# Viral Genetics of HIV-2 Infection

A thesis submitted for the degree of Doctor of Philosophy


Katherine Louise James
Hertford College

**Supervisors:**

Professor Sarah Rowland-Jones
Nuffield Department of Medicine

Professor Gilean McVean
Wellcome Trust Centre for Human Genetics

Nuffield Department of Medicine
University of Oxford

# Abstract

HIV-2 is a contemporary human retrovirus with the majority of infections localised to West Africa. Both HIV-1 and HIV-2 are able to cause AIDS; however, in contrast to HIV-1 infection, a common outcome following HIV-2 infection (~37% of patients in this study cohort) is long-term non-progression (LTNP), where patients remain aviraemic and asymptomatic in the absence of treatment, often for decades. HIV-1 and HIV-2 both arose following zoonotic transmission of SIVs from non-human primates at around the beginning of the 20$^{th}$ century and when patients develop AIDS caused by HIV-2 infection, it is clinically indistinguishable from AIDS following HIV-1 infection.

Whilst the estimated number of HIV-2 infections remains small in the context of the global HIV pandemic (HIV-2 ~2 million, HIV-1 group M ~75 million), the differences in pathogenicity between these two viruses has been a source of great interest, particularly the features of LTNPs that allow control of viral replication in the absence of anti-retroviral treatment.

The studies described in this thesis were carried out using samples collected from a well-characterised longitudinal community cohort in Caió, Guinea-Bissau. Chapter 3 of this thesis presents an investigation into the variation and evolution present in the HIV-2 specific accessory gene *vpx*. The data showed significantly increased signals of positive selection pressure in *vpx* in viraemic when compared to non-viraemic patients and also allowed the identification of novel variations at high frequencies (up to 22%) in this cohort that were previously un-described. Chapters 4 and 5 present a novel application of shotgun RNA sequencing (RNA-Seq) to HIV *ex vitro* and *ex vivo* samples. Chapter 4 demonstrates the divergence seen in a cultured viral isolate at the level of the whole genome, in the absence of many of the biases typically involved in sequencing of RNA viruses. Chapter 5 further extends this method to show the applicability of using RNA-Seq on primary patient HIV samples for the first time. Analysis of diversity estimates over the whole genome in the context of a low bias sequencing method show a high level of diversity in HIV-2 *pol* and low diversity in *vpx*.

The aim of this work was to combine traditional and novel sequencing methods to facilitate assessment of the variation and evolution acting on *vpx* and to generate an accurate picture of the genetic diversity over the whole genome of HIV-2.

# Table of Contents

## Acknowledgments

All genetic studies are impossible without two things, support and participants. I'd like to acknowledge the financial and pastoral support of the Wellcome Trust over the past 4 years and the participation of the people of Caió, whose involvement in multiple sero-surveys has made so much of our knowledge of HIV-2 possible.

I'd like to acknowledge the input into and support during this project from my supervisors, Sarah Rowland-Jones and Gilean McVean. Sarah, your knowledge of HIV-2 and the impact of the epidemic on the lives of West Africans is second to none. I hope some of your passion is reflected here and I thank you for all your kind words and encouragement over the past three years. Gil, very few people so actively push the limits of human genetics as you do and I'd like to thank you for initially suggesting RNA-Seq as a possibility and for guiding me through many of the analyses presented here. I'd also like to thank Karen Poxon, who has often been the calm in the eye of the storm.

I'd like to thank Jonathan Flint and Richard Mott, whose dedication to the Genomic Medicine and Statistics DTC programme has enabled this project. Special thanks must also go to Lorna Witty at the WTCHG Genomics Core and Stephen Taylor at the CBRG, without whose tireless work much of this project would not have been possible.

Members of the SRJ and McVean groups have provided invaluable input into this project. A big thank you to Sophie Andrews for her expert proofreading abilities and great friendship during the sometimes frantic final stages of this project. I'd also like to thank Louis-Marie Yindom, Miguel Garcia-Knight and Shokouh Makvandi-Nejad for being the older, wiser ears when I needed them. Glenn Wong and Shmona Simpson have provided help and companionship during late nights in the lab. I'd like to give special thanks to Thushan De Silva, whose immaculate sample collection and continued input and support have facilitated this work. I'd like to thank Kiran Garimella for lending me some of his rare talent for combining coding and design to allow the generation of some of the figures in this thesis. I'd also like to thank Luke Jostins, Ian Matheison and Denise Xifara for enlightening discussions and ready support when needed.

