIDS Res. Hum. Retroviruses 23, 923–933. https://doi.org/10.1089/aid.2007.0040
Powell, R.L.R., Zhao, J., Konings, F.A.J., Tang, S., Nanfack, A., Burda, S., Urbanski, M.M., Saa, D.R., Hewlett, I., Nyambi, P.N., 2007b. Identification of a novel circulating recombinant form (CRF) 36_cpx in Cameroon that combines two CRFs (01_AE and 02_AG) with ancestral lineages of subtypes A and G. AIDS Res. Hum. Retroviruses 23, 1008–1019. https://doi.org/10.1089/aid.2006.0289
Price, M.A., Wallis, C.L., Lakhi, S., Karita, E., Kamali, A., Anzala, O., Sanders, E.J., Bekker, L.-G., Twesigye, R., Hunter, E., Kaleebu, P., Kayitenkore, K., Allen, S., Ruzagira, E., Mwangome, M., Mutua, G., Amornkul, P.N., Stevens, G., Pond, S.L.K., Schaefer, M., Papathanasopoulos, M.A., Stevens, W., Gilmour, J., IAVI Early Infection Cohort Study Group, 2011. Transmitted HIV type 1 drug resistance among individuals with recent HIV infection in East and Southern Africa. AIDS Res. Hum. Retroviruses 27, 5–12. https://doi.org/10.1089/aid.2010.0030
Pritchard, L.K., Vasiljevic, S., Ozorowski, G., Seabright, G.E., Cupo, A., Ringe, R., Kim, H.J., Sanders, R.W., Doores, K.J., Burton, D.R., Wilson, I.A., Ward, A.B., Moore, J.P., Crispin, M., 2015. Structural Constraints Determine the Glycosylation of HIV-1 Envelope Trimers. Cell Rep. 11, 1604–1613. https://doi.org/10.1016/j.celrep.2015.05.017
Prufer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505, 43-49.
Prugnolle, F., Manica, A., Charpentier, M., Guégan, J.F., Guernier, V., Balloux, F., 2005. Pathogen-driven selection and worldwide HLA class I diversity. Curr Biol 15, 1022–1027. https://doi.org/10.1016/j.cub.2005.04.050
Quach, H., Quintana-Murci, L., 2017. Living in an adaptive world: Genomic dissection of the genus Homo and its immune response. Journal of Experimental Medicine 214, 877–894. https://doi.org/10.1084/jem.20161942

Quintana-Murci, L. (2019). Human Immunology through the Lens of Evolutionary Genetics. Cell 177, 184-199.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Raghavan, M., Skoglund, P., Graf, K.E., Metspalu, M., Albrechtsen, A., Moltke, I., Rasmussen, S., Stafford, T.W., Jr., Orlando, L., Metspalu, E., et al. (2014). Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 505, 87-91.

Raghwani, J., Redd, A.D., Longosz, A.F., Wu, C.-H., Serwadda, D., Martens, C., Kagaayi, J., Sewankambo, N., Porcella, S.F., Grabowski, M.K., Quinn, T.C., Eller, M.A., Eller, L.A., Wabwire-Mangen, F., Robb, M.L., Fraser, C., Lythgoe, K.A., 2018. Evolution of HIV-1 within untreated individuals and at the population scale in Uganda. PLOS Pathogens 14, e1007167. https://doi.org/10.1371/journal.ppat.1007167

Ragupathy,V., Voskanian-Kordi, A., Karagiannis, K., Wang, A., Zhao, J., Mbondji, C., Gao, F., Hora, B., Sanchez, A., Denny,T., Schito, M., Busch, M., Simonyan, V., Hewlett,I., 2016. PCR amplicon free whole genome charecterization of Human immunodeficiency virus type -1 genotypes. Unpublished direct submission to Genbank.

Rajapaksa, U.S., Li, D., Peng, Y.-C., McMichael, A.J., Dong, T., Xu, X.-N., 2012. HLA-B may be more protective against HIV-1 than HLA-A because it resists negative regulatory factor (Nef) mediated down-regulation. Proceedings of the National Academy of Sciences 109, 13353–13358. https://doi.org/10.1073/pnas.1204199109

Rakoff-Nahoum, S., Kuebler, P.J., Heymann, J.J., M, E.S., Ortiz, G.M., G, S.O., Barbour, J.D., Lenz, J., Steinfeld, A.D., and Nixon, D.F. (2006). Detection of T lymphocytes specific for human endogenous retrovirus K (HERV-K) in patients with seminoma. AIDS research and human retroviruses 22, 52-56.

Rambaut, A., Posada, D., Crandall, K.A., Holmes, E.C., 2004. The causes and consequences of HIV evolution. Nature Reviews Genetics 5, 52–61. https://doi.org/10.1038/nrg1246

Rambaut, A., Robertson, D.L., Pybus, O.G., Peeters, M., Holmes, E.C., 2001. Human immunodeficiency virus. Phylogeny and the origin of HIV-1. Nature 410, 1047–1048. https://doi.org/10.1038/35074179

Rannala, B., Yang, Z., 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. Journal of Molecular Evolution 43, 304–311. https://doi.org/10.1007/BF02338839

Raoult, D., 2013. TRUC or the Need for a New Microbial Classification. Intervirology 56, 349–353. https://doi.org/10.1159/000354269

Read, B.A., Kegel, J., Klute, M.J., Kuo, A., Lefebvre, S.C., Maumus, F., Mayer, C., Miller, J., Monier, A., Salamov, A., Young, J., Aguilar, M., Claverie, J.-M., Frickenhaus, S., Gonzalez, K., Herman, E.K., Lin, Y.-C., Napier, J., Ogata, H., Sarno, A.F., Shmutz, J., Schroeder, D., de Vargas, C., Verret, F., von Dassow,

P., Valentin, K., Van de Peer, Y., Wheeler, G., Dacks, J.B., Delwiche, C.F., Dyhrman, S.T., Glöckner, G., John, U., Richards, T., Worden, A.Z., Zhang, X., Grigoriev, I.V., 2013. Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature 499, 209–213. https://doi.org/10.1038/nature12221

Reche, P.A., and Reinherz, E.L. (2003). Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms. Journal of molecular biology 331, 623-641.

Reich, D. (2018). Who we are and how we got here (Oxford University Press).

Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L.F., Maricic, T., Good, J.M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E.E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M.V., Derevianko, A.P., Hublin, J.-J., Kelso, J., Slatkin, M., Pääbo, S., 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468, 1053–1060. https://doi.org/10.1038/nature09710

René, B., Mauffret, O., Fossé, P., 2018. Retroviral nucleocapsid proteins and DNA strand transfers. Biochim. Open 7, 10–25. https://doi.org/10.1016/j.biopen.2018.07.001

Reteno, D.G., Benamar, S., Khalil, J.B., Andreani, J., Armstrong, N., Klose, T., Rossmann, M., Colson, P., Raoult, D., La Scola, B., 2015. Faustovirus, an Asfarvirus-Related New Lineage of Giant Viruses Infecting Amoebae. Journal of Virology 89, 6585–6594. https://doi.org/10.1128/jvi.00115-15

Rife, B.D., Mavian, C., Chen, X., Ciccozzi, M., Salemi, M., Min, J., Prosperi, M.C., 2017. Phylodynamic applications in 21st century global infectious disease research. Global Health Research and Policy 2. https://doi.org/10.1186/s41256-017-0034-y

Riley, E., Olerup, O., 1992. HLA polymorphisms and evolution. Immunology Today 13, 333–335. https://doi.org/10.1016/0167-5699(92)90165-4

Robertson, D.L., Anderson, J.P., Bradac, J.A., Carr, J.K., Foley, B., Funkhouser, R.K., Gao, F., Hahn, B.H., Kalish, M.L., Kuiken, C., Learn, G.H., Leitner, T., McCutchan, F., Osmanov, S., Peeters, M., Pieniazek, D., Salminen, M., Sharp, P.M., Wolinsky, S., Korber, B., 2000. HIV-1 nomenclature proposal. Science 288, 55–56. https://doi.org/10.1126/science.288.5463.55d

Robinson, J., Guethlein, L.A., Cereb, N., Yang, S.Y., Norman, P.J., Marsh, S.G.E., Parham, P., 2017. Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles. PLOS Genetics 13, e1006862. https://doi.org/10.1371/journal.pgen.1006862

Robinson, J., Halliwell, J.A., Hayhurst, J.D., Flicek, P., Parham, P., Marsh, S.G.E., 2015. The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res 43, D423-431. https://doi.org/10.1093/nar/gku1161

Rodenburg, C.M., Li, Y., Trask, S.A., Chen, Y., Decker, J., Robertson, D.L., Kalish, M.L., Shaw, G.M., Allen, S., Hahn, B.H., Gao, F., UNAIDS and NIAID

Networks for HIV Isolation and Characterization, 2001. Near full-length clones and reference sequences for subtype C isolates of HIV type 1 from three different continents. AIDS Res. Hum. Retroviruses 17, 161–168. https://doi.org/10.1089/08892220150217247

Rodgers, M. A., Vallari, A.S., Harris, B., Yamaguchi, J., Holzmayer, V., Forberg, K., Berg, M.G., Kenmenge, J., Ngansop, C., Awazi, B., Mbanya, D., Kaptue, L., Brennan, C., Cloherty, G., Ndembi, N., 2017. Identification of rare HIV-1 Group N, HBV AE, and HTLV-3 strains in rural South Cameroon. Virology 504, 141–151. https://doi.org/10.1016/j.virol.2017.01.008