I've had the privilege of working with many international collaborators. Special thanks must go to Joakim Esbjörnsson for all his advice, support and barbequing skills that have made the past few years so enjoyable. I'd also like to thank Patrik Medstrand and Angelica Palm for being such gracious hosts during my time in Lund. I had the pleasure of hosting Takayuki Chikata for 3 months in 2013 and I look forward to the return leg soon!

I am lucky to be surrounded by a fantastic group of friends, many of who have been invaluable sources of support and joy during this project. It's impossible to name everyone but I'd like to thank Marki and Laz for more than a decade of japes, Bella, Cath, Kate and Beejal for often making the long trek out to Oxford, sometimes to go hiking through the driving snow! I'd like to thank Michael, Rachel, Niall, Colette and Mickey for making our time in Oxford and Edinburgh so enjoyable and I'd like to thank my travelling pals, Woody, Alex, Lizzie, Ben and

Ian for being such excellent company during my much needed breaks from Oxford. I cannot think of a better analogy for finishing a PhD than the feeling of finally freeing a particular white rental car from it's almost certain final resting place in amongst the cava vines! Special thanks must go to Katy Brown, your soothing words and blossoming Python skills have made the past few months a lot smoother than they could have been!

Ben, you have been my greatest cheerleader during this project and to come home to you at the end of everyday is a joy. You make me laugh so very often and it's at times like this I realise we're growing up together. I can't wait for our next big adventure (although I hope it has fewer words in it!).

Last but not least I'd like to acknowledge the support of my family. Philip and Maz, I'm so glad we got to overlap in Oxford and I thank you for your friendship, advice and almost limitless library of boxsets for loan! Finally, I would like to acknowledge my parents, Barbara Hatton and Roger James. You are my greatest and most loyal supporters and without your constant encouragement and (occasionally financial!) backing none of this would have been achievable. I thank you both (inadequately) here.

# List of Abbreviations

| | |
|---|---|
| +G | Gamma Distributed Substitution Rate Heterogeneity |
| +I | + A proportion of Invariant Sites |
| AGM | African Green Monkey |
| AGS | Aicardi-Goutières Syndrome |
| AIDS | Acquired Immune Deficiency Syndrome |
| AN | Accession Number |
| ART | Antiretroviral Therapy |
| BF | Bayes Factor |
| cDNA | Complementary DNA |
| CMV | Cytomegalovirus |
| CRF | Circulating Recombinant Form |
| CSW | Commercial Sex Worker |
| CTL | Cytotoxic T-Lymphocyte |
| DC | Dendritic Cell |
| df | Degrees of Freedom |
| DNA | Deoxyribonucleic Acid |
| dsDNA | Double Stranded DNA |
| EC | Elite Controller |
| Env | Envelope |
| FSW | Female Sex Worker |
| Gag | Group Specific Antigens |
| GM | The Gambia |
| GTR | General Time Reversible |
| GUM | Genito-Urinary Medicine |
| GW | Guinea-Bissau |
| GWAS | Genome-Wide Association Study |
| HCV | Hepatitis C Virus |
| HIV-1 | Human Immunodeficiency Virus Type 1 |
| HIV-2 | Human Immunodeficiency Virus Type 2 |
| HIV-D | HIV-1/HIV-2 Dual Infection |
| HKY85 | Hasegawa, Kishino and Yano 1985 |
| HLA | Human Leucocyte Antigen |
| IFN | Interferon |
| IGV | Integrative Genomics Viewer |
| IN | Integrase |
| IRF | Interferon Regulatory Factor |
| JC | Jukes Cantor |
| KIR | Killer-Cell Immunoglobulin-Like Receptor |
| LANL | Los Alamos National Laboratory |
| LB | Lysogeny Broth |