Rodgers, Mary A., Wilkinson, E., Vallari, A., McArthur, C., Sthreshley, L., Brennan, C.A., Cloherty, G., de Oliveira, T., 2017. Sensitive Next-Generation Sequencing Method Reveals Deep Genetic Diversity of HIV-1 in the Democratic Republic of the Congo. J. Virol. 91. https://doi.org/10.1128/JVI.01841-16

Rolland, M., Carlson, J.M., Manocheewa, S., Swain, J.V., Lanxon-Cookson, E., Deng, W., Rousseau, C.M., Raugi, D.N., Learn, G.H., Maust, B.S., Coovadia, H., Ndung'u, T., Goulder, P.J.R., Walker, B.D., Brander, C., Heckerman, D.E., Mullins, J.I., 2010. Amino-acid co-variation in HIV-1 Gag subtype C: HLA-mediated selection pressure and compensatory dynamics. PloS One 5. https://doi.org/10.1371/journal.pone.0012463

Roques, P., Robertson, D.L., Souquière, S., Apetrei, C., Nerrienet, E., Barré-Sinoussi, F., Müller-Trutwin, M., Simon, F., 2004. Phylogenetic characteristics of three new HIV-1 N strains and implications for the origin of group N. AIDS 18, 1371–1381. https://doi.org/10.1097/01.aids.0000125990.86904.28

Roques, P., Robertson, D.L., Souquière, S., Damond, F., Ayouba, A., Farfara, I., Depienne, C., Nerrienet, E., Dormont, D., Brun-Vézinet, F., Simon, F., Mauclère, P., 2002. Phylogenetic Analysis of 49 Newly Derived HIV-1 Group O Strains: High Viral Diversity but No Group M-like Subtype Structure. Virology 302, 259–273. https://doi.org/10.1006/viro.2002.1430

Rose, J.A., Berns, K.I., Hoggan, M.D., Koczot, F.J., 1969. Evidence for a single-stranded adenovirus-associated virus genome: formation of a DNA density hybrid on release of viral DNA. Proceedings of the National Academy of Sciences 64, 863–869. https://doi.org/10.1073/pnas.64.3.863

Rose, J.A., Koczot, F., 1971. Adenovirus-associated virus multiplication. VI. Base compostion of the deoxyribonucleic acid strand species and strand-specific in vivo transcription. J. Virol. 8, 771–777.

Ross, J.R., 1967. Constraints on variables in syntax. Massachusetts Institute of Technology.

Rousseau, C.M., Birditt, B.A., McKay, A.R., Stoddard, J.N., Lee, T.C., McLaughlin, S., Moore, S.W., Shindo, N., Learn, G.H., Korber, B.T., Brander, C., Goulder, P.J.R., Kiepiela, P., Walker, B.D., Mullins, J.I., 2006. Large-scale amplification, cloning and sequencing of near full-length HIV-1 subtype C genomes. J. Virol. Methods 136, 118–125. https://doi.org/10.1016/j.jviromet.2006.04.009

Rycaj, K., Plummer, J.B., Yin, B., Li, M., Garza, J., Radvanyi, L., Ramondetta, L.M., Lin, K., Johanning, G.L., Tang, D.G., et al. (2015). Cytotoxicity of human endogenous retrovirus K-specific T cells toward autologous ovarian cancer cells. Clin Cancer Res 21, 471-483.

Saina, M.C., Bi, X., Lihana, R., Lwembe, R., Ishizaki, A., Panikulam, A., Palakudy, T., Musoke, R., Owens, M., Songok, E.M., Ichimura, H., 2015. Comparison of HIV-1 nef and gag Variations and Host HLA Characteristics as Determinants of Disease Progression among HIV-1 Vertically Infected Kenyan Children. PLoS ONE 10, e0137140. https://doi.org/10.1371/journal.pone.0137140

Saitou, N., 1989. A Theoretical Study of the Underestimation of Branch Lengths by the Maximum Parsimony Principle. Systematic Zoology 38, 1. https://doi.org/10.2307/2992431

Saitou, N., 2018. Introduction to Evolutionary Genomics, Computational Biology. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-92642-1

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4, 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454

Sakaguchi, S., 2005. Naturally arising Foxp3-expressing CD25+CD4+ regulatory T cells in immunological tolerance to self and non-self. Nature Immunology 6, 345–352. https://doi.org/10.1038/ni1178

Sakamoto, Y., Takekawa, N., Ampofo, W., Sata, T., Tatsumi, M., 2006. Construction and characterization of HIV-1 CRF02_AG infectious molecular clones derived from Ghanaian patients. Unpublished direct submission to Genbank.

Sakamoto,Y., Takekawa, N, Tatsumi, M., 2006a. Construction and characterization of HIV-1 subtype A and CRF01_AE infectious molecular clones. Unpublished direct submission to Genbank.

Sakamoto,Y., Tatsumi,M., Takekawa,N., 2006b. Construction and characterization of naive HIV-1 subtype A infectious molecular clones derived from Rwanda patients for reference. Unpublished direct submission to Genbank.

Salazar-Gonzalez, J.F., Salazar, M.G., Keele, B.F., Learn, G.H., Giorgi, E.E., Li, H., Decker, J.M., Wang, S., Baalwa, J., Kraus, M.H., Parrish, N.F., Shaw, K.S., Guffey, M.B., Bar, K.J., Davis, K.L., Ochsenbauer-Jambor, C., Kappes, J.C., Saag, M.S., Cohen, M.S., Mulenga, J., Derdeyn, C.A., Allen, S., Hunter, E., Markowitz, M., Hraber, P., Perelson, A.S., Bhattacharya, T., Haynes, B.F., Korber, B.T., Hahn, B.H., Shaw, G.M., 2009. Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. J. Exp. Med. 206, 1273–1289. https://doi.org/10.1084/jem.20090378

Salemi, M., Strimmer, K., Hall, W.W., Duffy, M., Delaporte, E., Mboup, S., Peeters, M., Vandamme, A.M., 2001. Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution. FASEB J. 15, 276–278.

https://doi.org/10.1096/fj.00-0449fje
Salminen, M.O., Carr, J.K., Robertson, D.L., Hegerich, P., Gotte, D., Koch, C., Sanders-Buell, E., Gao, F., Sharp, P.M., Hahn, B.H., Burke, D.S., McCutchan, F.E., 1997. Evolution and probable transmission of intersubtype recombinant human immunodeficiency virus type 1 in a Zambian couple. J. Virol. 71, 2647–2655.
Salminen, M.O., Ehrenberg, P.K., Mascola, J.R., Dayhoff, D.E., Merling, R., Blake, B., Louder, M., Hegerich, S., Polonis, V.R., Birx, D.L., Robb, M.L., McCutchan, F.E., Michael, N.L., 2000. Construction and biological characterization of infectious molecular clones of HIV-1 subtypes B and E (CRF01_AE) generated by the polymerase chain reaction. Virology 278, 103–110. https://doi.org/10.1006/viro.2000.0640
Salminen, M.O., Johansson, B., Sönnerborg, A., Ayehunie, S., Gotte, D., Leinikki, P., Burke, D.S., McCutchan, F.E., 1996. Full-length sequence of an ethiopian human immunodeficiency virus type 1 (HIV-1) isolate of genetic subtype C. AIDS Res. Hum. Retroviruses 12, 1329–1339. https://doi.org/10.1089/aid.1996.12.1329
Salmons, B., Lawson, J.S., and Gunzburg, W.H. (2014). Recent developments linking retroviruses to human breast cancer: infectious agent, enemy within or both? J Gen Virol 95, 2589-2593.
Sanchez, A.M., DeMarco, C.T., Hora, B., Keinonen, S., Chen, Y., Brinkley, C., Stone, M., Tobler, L., Keating, S., Schito, M., Busch, M.P., Gao, F., Denny, T.N., 2014. Development of a contemporary globally diverse HIV viral panel by the EQAPOL program. J. Immunol. Methods 409, 117–130. https://doi.org/10.1016/j.jim.2014.01.004
Sanchez-Mazas, A., Fernandez-Viña, M., Middleton, D., Hollenbach, J.A., Buhler, S., Di, D., Rajalingam, R., Dugoujon, J.-M., Mack, S.J., Thorsby, E., 2011. Immunogenetics as a tool in anthropological studies. Immunology 133, 143–164. https://doi.org/10.1111/j.1365-2567.2011.03438.x
Sandaa, R.-A., Heldal, M., Castberg, T., Thyrhaug, R., Bratbak, G., 2001. Isolation and Characterization of Two Viruses with Large Genome Size Infecting Chrysochromulina ericina (Prymnesiophyceae) and Pyramimonas orientalis (Prasinophyceae). Virology 290, 272–280. https://doi.org/10.1006/viro.2001.1161
Sanger, F., Brownlee, G.G., Barrell, B.G., 1965. A two-dimensional fractionation procedure for radioactive nucleotides. J Mol Biol 13, 373–398. https://doi.org/10.1016/s0022-2836(65)80104-8
Sanghvi, L.D., 1953. Comparison of genetical and morphological methods for a study of biological differences. American Journal of Physical Anthropology 11, 385–404. https://doi.org/10.1002/ajpa.1330110313
Sanjuan, R., Nebot, M.R., Chirico, N., Mansky, L.M., Belshaw, R., 2010. Viral Mutation Rates. Journal of Virology 84, 9733–9748. https://doi.org/10.1128/JVI.00694-10

Santini, S., Jeudy, S., Bartoli, J., Poirot, O., Lescot, M., Abergel, C., Barbe, V., Wommack, K.E., Noordeloos, A.A.M., Brussaard, C.P.D., Claverie, J.-M., 2013. Genome of Phaeocystis globosa virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. Proceedings of the National Academy of Sciences 110, 10800–10805. https://doi.org/10.1073/pnas.1303251110

Sarr, A.D., Eisen, G., Guèye-Ndiaye, A., Mullins, C., Traoré, I., Dia, M.C., Sankalé, J.-L., Faye, D., Mboup, S., Kanki, P., 2005. Viral dynamics of primary HIV-1 infection in Senegal, West Africa. J Infect Dis 191, 1460–1467. https://doi.org/10.1086/429409

Sauter, D., Schindler, M., Specht, A., Landford, W.N., Münch, J., Kim, K.-A., Votteler, J., Schubert, U., Bibollet-Ruche, F., Keele, B.F., Takehisa, J., Ogando, Y., Ochsenbauer, C., Kappes, J.C., Ayouba, A., Peeters, M., Learn, G.H., Shaw, G., Sharp, P.M., Bieniasz, P., Hahn, B.H., Hatziioannou, T., Kirchhoff, F., 2009. Tetherin-driven adaptation of Vpu and Nef function and the evolution of pandemic and nonpandemic HIV-1 strains. Cell Host Microbe 6, 409–421. https://doi.org/10.1016/j.chom.2009.10.004

Schiavetti, F., Thonnard, J., Colau, D., Boon, T., and Coulie, P.G. (2002). A human endogenous retroviral sequence encoding an antigen recognized on melanoma by cytolytic T lymphocytes. Cancer Res 62, 5510-5516.

Schleper, C., Kubo, K., Zillig, W., 1992. The particle SSV1 from the extremely thermophilic archaeon Sulfolobus is a virus: demonstration of infectivity and of transfection with viral DNA. Proceedings of the National Academy of Sciences 89, 7645–7649. https://doi.org/10.1073/pnas.89.16.7645

Schlub, T.E., Grimm, A.J., Smyth, R.P., Cromer, D., Chopra, A., Mallal, S., Venturi, V., Waugh, C., Mak, J., Davenport, M.P., 2014. Fifteen to twenty percent of HIV substitution mutations are associated with recombination. J. Virol. 88, 3837–3849. https://doi.org/10.1128/JVI.03136-13

Schneidewind, A., Brockman, M.A., Sidney, J., Wang, Y.E., Chen, H., Suscovich, T.J., Li, B., Adam, R.I., Allgaier, R.L., Mothé, B.R., Kuntzen, T., Oniangue-Ndza, C., Trocha, A., Yu, X.G., Brander, C., Sette, A., Walker, B.D., Allen, T.M., 2008. Structural and functional constraints limit options for cytotoxic T-lymphocyte escape in the immunodominant HLA-B27-restricted epitope in human immunodeficiency virus type 1 capsid. J. Virol. 82, 5594–5605. https://doi.org/10.1128/JVI.02356-07

Schopf, J.W., Kudryavtsev, A.B., Czaja, A.D., Tripathi, A.B., 2007. Evidence of Archean life: Stromatolites and microfossils. Precambrian Research 158, 141–155. https://doi.org/10.1016/j.precamres.2007.04.009

Schuster, F.L., Dunnebacke, T.H., 1974. Virus-like particles and an unassociated infectious agent in amoebae of the genus Naegleria. Ann Soc Belg Med Trop 54, 359–370.

Schuster, F.L., Dunnebacke, T.H., 1976. Development and release of virus-like particles in naegleria gruberi EGS. Cytobiologie 14, 131–147.

Schwartz, R., 2019. Computational Models for Cancer Phylogenetics, in: Warnow, T. (Ed.), Bioinformatics and Phylogenetics, Computational Biology. Springer International Publishing, Cham, pp. 243–275. https://doi.org/10.1007/978-3-030-10837-3_11

Scriba, T.J., de Villiers, T., Treurnicht, F.K., zur Megede, J., Barnett, S.W., Engelbrecht, S., van Rensburg, E.J., 2002. Characterization of the South African HIV type 1 subtype C complete 5' long terminal repeat, nef, and regulatory genes. AIDS Res. Hum. Retroviruses 18, 149–159. https://doi.org/10.1089/08892220252779692

Segurel, L., Thompson, E.E., Flutre, T., Lovstad, J., Venkat, A., Margulis, S.W., Moyse, J., Ross, S., Gamble, K., Sella, G., et al. (2012). The ABO blood group is a trans-species polymorphism in primates. Proceedings of the National Academy of Sciences of the United States of America 109, 18493-18498.

Serwanga, J., Mugaba, S., Pimego, E., Nanteza, B., Lyagoba, F., Nakubulwa, S., Heath, L., Nsubuga, R.N., Ndembi, N., Gotch, F., Kaleebu, P., 2012. Profile of T cell recognition of HIV type 1 consensus group M Gag and Nef peptides in a clade A1- and D-infected Ugandan population. AIDS Res. Hum. Retroviruses 28, 384–392. https://doi.org/10.1089/AID.2011.0175

Sewell, A.K., 2012. Why must T cells be cross-reactive? Nat Rev Immunol 12, 669–677. https://doi.org/10.1038/nri3279

Seymour, R.M., Allan, M.J., Pomiankowski, A., and Gustafsson, K. (2004). Evolution of the human ABO polymorphism by two complementary selective pressures. Proc Biol Sci 271, 1065-1072.

Shanks, G.D., Hu, Z., Waller, M., Lee, S.E., Terfa, D., Howard, A., van Heyningen, E., and Brundage, J.F. (2014). Measles epidemics of variable lethality in the early 20th century. Am J Epidemiol 179, 413-422.

Shannon, C.E., 1948. A mathematical theory of communication. ACM SIGMOBILE Mob. Comput. Commun. Rev. 5, 3. https://doi.org/10.1145/584091.584093

Sharp, P.M., Hahn, B.H., 2008. AIDS: prehistory of HIV-1. Nature 455, 605–606. https://doi.org/10.1038/455605a

Sherefa, K., Johansson, B., Salminen, M., Sönnerborg, A., 1998. Full-length sequence of human immunodeficiency virus type 1 subtype A, recombined with subtype C in the env V3 domain. AIDS Res. Hum. Retroviruses 14, 289–292. https://doi.org/10.1089/aid.1998.14.289

Shiina, T., Inoko, H., Kulski, J.K., 2004. An update of the HLA genomic region, locus information and disease associations: 2004. Tissue Antigens 64, 631–649. https://doi.org/10.1111/j.1399-0039.2004.00327.x

Shiina, T., Ota, M., Shimizu, S., Katsuyama, Y., Hashimoto, N., Takasu, M., Anzai, T., Kulski, J.K., Kikkawa, E., Naruse, T., Kimura, N., Yanagiya, K., Watanabe, A., Hosomichi, K., Kohara, S., Iwamoto, C., Umehara, Y., Meyer, A., Wanner, V., Sano, K., Macquin, C., Ikeo, K., Tokunaga, K., Gojobori, T., Inoko, H., Bahram, S., 2006. Rapid evolution of major histocompatibility complex class I genes in primates generates new disease alleles in humans via hitchhiking

diversity. Genetics 173, 1555–1570. https://doi.org/10.1534/genetics.106.057034
Sicko-Goad, L., Walker, G., 1979. Viroplasm and large virus-like particles in the dinoflagellateGymnodinium uberrimum. Protoplasma 99, 203–210. https://doi.org/10.1007/bf01275735
Sievert, C., 2020. Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC.
Simas, A.B., Barreto-Souza, W., Rocha, A.V., 2010. Improved estimators for a general class of beta regression models. Computational Statistics & Data Analysis 54, 348–366. https://doi.org/10.1016/j.csda.2009.08.017
Simon, F., Mauclère, P., Roques, P., Loussert-Ajaka, I., Müller-Trutwin, M.C., Saragosti, S., Georges-Courbot, M.C., Barré-Sinoussi, F., Brun-Vézinet, F., 1998. Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. Nat. Med. 4, 1032–1037. https://doi.org/10.1038/2017
Simon-Loriere, E., Martin, D.P., Weeks, K.M., Negroni, M., 2010. RNA Structures Facilitate Recombination-Mediated Gene Swapping in HIV-1. Journal of Virology 84. https://doi.org/10.1128/JVI.01302-10
Simpson, A.G.B., Roger, A.J., 2004. The real 'kingdoms' of eukaryotes. Current Biology 14, R693–R696. https://doi.org/10.1016/j.cub.2004.08.038
Skoglund, P., and Reich, D. (2016). A genomic view of the peopling of the Americas. Curr Opin Genet Dev 41, 27-35.
Slon, V., Mafessoni, F., Vernot, B., de Filippo, C., Grote, S., Viola, B., Hajdinjak, M., Peyregne, S., Nagel, S., Brown, S., et al. (2018). The genome of the offspring of a Neanderthal mother and a Denisovan father. Nature 561, 113-116.
Smithson, M., Verkuilen, J., 2006. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychological Methods 11, 54–71. https://doi.org/10.1037/1082-989X.11.1.54
Smyth, R.P., Schlub, T.E., Grimm, A.J., Waugh, C., Ellenberg, P., Chopra, A., Mallal, S., Cromer, D., Mak, J., Davenport, M.P., Hahn, B.H., 2014. Identifying Recombination Hot Spots in the HIV-1 Genome. Journal of Virology 88, 2891–2902. https://doi.org/10.1128/JVI.03014-13
Sokal, R.R., Jacquez, G.M., 1991. Testing Inferences about Microevolutionary Processes by Means of Spatial Autocorrelation Analysis. Evolution 45, 152. https://doi.org/10.2307/2409490
Sokal, R.R., Michener, C.D., 1958. A statistical method for evaluating systematic relationships. Univ. Kansas, Sci. Bull. 38, 1409–1438.
Sokal, R.R., Wartenberg, D.E., 1983. A Test of Spatial Autocorrelation Analysis Using an Isolation-by-Distance Model. Genetics 105, 219–237.
Sollerkvist, L.P., Gaseitsiwe, S., Mine, M., Sebetso, G., Mphoyakgosi, T., Diphoko, T., Essex, M., Ehrnst, A., 2014. Increased CXCR4 use of HIV-1 subtype C identified by population sequencing in patients failing antiretroviral treatment compared with treatment-naive patients in Botswana. AIDS Res. Hum.