| | |
|---|---|
| LRT | Likelihood Ratio Test |
| LTNP | Long Term Non-Progressor |
| LTR | Long Terminal Repeat |
| MCCT | Maximum Clade Credibility Tree |
| MDM | Monocyte-Derived Macrophages |
| MHC | Major Histocompatibility Complex |
| ML | Maximum Likelihood |
| MLV | Murine Leukaemia Virus |
| MRC | Medical Research Council |
| mRNA | Messenger RNA |
| MSA | Multiple Sequence Alignment |
| Nab | Neutralising Antibody |
| NC | Nucleocapsid |
| Nef | Negative Regulatory Factor |
| NGS | Next Generation Sequencing |
| NHP | Non-Human Primate |
| NK | Natural Killer (Cell) |
| NLS | Nuclear Localisation Signal |
| ORF | Open Reading Frame |
| OWM | Old World Monkey |
| PBMC | Peripheral Blood Mononuclear Cell |
| PBS | Phosphate Buffered Saline |
| PCR | Polymerase Chain Reaction |
| PEG | Polyethylene Glycol |
| PIC | Pre-Integration Complex |
| Pol | Polymerase |
| PPM | Poly-Proline Motif |
| PPT | Polypurine Tract |
| PR | Protease |
| Q-score | Quality-Score |
| R | Repeat |
| Rev | Regulator of Expression of Viral Proteins |
| RH | Random Hexamer |
| RNA | Ribonucleic Acid |
| RNA-Seq | Shotgun RNA Sequencing |
| RPMI | Roswell Park Memorial Institute (medium) |
| RRE | Rev Responsive Element |
| RT | Reverse Transcriptase |
| SH | Shimodaira-Hasegawa |
| shRNA | Short-Hairpin RNA |
| SIV | Simian Immunodeficiency Virus |

| | |
|---|---|
| SN | Senegal |
| SNP | Single Nucleotide Polymorphism |
| SOC | Super Optimal Broth with Catabolite Repression |
| ssRNA | Single-Stranded RNA |
| SU | Surface |
| TAE | Tris-Acetate-EDTA |
| TAR | Trans-Activation Response Element |
| Tat | Transactivator of Transcription |
| TCID | Tissue Culture Infective Dose |
| TCR | T-Cell Receptor |
| TE | Tris-EDTA |
| TF | Transcription Factor |
| TM | Trans-Membrane |
| tRNA | Transfer RNA |
| U3 | Untranslated 3' |
| U5 | Untranslated 5' |
| UV | Ultraviolet |
| Vif | Viral Infectivity Factor |
| VL | Viral Load |
| Vpr | Viral Protein R |
| Vpx | Viral Protein X |
| WHO | World Health Organisation |
| WT | Wild Type |

# List of Figures

# List of Tables

# Chapter 1: Introduction

## 1.1 The global HIV/AIDS epidemic

The HIV/AIDS epidemic is one of the defining social histories of the 20th century. Acquired Immunodeficiency Syndrome (AIDS) was initially described in 1981 following an unexpected increase in young homosexual men presenting with opportunistic infections and rare malignancies, conditions indicative of immunodeficiency [1]. A novel human retrovirus, Human Immunodeficiency Virus Type-1 (HIV-1) was identified as the causative agent of AIDS in 1981[2] and a second causative retrovirus, Human Immunodeficiency Virus Type-2 (HIV-2) was subsequently identified in 1986 [3],[4],[5].

In the three decades since the initial description of HIV/AIDS, an estimated 75 million people have been infected and 35.3 million people are currently living with HIV infection [6]. Global HIV prevalence varies across countries and continents but HIV is seen in almost every country (**Figure 1.1**)[7].