Retroviruses 30, 436–445. https://doi.org/10.1089/AID.2013.0203
Sommer, S., 2005. The importance of immune gene variability (MHC) in evolutionary ecology and conservation. Front. Zool. 2, 16. https://doi.org/10.1186/1742-9994-2-16
Song, W., He, D., Brill, I., Malhotra, R., Mulenga, J., Allen, S., Hunter, E., Tang, J., Kaslow, R.A., 2011. Disparate associations of HLA class I markers with HIV-1 acquisition and control of viremia in an African population. PloS One 6, e23469. https://doi.org/10.1371/journal.pone.0023469
Song,H., Cai,F., Chen,S., Hoper,J., Jiang,C., Gao,F., 2012. Analysis of recombinant viral genomes in individuals infected with multiple transmitted/founder HIV-1 strains. Unpublished direct submission to Genbank.
Specthrie, L., Bullitt, E., Horiuchi, K., Model, P., Russel, M., Makowski, L., 1992. Construction of a microphage variant of filamentous bacteriophage. Journal of Molecular Biology 228, 720–724. https://doi.org/10.1016/0022-2836(92)90858-h
Spire, B., Sire, J., Zachar, V., Rey, F., Barré-Sinoussi, F., Galibert, F., Hampe, A., Chermann, J.C., 1989. Nucleotide sequence of HIV1-NDK: a highly cytopathic strain of the human immunodeficiency virus. Gene 81, 275–284. https://doi.org/10.1016/0378-1119(89)90188-1
Srinivasan, A., Anand, R., York, D., Ranganathan, P., Feorino, P., Schochetman, G., Curran, J., Kalyanaraman, V.S., Luciw, P.A., Sanchez-Pescador, R., 1987. Molecular characterization of human immunodeficiency virus from Zaire: nucleotide sequence analysis identifies conserved and variable domains in the envelope gene. Gene 52, 71–82. https://doi.org/10.1016/0378-1119(87)90396-9
Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033
Stamatakis, A., Ludwig, T., Meier, H., 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21, 456–463. https://doi.org/10.1093/bioinformatics/bti191
Stanier, R.Y., Niel, C.B., 1962. The concept of a bacterium. Archiv für Mikrobiologie 42, 17–35. https://doi.org/10.1007/BF00425185
Steain, M.C., Wang, B., Yang, C., Shi, Y.-P., Nahlen, B., Lal, R.B., Saksena, N.K., 2005. HIV type 1 sequence diversity and dual infections in Kenya. AIDS Res. Hum. Retroviruses 21, 882–885. https://doi.org/10.1089/aid.2005.21.882
Steel, M., Huson, D., Lockhart, P.J., 2000. Invariable sites models and their use in phylogeny reconstruction. Syst Biol 49, 225–232.
Stephens, P.R., Altizer, S., Smith, K.F., Alonso Aguirre, A., Brown, J.H., Budischak, S.A., Byers, J.E., Dallas, T.A., Jonathan Davies, T., Drake, J.M., et al. (2016). The macroecology of infectious diseases: a new perspective on global-scale drivers of pathogen distributions and impacts. Ecol Lett 19, 1159-1171.
Stevens, P.F., 1991. Character States, Morphological Variation, and Phylogenetic Analysis: A Review. Systematic Botany 16, 553.

https://doi.org/10.2307/2419343
Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J., Rambaut, A., 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evolution 4. https://doi.org/10.1093/ve/vey016
Summers, J., Mason, W.S., 1982. Replication of the genome of a hepatitis B-like virus by reverse transcription of an RNA intermediate. Cell 29, 403–415. https://doi.org/10.1016/0092-8674(82)90157-x
Sun, C., Feschotte, C., Wu, Z., Mueller, R.L., 2015. DNA transposons have colonized the genome of the giant virus Pandoravirus salinus. BMC Biology 13. https://doi.org/10.1186/s12915-015-0145-1
Susko, E., 2010. First-Order Correct Bootstrap Support Adjustments for Splits that Allow Hypothesis Testing When Using Maximum Likelihood Estimation. Molecular Biology and Evolution 27, 1621–1629. https://doi.org/10.1093/molbev/msq048
Sutherland, K.A., Goodall, R.L., McCormick, A., Kapaata, A., Lyagoba, F., Kaleebu, P., Thiltgen, G., Gilks, C.F., Spyer, M., Kityo, C., Pillay, D., Dunn, D., Gupta, R.K., DART Virology Group, DART Trial Team, 2015. Gag-Protease Sequence Evolution Following Protease Inhibitor Monotherapy Treatment Failure in HIV-1 Viruses Circulating in East Africa. AIDS Res. Hum. Retroviruses 31, 1032–1037. https://doi.org/10.1089/aid.2015.0138
Suttle, C.A., 2007. Marine viruses — major players in the global ecosystem. Nature Reviews Microbiology 5, 801–812. https://doi.org/10.1038/nrmicro1750
Suzuki, Y., Glazko, G.V., Nei, M., 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc Natl Acad Sci U S A 99, 16138–16143. https://doi.org/10.1073/pnas.212646199
Swofford, D.L., Olsen, G.J., Waddell, P.J., Hillis, D.M., 1996. Phylogenetic Inference, in: Hillis, David M., Moritz, C., Mable, B.K. (Eds.), Molecular Systematics. Sinauer Associates, Sunderland, Mass, pp. 407–514.
Sydow, J.F., Cramer, P., 2009. RNA polymerase fidelity and transcriptional proofreading. Current Opinion in Structural Biology 19, 732–739. https://doi.org/10.1016/j.sbi.2009.10.009
Szolek, A., Schubert, B., Mohr, C., Sturm, M., Feldhahn, M., and Kohlbacher, O. (2014). OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics 30, 3310-3316.
Szpakowski, S., Sun, X., Lage, J.M., Dyer, A., Rubinstein, J., Kowalski, D., Sasaki, C., Costa, J., and Lizardi, P.M. (2009). Loss of epigenetic silencing in tumors preferentially affects primate-specific retroelements. Gene 448, 151-167.
Takahashi, Y., Harashima, N., Kajigaya, S., Yokoyama, H., Cherkasova, E., McCoy, J.P., Hanada, K., Mena, O., Kurlander, R., Tawab, A., et al. (2008). Regression of human kidney cancer following allogeneic stem cell transplantation is associated with recognition of an HERV-E antigen by T cells. The Journal of clinical investigation 118, 1099-1109.
Takahoko, M., Tobiume, M., Ishikawa, K., Ampofo, W., Yamamoto, N., Matsuda, M.,

Tatsumi, M., 2001. Infectious DNA clone of HIV type 1 A/G recombinant (CRF02_AG) replicable in peripheral blood mononuclear cells. AIDS Res. Hum. Retroviruses 17, 1083–1087. https://doi.org/10.1089/088922201300343771

Takekawa, N., Sakamoto, Y., Kinomoto, M., Nii-Trebi, N., Tokunaga, K., Ampofo, W., Ofori-Adjei, D., Kurata, T., Sata, T., Tatsumi, M., 2009. A new generation of recombinants between CRF02_AG and subtype A detected in Ghana : analysis by infectious molecular clones. Unpublished direct submission to Genbank.

Takekawa, N., Tatsumi, M., 2009. A comprehensive panel of infectious molecular clones derived from HIV isolates of BBI subtype infectivity panel. Unpublished direct submission to Genbank.

Takezaki, N., Nei, M., 1994. Inconsistency of the maximum parsimony method when the rate of nucleotide substitution is constant. J Mol Evol 39, 210–218. https://doi.org/10.1007/BF00163810

Tandon, R., SenGupta, D., Ndhlovu, L.C., Vieira, R.G., Jones, R.B., York, V.A., Vieira, V.A., Sharp, E.R., Wiznia, A.A., Ostrowski, M.A., et al. (2011). Identification of human endogenous retrovirus-specific T cell responses in vertically HIV-1-infected subjects. Journal of virology 85, 11526-11531.

Tarosso, L.F., Vieira, V.A., Sauer, M.M., Tomiyama, H.I., Kalil, J., Kallas, E.G., 2017. Conserved HIV-1 Gag p24 Epitopes Elicit Cellular Immune Responses That Impact Disease Outcome. AIDS Res. Hum. Retroviruses 33, 832–842. https://doi.org/10.1089/AID.2016.0168

Tatsumi, M., Hara, T., Sakamoto, Y., Terunuma, H., Ichiyama, H., Kasolo, M., Honda, M., Yamamoto, N., 2006. Construction and characterization of HIV-1 subtype C infectious molecular clones derived from Zambian isolates. Unpublished direct submission to Genbank.

Tatsumi, M., Kinomoto, M., Nii-Trebi, N., Sakamoto, Y., Tokunaga, K., Ampofo, W., Ofori-Adjei, D., Kurata, T., Sata, T., 2005. Infectious DNA clones of subtype G and CRF02_AG HIV-1 derived from Ghananian isolates. Unpublished direct submission to Genbank.

Tatsumi, M.,, Matsuda, M., 2002. Infectious DNA clones of HIV-1 subtype A. Unpublished direct submission to Genbank.

Tatsumi,M., Sakamoto,Y., Takekawa,N., 2006. Construction and characterization of HIV-1 subtype A infectious molecular clones. Unpublished direct submission to Genbank.

Taylor, B.S., Sobieszczyk, M.E., McCutchan, F.E., Hammer, S.M., 2008. The challenge of HIV-1 subtype diversity. N Engl J Med 358, 1590–1602. https://doi.org/10.1056/NEJMra0706737

Tebit, D.M., Arts, E.J., 2011. Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. The Lancet Infectious Diseases 11, 45–56. https://doi.org/10.1016/S1473-3099(10)70186-9

Tebit, D.M., Zekeng, L., Kaptué, L., Gürtler, L., Fackler, O.T., Keppler, O.T.,

Herchenröder, O., Kräusslich, H.-G., 2004. Construction and characterization of an HIV-1 group O infectious molecular clone and analysis of vpr- and nef-negative derivatives. Virology 326, 329–339. https://doi.org/10.1016/j.virol.2004.05.027

Tebit, D.M., Zekeng, L., Kaptué, L., Kräusslich, H.-G., Herchenröder, O., 2003. Construction and characterisation of a full-length infectious molecular clone from a fast replicating, X4-tropic HIV-1 CRF02.AG primary isolate. Virology 313, 645–652. https://doi.org/10.1016/s0042-6822(03)00381-7

Tebit, D.M., Zekeng, L., Kaptué, L., Salminen, M., Kräusslich, H.-G., Herchenröder, O., 2002. Genotypic and phenotypic analysis of HIV type 1 primary isolates from western Cameroon. AIDS Res. Hum. Retroviruses 18, 39–48. https://doi.org/10.1089/088922202753394709

Temin, H.M., 1985. Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. Mol. Biol. Evol. 2, 455–468. https://doi.org/10.1093/oxfordjournals.molbev.a040365

Tenzer, S., Crawford, H., Pymm, P., Gifford, R., Sreenu, V.B., Weimershaus, M., de Oliveira, T., Burgevin, A., Gerstoft, J., Akkad, N., Lunn, D., Fugger, L., Bell, J., Schild, H., van Endert, P., Iversen, A.K.N., 2014. HIV-1 adaptation to antigen processing results in population-level immune evasion and affects subtype diversification. Cell Rep. 7, 448–463. https://doi.org/10.1016/j.celrep.2014.03.031

The UniProt Consortium, 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research 47, D506–D515. https://doi.org/10.1093/nar/gky1049

Theodore, T., Buckler-White, A.J., 1988. Human immunodeficiency virus type 1 (HIV-1), isolate Z2, unintegrated circular proviral DNA. Unpublished direct submission to Genbank.

Thomson, M.M., Casado, G., Posada, D., Sierra, M., Nájera, R., 2005. Identification of a novel HIV-1 complex circulating recombinant form (CRF18_cpx) of Central African origin in Cuba. AIDS 19, 1155–1163. https://doi.org/10.1097/01.aids.0000176215.95119.1d

Thorsby, E., 2009. A short history of HLA. Tissue Antigens 74, 101–116. https://doi.org/10.1111/j.1399-0039.2009.01291.x

Tidona, C.A., Darai, G., 1997. The Complete DNA Sequence of Lymphocystis Disease Virus. Virology 230, 207–216. https://doi.org/10.1006/viro.1997.8456

Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.-M., Doumbo, O., Ibrahim, M., Juma, A.T., Kotze, M.J., Lema, G., Moore, J.H., Mortensen, H., Nyambo, T.B., Omar, S.A., Powell, K., Pretorius, G.S., Smith, M.W., Thera, M.A., Wambebe, C., Weber, J.L., Williams, S.M., 2009. The genetic structure and history of Africans and African Americans. Science 324, 1035–1044. https://doi.org/10.1126/science.1172257

Tobler, W.R., 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. Economic Geography 46, 234. https://doi.org/10.2307/143141

Tongo, M., Dorfman, J.R., Abrahams, M.-R., Mpoudi-Ngole, E., Burgers, W.A., Martin, D.P., 2015. Near full-length HIV type 1M genomic sequences from Cameroon : Evidence of early diverging under-sampled lineages in the country. Evol Med Public Health 2015, 254–265. https://doi.org/10.1093/emph/eov022

Tongo, M., Martin, D.P., Zembe, L., Mpoudi-Ngole, E., Williamson, C., Burgers, W.A., 2013. Characterization of HIV-1 gag and nef in Cameroon: further evidence of extreme diversity at the origin of the HIV-1 group M epidemic. Virol. J. 10, 29. https://doi.org/10.1186/1743-422X-10-29

Toni, T., Adjé-Touré, C., Vidal, N., Minga, A., Huet, C., Borger, M.-Y., Recordon-Pinson, P., Masquelier, B., Nolan, M., Nkengasong, J., Fleury, H.J., Delaporte, E., Peeters, M., 2005. Presence of CRF09_cpx and complex CRF02_AG/CRF09_cpx recombinant HIV type 1 strains in Côte d'Ivoire, West Africa. AIDS Res. Hum. Retroviruses 21, 667–672. https://doi.org/10.1089/aid.2005.21.667

Toure-Kane, C., Montavon, C., Faye, M.A., Gueye, P.M., Sow, P.S., Ndoye, I., Gaye-Diallo, A., Delaporte, E., Peeters, M., Mboup, S., 2000. Identification of all HIV type 1 group M subtypes in Senegal, a country with low and stable seroprevalence. AIDS Res. Hum. Retroviruses 16, 603–609. https://doi.org/10.1089/088922200309025

Tovanabutra, S., Sanders-Buell, E., Arroyo, M.A., Maboko, L., Hoffman, O., Bose, M., Crossler, J., Bradfield, A., Geis, S., Birx, D.L., Robb, M.L., Michael, N.L., Hoelscher, M., McCutchan, F.E., 2009. HIV-1 subtype C gag gene from an incident infection identified in Tanzania. Unpublished direct submission to Genbank.

Treurnicht, F.K., Seoighe, C., Martin, D.P., Wood, N., Abrahams, M.-R., Rosa, D. de A., Bredell, H., Woodman, Z., Hide, W., Mlisana, K., Karim, S.A., Gray, C.M., Williamson, C., 2010. Adaptive changes in HIV-1 subtype C proteins during early infection are driven by changes in HLA-associated immune pressure. Virology 396, 213–225. https://doi.org/10.1016/j.virol.2009.10.002

Triques, K., Bourgeois, A., Vidal, N., Mpoudi-Ngole, E., Mulanga-Kabeya, C., Nzilambi, N., Torimiro, N., Saman, E., Delaporte, E., Peeters, M., 2000. Near-full-length genome sequencing of divergent African HIV type 1 subtype F viruses leads to the identification of a new HIV type 1 subtype designated K. AIDS Res. Hum. Retroviruses 16, 139–151. https://doi.org/10.1089/088922200309485

Trkola, A., Paxton, W.A., Monard, S.P., Hoxie, J.A., Siani, M.A., Thompson, D.A., Wu, L., Mackay, C.R., Horuk, R., Moore, J.P., 1998. Genetic subtype-independent inhibition of human immunodeficiency virus type 1 replication by CC and CXC chemokines. J. Virol. 72, 396–404.

Tu, X., Li, S., Zhao, L., Xiao, R., Wang, X., and Zhu, F. (2017). Human leukemia antigen-A*0201-restricted epitopes of human endogenous retrovirus W family envelope (HERV-W env) induce strong cytotoxic T lymphocyte responses.