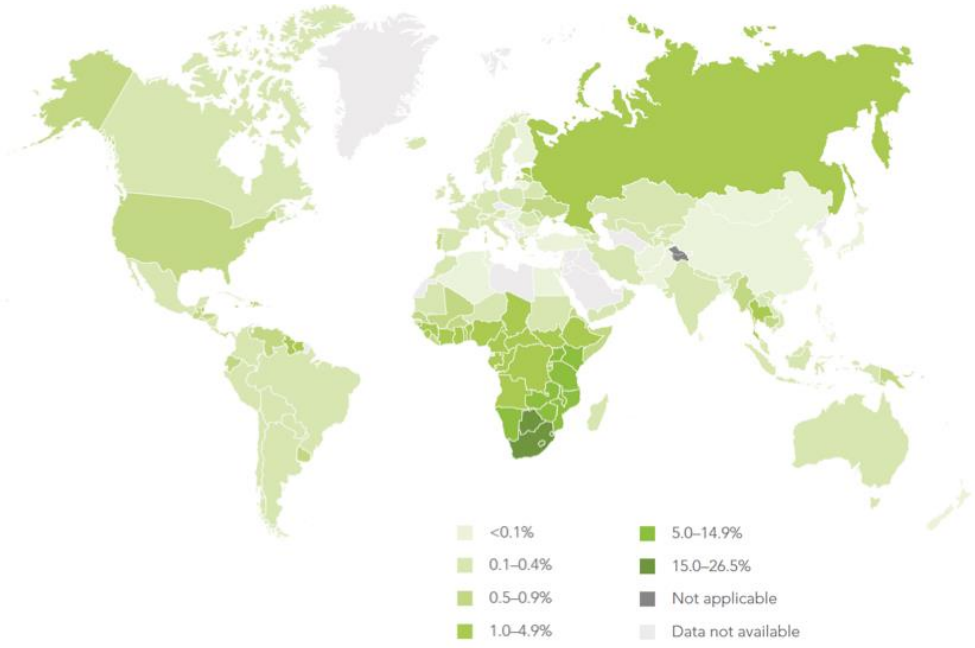



Figure 1.1: Global HIV prevalence.
Countries are coloured according to the estimated prevalence of HIV (adapted from[6])

In spite of advances in treatment and targeted public health campaigns, the HIV/AIDS epidemic remains a leading cause of disease and death worldwide. In

2010 there were 2.7 million new infections, representing a fall of 21% from the peak in 1997[7]. The World Health Organisation (WHO) estimates that less than 10 million of the 28.6 million people eligible for antiretroviral treatment (ART) in low and middle-income countries are currently receiving treatment and in 2012 there were 1.6 million AIDS-related deaths worldwide.

Furthermore, in addition to the gaps in ART coverage, there remains no effective HIV vaccine or cure. A lack of knowledge about the host and viral genetic factors involved in natural control of HIV in the absence of ART in a small proportion of patients known as elite controllers (EC) or long-term non-progressors (LTNP) is surely a major obstacle in the development of an HIV vaccine.

## 1.2 HIV-1 and HIV-2 origins

HIV-1 and HIV-2 both entered human populations following a zoonotic (cross-species) transmission of non-human primate (NHP) lentiviruses (**Figure 1.2**)[8]. Approximately 40 primate species are infected with Simian Immunodeficiency Viruses (SIVs) and in the majority of cases SIV infection is non-pathogenic[9]. Two exceptions are SIVcpz, infecting 2 subspecies of chimpanzees[10], and SIVmac infecting rhesus macaques[11]. Both SIVmac and SIVcpz are able to cause AIDS and SIVcpz infection is associated with a 10 to 16-fold higher age-corrected death hazard in free-ranging chimpanzees[10]. Neither chimpanzees nor rhesus macaques appear to be the natural hosts for SIV but were infected following zoonotic transmission of SIV from another primate. SIVmac can be traced to a recent and inadvertent origin in US primate centres following inoculation of various macaque species with blood from sooty mangabeys infected with SIVsmm[12]. SIVcpz is a complex mosaic virus generated by recombination between SIVrcm (infecting several species of red-capped mangabeys) and a virus closely related to SIVs infecting several *Cercopithecus* monkey species[13]. The mosaic nature of SIVcpz most likely results from the complex predation patterns of chimpanzees, incorporating multiple species dependent on overlapping habitats[14].