Virol Sin 32, 280-289.
Tzitzivacos, D.B., Tiemessen, C.T., Stevens, W.S., Papathanasopoulos, M.A., 2009. Viral genetic determinants of nonprogressive HIV type 1 subtype C infection in antiretroviral drug-naive children. AIDS Res. Hum. Retroviruses 25, 1141–1148. https://doi.org/10.1089/aid.2009.0080
UNAIDS, 2019. Global HIV & AIDS statistics — 2019 fact sheet.
Urban, M.C., 2011. The evolution of species interactions across natural landscapes: Urban - evolution of species interactions. Ecology Letters 14, 723–732. https://doi.org/10.1111/j.1461-0248.2011.01632.x
Vaidyanathan, R., Xie, Y., Allaire, J.J., Cheng, J., Sievert, C., Russell, K., 2020. htmlwidgets: HTML Widgets for R.
Vallari, A., Bodelle, P., Ngansop, C., Makamche, F., Ndembi, N., Mbanya, D., Kaptué, L., Gürtler, L.G., McArthur, C.P., Devare, S.G., Brennan, C.A., 2010. Four new HIV-1 group N isolates from Cameroon: Prevalence continues to be low. AIDS Res. Hum. Retroviruses 26, 109–115. https://doi.org/10.1089/aid.2009.0178
Vallari, A., Holzmayer, V., Harris, B., Yamaguchi, J., Ngansop, C., Makamche, F., Mbanya, D., Kaptué, L., Ndembi, N., Gürtler, L., Devare, S., Brennan, C.A., 2011. Confirmation of putative HIV-1 group P in Cameroon. J. Virol. 85, 1403–1407. https://doi.org/10.1128/JVI.02005-10
Vallejo, A., Heredia, A., Mas, A., Lee, S.F., Epstein, J.S., Soriano, V., Hewlett, I.K., 1998. Tropism, coreceptor use, and phylogenetic analysis of both the V3 loop and the protease gene of three novel HIV-1 group O isolates. J. Acquir. Immune Defic. Syndr. Hum. Retrovirol. 18, 417–425.
Van Etten, J.L., Lane, L.C., Dunigan, D.D., 2010. DNA Viruses: The Really Big Ones (Giruses). Annual Review of Microbiology 64, 83–99. https://doi.org/10.1146/annurev.micro.112408.134338
van Harmelen, J., Williamson, C., Kim, B., Morris, L., Carr, J., Karim, S.S., McCutchan, F., 2001. Characterization of full-length HIV type 1 subtype C sequences from South Africa. AIDS Res. Hum. Retroviruses 17, 1527–1531. https://doi.org/10.1089/08892220152644232
Van Heuverswyn, F., Li, Y., Bailes, E., Neel, C., Lafay, B., Keele, B.F., Shaw, K.S., Takehisa, J., Kraus, M.H., Loul, S., Butel, C., Liegeois, F., Yangda, B., Sharp, P.M., Mpoudi-Ngole, E., Delaporte, E., Hahn, B.H., Peeters, M., 2007. Genetic diversity and phylogeographic clustering of SIVcpzPtt in wild chimpanzees in Cameroon. Virology 368, 155–171. https://doi.org/10.1016/j.virol.2007.06.018
Van Valen, L., 1973. A new evolutionary law. Evolutionary theory 1, 1–30.
Vanspauwen, M.J., Schnabel, R.M., Bruggeman, C.A., Drent, M., van Mook, W.N.K.A., Bergmans, D.C.J.J., Linssen, C.F.M., 2013. Mimivirus is not a frequent cause of ventilator-associated pneumonia in critically ill patients. Journal of Medical Virology 85, 1836–1841. https://doi.org/10.1002/jmv.23655
Veillette, M., Coutu, M., Richard, J., Batraville, L.-A., Dagher, O., Bernard, N., Tremblay, C., Kaufmann, D.E., Roger, M., Finzi, A., 2015. The HIV-1 gp120

CD4-bound conformation is preferentially targeted by antibody-dependent cellular cytotoxicity-mediating antibodies in sera from HIV-1-infected individuals. J. Virol. 89, 545–551. https://doi.org/10.1128/JVI.02868-14

Vergne, L., Bourgeois, A., Mpoudi-Ngole, E., Mougnutou, R., Mbuagbaw, J., Liegeois, F., Laurent, C., Butel, C., Zekeng, L., Delaporte, E., Peeters, M., 2003. Biological and genetic characteristics of HIV infections in Cameroon reveals dual group M and O infections and a correlation between SI-inducing phenotype of the predominant CRF02_AG variant and disease stage. Virology 310, 254–266. https://doi.org/10.1016/s0042-6822(03)00167-3

Vergne, L., Peeters, M., Mpoudi-Ngole, E., Bourgeois, A., Liegeois, F., Toure-Kane, C., Mboup, S., Mulanga-Kabeya, C., Saman, E., Jourdan, J., Reynes, J., Delaporte, E., 2000. Genetic diversity of protease and reverse transcriptase sequences in non-subtype-B human immunodeficiency virus type 1 strains: evidence of many minor drug resistance mutations in treatment-naive patients. J. Clin. Microbiol. 38, 3919–3925.

Vidal, N., Bazepeo, S.E., Mulanga, C., Delaporte, E., Peeters, M., 2009. Genetic characterization of eight full-length HIV type 1 genomes from the Democratic Republic of Congo (DRC) reveal a new subsubtype, A5, in the A radiation that predominates in the recombinant structure of CRF26_A5U. AIDS Res. Hum. Retroviruses 25, 823–832. https://doi.org/10.1089/aid.2008.0283

Vidal, N., Frange, P., Chaix, M.-L., Mulanga, C., Lepira, F., Bazepeo, S.E., Goujard, C., Meyer, L., Rouzioux, C., Delaporte, E., Peeters, M., 2008. Characterization of an old complex circulating recombinant form, CRF27_cpx, originating from the Democratic Republic of Congo (DRC) and circulating in France. AIDS Res. Hum. Retroviruses 24, 315–321. https://doi.org/10.1089/aid.2007.0241

Vidal, N., Koyalta, D., Richard, V., Lechiche, C., Ndinaromtan, T., Djimasngar, A., Delaporte, E., Peeters, M., 2003. High genetic diversity of HIV-1 strains in Chad, West Central Africa. J. Acquir. Immune Defic. Syndr. 33, 239–246. https://doi.org/10.1097/00126334-200306010-00020

Vidal, N., Mulanga, C., Bazepeo, S.E., Lepira, F., Delaporte, E., Peeters, M., 2006. Identification and molecular characterization of subsubtype A4 in central Africa. AIDS Res. Hum. Retroviruses 22, 182–187. https://doi.org/10.1089/aid.2006.22.182

Vidal, N., Mulanga-Kabeya, C., Nzilambi, N., Delaporte, E., Peeters, M., 2000. Identification of a complex env subtype E HIV type 1 virus from the democratic republic of congo, recombinant with A, G, H, J, K, and unknown subtypes. AIDS Res. Hum. Retroviruses 16, 2059–2064. https://doi.org/10.1089/088922200750054800

Vidal, N., Peeters, M., Mulanga-Kabeya, C., Nzilambi, N., Robertson, D., Ilunga, W., Sema, H., Tshimanga, K., Bongo, B., Delaporte, E., 2000. Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1

pandemic originated in Central Africa. J. Virol. 74, 10498–10507. https://doi.org/10.1128/jvi.74.22.10498-10507.2000

Vidal,N., 2002a. Genetic variability of HIV-1 strains in Chad: high prevalence of recombinant strains and identifictauion [sic] of a new D' variant. Unpublished direct submission to Genbank.

Vidal,N., 2002b. Genetic variability of HIV-1 strains in chad: high prevalence of recombinant strains and identification of a new D'' variant. Unpublished direct submission to Genbank.

Vinner, L., Holmgren, B., Jensen, K.J., Esbjornsson, J., Borggren, M., Hentze, J.L., Karlsson, I., Andresen, B.S., Gram, G.J., Kloverpris, H., Aaby, P., Da Silva, Z.J., Fenyö, E.-M., Fomsgaard, A., 2011. Sequence analysis of HIV-1 isolates from Guinea-Bissau: selection of vaccine epitopes relevant in both West African and European countries. APMIS 119, 487–497. https://doi.org/10.1111/j.1600-0463.2011.02763.x

Visawapoka, U., Tovanabutra, S., Currier, J.R., Cox, J.H., Mason, C.J., Wasunna, M., Ponglikitmongkol, M., Dowling, W.E., Robb, M.L., Birx, D.L., McCutchan, F.E., 2006. Circulating and unique recombinant forms of HIV type 1 containing subsubtype A2. AIDS Res. Hum. Retroviruses 22, 695–702. https://doi.org/10.1089/aid.2006.22.695

Voet, D., Voet, J.G., 2011. Biochemistry, 4th ed. ed. John Wiley & Sons, Hoboken, NJ.

Volkman, H.E., and Stetson, D.B. (2014). The enemy within: endogenous retroelements and autoimmune disease. Nature immunology 15, 415-422.

Wang, C., Szpiech, Z.A., Degnan, J.H., Jakobsson, M., Pemberton, T.J., Hardy, J.A., Singleton, A.B., and Rosenberg, N.A. (2010). Comparing spatial maps of human population-genetic variation using Procrustes analysis. Stat Appl Genet Mol Biol 9, Article 13.

Wang, L., Zhang, K., Zhang, L., 2001. Perfect phylogenetic networks with recombination. J Comput Biol 8, 69–78. https://doi.org/10.1089/106652701300099119

Wang, Z., Guo, W., Li, L., Luk'yanchuk, B., Khan, A., Liu, Z., Chen, Z., Hong, M., 2011. Optical virtual imaging at 50 nm lateral resolution with a white-light nanoscope. Nature Communications 2. https://doi.org/10.1038/ncomms1211

Wang-Johanning, F., Radvanyi, L., Rycaj, K., Plummer, J.B., Yan, P., Sastry, K.J., Piyathilake, C.J., Hunt, K.K., and Johanning, G.L. (2008). Human endogenous retrovirus K triggers an antigen-specific immune response in breast cancer patients. Cancer Res 68, 5869-5877.

Weiss, R.A., 1969a. The Host Range of BRYAN Strain Rous Sarcoma Virus Synthesized in the Absence of Helper Virus. Journal of General Virology 5, 511–528. https://doi.org/10.1099/0022-1317-5-4-511

Weiss, R.A., 1969b. Interference and Neutralization Studies with BRYAN Strain Rous Sarcoma Virus Synthesized in the Absence of Helper Virus. Journal of General Virology 5, 529–539. https://doi.org/10.1099/0022-1317-5-4-529

Weiss, R.A., 2006. The discovery of endogenous retroviruses. Retrovirology 3.

https://doi.org/10.1186/1742-4690-3-67
WHO network for HIV isolation and characterization, 1994. HIV type 1 variation in World Health Organization-sponsored vaccine evaluation sites: genetic screening, sequence analysis, and preliminary biological characterization of selected viral strains. WHO Network for HIV Isolation and Characterization. AIDS Res Hum Retroviruses 10, 1327–1343. https://doi.org/10.1089/aid.1994.10.1327
Wickham, H., 2009. Ggplot2: elegant graphics for data analysis, Use R! Springer, New York.
Wieczorek, M., Abualrous, E.T., Sticht, J., Álvaro-Benito, M., Stolzenberg, S., Noé, F., Freund, C., 2017. Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. Frontiers in Immunology 8. https://doi.org/10.3389/fimmu.2017.00292
Wilbe, K., Casper, C., Albert, J., Leitner, T., 2002. Identification of two CRF11-cpx genomes and two preliminary representatives of a new circulating recombinant form (CRF13-cpx) of HIV type 1 in Cameroon. AIDS Res. Hum. Retroviruses 18, 849–856. https://doi.org/10.1089/08892220260190326
Wilbe, K., Salminen, M., Laukkanen, T., McCutchan, F., Ray, S.C., Albert, J., Leitner, T., 2003. Characterization of novel recombinant HIV-1 genomes using the branching index. Virology 316, 116–125. https://doi.org/10.1016/j.virol.2003.08.004
Wildschutte, J.H., Williams, Z.H., Montesion, M., Subramanian, R.P., Kidd, J.M., and Coffin, J.M. (2016). Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proceedings of the National Academy of Sciences of the United States of America 113, E2326-2334.
Wilkinson, E., Engelbrecht, S., 2009. Molecular characterization of non-subtype C and recombinant HIV-1 viruses from Cape Town, South Africa. Infect. Genet. Evol. 9, 840–846. https://doi.org/10.1016/j.meegid.2009.05.001
Wilkinson, E., Holzmayer, V., Jacobs, G.B., de Oliveira, T., Brennan, C.A., Hackett, J., van Rensburg, E.J., Engelbrecht, S., 2015. Sequencing and phylogenetic analysis of near full-length HIV-1 subtypes A, B, G and unique recombinant AC and AD viral strains identified in South Africa. AIDS Res. Hum. Retroviruses 31, 412–420. https://doi.org/10.1089/AID.2014.0230
Wilks, S.S., 1938. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. The Annals of Mathematical Statistics 9, 60–62. https://doi.org/10.1214/aoms/1177732360
Williams, T.A., Embley, T.M., Heinz, E., 2011. Informational Gene Phylogenies Do Not Support a Fourth Domain of Life for Nucleocytoplasmic Large DNA Viruses. PLoS ONE 6, e21080. https://doi.org/10.1371/journal.pone.0021080
Williams, T.A., Heaps, S.E., 2014. An Introduction to Phylogenetics and the Tree of Life, in: Methods in Microbiology. Elsevier, pp. 13–44. https://doi.org/10.1016/bs.mim.2014.05.001
Williamson, C., Morris, L., Maughan, M.F., Ping, L.-H., Dryga, S.A., Thomas, R.,

Reap, E.A., Cilliers, T., van Harmelen, J., Pascual, A., Ramjee, G., Gray, G., Johnston, R., Karim, S.A., Swanstrom, R., 2003. Characterization and selection of HIV-1 subtype C isolates for use in vaccine development. AIDS Res. Hum. Retroviruses 19, 133–144. https://doi.org/10.1089/088922203762688649

Williamson, K.E., Corzo, K.A., Drissi, C.L., Buckingham, J.M., Thompson, C.P., Helton, R.R., 2013. Estimates of viral abundance in soils are strongly influenced by extraction and enumeration methods. Biology and Fertility of Soils 49, 857–869. https://doi.org/10.1007/s00374-013-0780-z

Woese, C.R., Fox, G.E., 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74, 5088–5090. https://doi.org/10.1073/pnas.74.11.5088

Wolf, Y.I., Rogozin, I.B., Grishin, N.V., Koonin, E.V., 2002. Genome trees and the tree of life. Trends in Genetics 18, 472–479. https://doi.org/10.1016/S0168-9525(02)02744-0

Wolfe, N.D., Switzer, W.M., Carr, J.K., Bhullar, V.B., Shanmugam, V., Tamoufe, U., Prosser, A.T., Torimiro, J.N., Wright, A., Mpoudi-Ngole, E., McCutchan, F.E., Birx, D.L., Folks, T.M., Burke, D.S., Heneine, W., 2004. Naturally acquired simian retrovirus infections in central African hunters. Lancet 363, 932–937. https://doi.org/10.1016/S0140-6736(04)15787-5

Wommack, K.E., Williamson, S.J., Sundbergh, A., Helton, R.R., Glazer, B.T., Portune, K., Craig Cary, S., 2004. An instrument for collecting discrete large-volume water samples suitable for ecological studies of microorganisms. Deep Sea Research Part I: Oceanographic Research Papers 51, 1781–1792. https://doi.org/10.1016/j.dsr.2004.05.011

Woo, J., Robertson, D.L., Lovell, S.C., 2010. Constraints on HIV-1 diversity from protein structure. J. Virol. 84, 12995–13003. https://doi.org/10.1128/JVI.00702-10

Worboys, M., 1994. The comparative history of sleeping sickness in east and central Africa. Hist Sci 32, 89–102. https://doi.org/10.1177/007327539403200103

World Bank, 2019. World Development Indicators [WWW Document]. URL https://databank.worldbank.org

Worobey, M., Gemmel, M., Teuwen, D.E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J.-J., Kabongo, J.-M.M., Kalengayi, R.M., Van Marck, E., Gilbert, M.T.P., Wolinsky, S.M., 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature 455, 661–664. https://doi.org/10.1038/nature07390

Wright, E.A., Heckel, T., Groenendijk, J., Davies, J.W., Boulton, M.I., 1997. Splicing features in maize streak virus virion- and complementary-sense gene expression. The Plant Journal 12, 1285–1297. https://doi.org/10.1046/j.1365-313x.1997.12061285.x

Wright, J.K., Brumme, Z.L., Carlson, J.M., Heckerman, D., Kadie, C.M., Brumme, C.J., Wang, B., Losina, E., Miura, T., Chonco, F., van der Stok, M., Mncube,

Z., Bishop, K., Goulder, P.J.R., Walker, B.D., Brockman, M.A., Ndung'u, T., 2010. Gag-protease-mediated replication capacity in HIV-1 subtype C chronic infection: associations with HLA type and clinical parameters. J. Virol. 84, 10820–10831. https://doi.org/10.1128/JVI.01084-10

Wright, J.K., Naidoo, V.L., Brumme, Z.L., Prince, J.L., Claiborne, D.T., Goulder, P.J.R., Brockman, M.A., Hunter, E., Ndung'u, T., 2012. Impact of HLA-B*81-associated mutations in HIV-1 Gag on viral replication capacity. J. Virol. 86, 3193–3199. https://doi.org/10.1128/JVI.06682-11

Wright, J.K., Novitsky, V., Brockman, M.A., Brumme, Z.L., Brumme, C.J., Carlson, J.M., Heckerman, D., Wang, B., Losina, E., Leshwedi, M., van der Stok, M., Maphumulo, L., Mkhwanazi, N., Chonco, F., Goulder, P.J.R., Essex, M., Walker, B.D., Ndung'u, T., 2011. Influence of Gag-protease-mediated replication capacity on disease progression in individuals recently infected with HIV-1 subtype C. J. Virol. 85, 3996–4006. https://doi.org/10.1128/JVI.02520-10

Wright, S., 1932. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. Proceedings of the Sixth International Congress of Genetics 356–366.

Wright, W.J., Irvine, K.M., Warren, J.M., Barnett, J.K., 2017. Statistical design and analysis for plant cover studies with multiple sources of observation errors. Methods Ecol Evol 8, 1832–1841. https://doi.org/10.1111/2041-210X.12825

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E.C., Zhang, Y.-Z., 2020. A new coronavirus associated with human respiratory disease in China. Nature 579, 265–269. https://doi.org/10.1038/s41586-020-2008-3

Yamaguchi, J., Bodelle, P., Kaptué, L., Zekeng, L., Gürtler, L.G., Devare, S.G., Brennan, C.A., 2003. Near full-length genomes of 15 HIV type 1 group O isolates. AIDS Res. Hum. Retroviruses 19, 979–988. https://doi.org/10.1089/088922203322588332

Yamaguchi, J., Bodelle, P., Vallari, A.S., Coffey, R., McArthur, C.P., Schochetman, G., Devare, S.G., Brennan, C.A., 2004. HIV infections in northwestern Cameroon: identification of HIV type 1 group O and dual HIV type 1 group M and group O infections. AIDS Res. Hum. Retroviruses 20, 944–957. https://doi.org/10.1089/aid.2004.20.944

Yamaguchi, J., Coffey, R., Vallari, A., Ngansop, C., Mbanya, D., Ndembi, N., Kaptué, L., Gürtler, L.G., Bodelle, P., Schochetman, G., Devare, S.G., Brennan, C.A., 2006. Identification of HIV type 1 group N infections in a husband and wife in Cameroon: viral genome sequences provide evidence for horizontal transmission. AIDS Res. Hum. Retroviruses 22, 83–92. https://doi.org/10.1089/aid.2006.22.83

Yamaguchi, J., Ndembi, N., Ngansop, C., Mbanya, D., Kaptué, L., Gürtler, L.G., Devare, S.G., Brennan, C.A., 2009. HIV type 1 group M subtype G in

Cameroon: five genome sequences. AIDS Res. Hum. Retroviruses 25, 469–473. https://doi.org/10.1089/aid.2008.0296

Yamaguchi, J., Vallari, A., Ngansop, C., Makamche, F., Ndembi, N., Mbanya, D., Kaptué, L., Gürtler, L.G., Devare, S.G., Brennan, C.A., 2010. Near full-length sequence of HIV type 1 subtype J strain 04CMU11421 from Cameroon. AIDS Res. Hum. Retroviruses 26, 693–697. https://doi.org/10.1089/aid.2009.0305

Yamaguchi, J., Vallari, A.S., Swanson, P., Bodelle, P., Kaptué, L., Ngansop, C., Zekeng, L., Gürtler, L.G., Devare, S.G., Brennan, C.A., 2002. Evaluation of HIV type 1 group O isolates: identification of five phylogenetic clusters. AIDS Res. Hum. Retroviruses 18, 269–282. https://doi.org/10.1089/088922202753472847

Yang, Z., 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10, 1396–1401. https://doi.org/10.1093/oxfordjournals.molbev.a040082

Yang, Z., 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol 11, 367–372. https://doi.org/10.1016/0169-5347(96)10041-0

Yang, Z., 2006. Computational molecular evolution, Oxford series in ecology and evolution. Oxford University Press, Oxford.

Yang, Z., Rannala, B., 2005. Branch-Length Prior Influences Bayesian Posterior Probability of Phylogeny. Systematic Biology 54, 455–470. https://doi.org/10.1080/10635150590945313

Yang, Z., Rannala, B., 2012. Molecular phylogenetics: principles and practice. Nature Reviews Genetics 13, 303–314. https://doi.org/10.1038/nrg3186

Yau, S., Lauro, F.M., DeMaere, M.Z., Brown, M.V., Thomas, T., Raftery, M.J., Andrews-Pfannkoch, C., Lewis, M., Hoffman, J.M., Gibson, J.A., Cavicchioli, R., 2011. Virophage control of antarctic algal host-virus dynamics. Proceedings of the National Academy of Sciences 108, 6163–6168. https://doi.org/10.1073/pnas.1018221108

Yu, P., Lübben, W., Slomka, H., Gebler, J., Konert, M., Cai, C., Neubrandt, L., Prazeres da Costa, O., Paul, S., Dehnert, S., Döhne, K., Thanisch, M., Storsberg, S., Wiegand, L., Kaufmann, A., Nain, M., Quintanilla-Martinez, L., Bettio, S., Schnierle, B., Kolesnikova, L., Becker, S., Schnare, M., Bauer, S., 2012. Nucleic acid-sensing Toll-like receptors are essential for the control of endogenous retrovirus viremia and ERV-induced tumors. Immunity 37, 867–879. https://doi.org/10.1016/j.immuni.2012.07.018

Yu, V.L., Edberg, S.C., 2005. Global Infectious Diseases and Epidemiology Network (GIDEON): A World Wide Web-Based Program for Diagnosis and Informatics in Infectious Diseases. Clinical Infectious Diseases 40, 123–126. https://doi.org/10.1086/426549

Yuan, T., Li, J., Zhang, M.-Y., 2013. HIV-1 envelope glycoprotein variable loops are indispensable for envelope structural integrity and virus entry. PloS One 8, e69789. https://doi.org/10.1371/journal.pone.0069789

Yue, L., Pfafferott, K.J., Baalwa, J., Conrod, K., Dong, C.C., Chui, C., Rong, R.,

Claiborne, D.T., Prince, J.L., Tang, J., Ribeiro, R.M., Cormier, E., Hahn, B.H., Perelson, A.S., Shaw, G.M., Karita, E., Gilmour, J., Goepfert, P., Derdeyn, C.A., Allen, S.A., Borrow, P., Hunter, E., 2015. Transmitted virus fitness and host T cell responses collectively define divergent infection outcomes in two HIV-1 recipients. PLoS Pathog. 11, e1004565. https://doi.org/10.1371/journal.ppat.1004565

Yutin, N., Kapitonov, V.V., Koonin, E.V., 2015a. A new family of hybrid virophages from an animal gut metagenome. Biology Direct 10. https://doi.org/10.1186/s13062-015-0054-9

Yutin, N., Raoult, D., Koonin, E.V., 2013. Virophages, polintons, and transpovirons: a complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. Virol. J. 10, 158. https://doi.org/10.1186/1743-422X-10-158

Yutin, N., Shevchenko, S., Kapitonov, V., Krupovic, M., Koonin, E.V., 2015b. A novel group of diverse Polinton-like viruses discovered by metagenome analysis. BMC Biology 13. https://doi.org/10.1186/s12915-015-0207-4

Yutin, N., Wolf, Y.I., Koonin, E.V., 2014. Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. Virology 466–467, 38–52. https://doi.org/10.1016/j.virol.2014.06.032

Zade, A., Sengupta, M., Kondabagil, K., 2015. Extensive in silico analysis of Mimivirus coded Rab GTPase homolog suggests a possible role in virion membrane biogenesis. Frontiers in Microbiology 6. https://doi.org/10.3389/fmicb.2015.00929

Zamani, C., Elzey, J.D., Hildreth, J.E., 2013. Greater ethnic diversity correlates with lower HIV prevalence in Africa: justification for an alloimmunity vaccine. HIVAIDS Auckl. NZ 5, 75–80. https://doi.org/10.2147/HIV.S38922

Zeberg, H., Pääbo, S., 2021. A genomic region associated with protection against severe COVID-19 is inherited from Neandertals. Proc Natl Acad Sci USA 118, e2026309118. https://doi.org/10.1073/pnas.2026309118

Zeberg, H., Pääbo, S., 2020. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. Nature 587, 610–612. https://doi.org/10.1038/s41586-020-2818-3

Zeileis, A., 2016. variance vs. standard deviation in beta regression. Cross Validated (accessed 12-06-2020).

Zeng, M., Hu, Z., Shi, X., Li, X., Zhan, X., Li, X.D., Wang, J., Choi, J.H., Wang, K.W., Purrington, T., et al. (2014). MAVS, cGAS, and endogenous retroviruses in T-independent B cell responses. Science 346, 1486-1492.

Zhang, X., Sun, S., Xiang, Y., Wong, J., Klose, T., Raoult, D., Rossmann, M.G., 2012. Structure of Sputnik, a virophage, at 3.5-A resolution. Proceedings of the National Academy of Sciences 109, 18431–18436. https://doi.org/10.1073/pnas.1211702109

Zhao, J., Tang, S., Ragupathy, V., Carr, J.K., Wolfe, N.D., Awazi, B., Hewlett, I., 2010. Identification and genetic characterization of a novel CRF22_01A1

recombinant form of HIV type 1 in Cameroon. AIDS Res. Hum. Retroviruses 26, 1033–1045. https://doi.org/10.1089/aid.2009.0197

Zhao, J., Tang, S., Ragupathy, V., Gaddam, D., Wang, X., Zhang, P., Nyambi, P.N., Hewlett, I., 2012. CRF22_01A1 is involved in the emergence of new HIV-1 recombinants in Cameroon. J. Acquir. Immune Defic. Syndr. 60, 344–350. https://doi.org/10.1097/QAI.0b013e318258c7e3

Zhao, J., Tang, S., Viswanath, R., Wang, X., Awazi, B., Hewlett,I., 2008. Identification of new recombinant of CRF22 with subtype A1 or CRF02_AG of HIV-1 in Cameroon. Unpublished direct submission to Genbank.

Zharkikh, A., Li, W.-H., 1993. Inconsistency of the Maximum-parsimony Method: the Case of Five Taxa With a Molecular Clock. Systematic Biology 42, 113–125. https://doi.org/10.1093/sysbio/42.2.113

Zheng, N.N., Daniels, R.S., 2004. Substitution of hiv type-1 non-B env genes in C2, a subtype B cassette system, results in functional chimeric viruses. AIDS Res. Hum. Retroviruses 20, 964–971. https://doi.org/10.1089/aid.2004.20.964

Zhou, J., Sun, D., Childers, A., McDermott, T.R., Wang, Y., Liles, M.R., 2014. Three Novel Virophage Genomes Discovered from Yellowstone Lake Metagenomes. Journal of Virology 89, 1278–1285. https://doi.org/10.1128/jvi.03039-14

Zhou, J., Zhang, W., Yan, S., Xiao, J., Zhang, Y., Li, B., Pan, Y., Wang, Y., 2013. Diversity of Virophages in Metagenomic Data Sets. Journal of Virology 87, 4225–4236. https://doi.org/10.1128/jvi.03398-12

Zhu, J., Paul, W.E., 2010. Heterogeneity and plasticity of T helper cells. Cell Research 20, 4–12. https://doi.org/10.1038/cr.2009.138

Zimmermann, C., Kowalewski, D., Bauersfeld, L., Hildenbrand, A., Gerke, C., Schwarzmüller, M., Le-Trilling, V.T.K., Stevanovic, S., Hengel, H., Momburg, F., Halenius, A., 2019. HLA-B locus products resist degradation by the human cytomegalovirus immunoevasin US11. PLOS Pathogens 15, e1008040. https://doi.org/10.1371/journal.ppat.1008040

Zuñiga, R., Lucchetti, A., Galvan, P., Sanchez, S., Sanchez, C., Hernandez, A., Sanchez, H., Frahm, N., Linde, C.H., Hewitt, H.S., Hildebrand, W., Altfeld, M., Allen, T.M., Walker, B.D., Korber, B.T., Leitner, T., Sanchez, J., Brander, C., 2006. Relative dominance of Gag p24-specific cytotoxic T lymphocytes is associated with human immunodeficiency virus control. J. Virol. 80, 3122–3125. https://doi.org/10.1128/JVI.80.6.3122-3125.2006