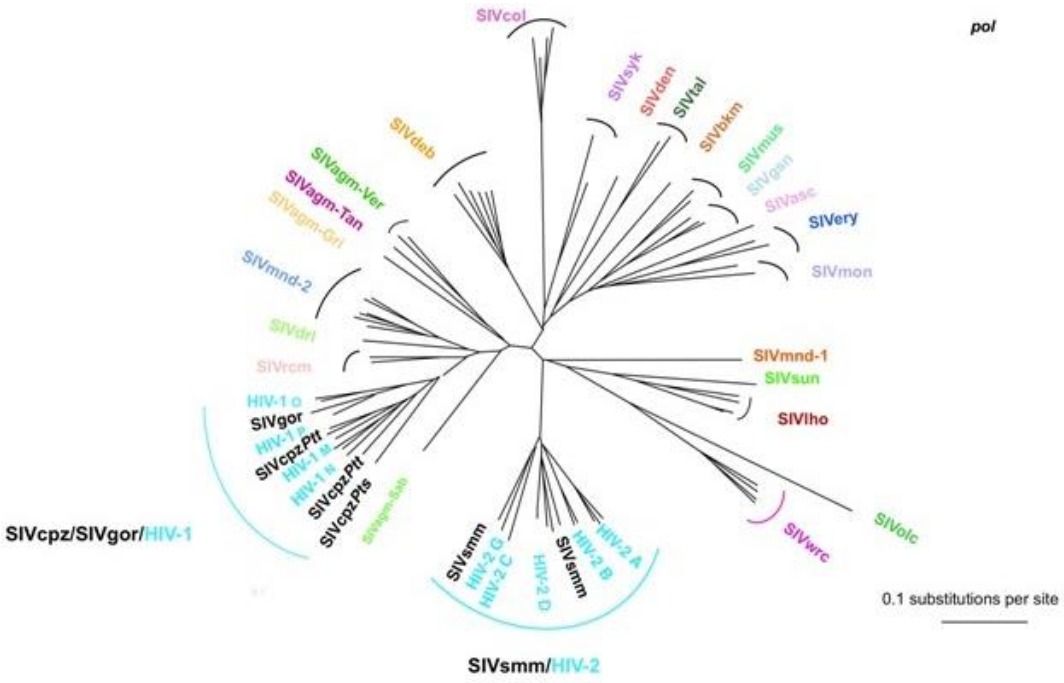



Figure 1.2: Evolutionary relationship between SIV and HIV lineages (based on *pol* sequence data). HIV-1 and HIV-2 sequences are interspersed with SIVcpz/SIVgor and SIVsmm respectively, reflecting their simian origins (adapted from [15])

HIV-1 comprises 4 distinct lineages, known as groups M (Major), N (Non-M, Non-O), O (Outlier) and P (Pending). Each group is the result of an independent transmission event of either SIVcpz or SIVgor (infecting gorillas). HIV-1 group M is the pandemic form of the virus and responsible for the majority of HIV infections worldwide. Analysis of contemporary HIV-1 M sequences, sequences from paraffin-embedded lymph node samples from 1959 and 1960 and SIVcpzPtt sequences from wild *Pan troglodytes troglodytes* chimpanzees indicates that HIV-1 group M entered human populations around 1900 (combined confidence interval 1873-1933) in West Central Africa[8,16]. HIV-1 group O is less prevalent than group M and accounts for less than 1% of infections worldwide, mostly restricted to Cameroon and Gabon[17]. The origin of HIV-1 O remains unclear due to the lack of a particularly closely related SIV sequence but it clusters with SIVcpz and SIVgor, suggesting one may be the source[9]. Along with HIV-1 M, HIV-1 O appears to have entered human populations around the start of the 20$^{th}$ century[17]. HIV-1 group N has only been documented in 13 individuals, all from Cameroon[18], and phylogenetic analysis suggests it arose via zoonotic transmission of SIVcpzPtt[19].

HIV-1 group P has only been reported in two individuals, both from Cameroon. HIV-1 P appears to have a gorilla origin, but there are not enough SIVgor sequences available to estimate a date or location for the zoonosis[20].

Following the initial transmission into humans, probably as the result of exposure to contaminated blood during Bush meat preparation[8], HIV-1 M has rapidly evolved. The heterogeneity of HIV-1 is due to many factors including an error-prone reverse transcriptase lacking proofreading activity, a rapid viral life cycle leading to a high population turnover and template switching by reverse transcriptase which leads to the production of recombinant progeny viruses following infection with one or more strains (superinfection)[21],[22],[23]. HIV-1 M has evolved into 9 distinct subtypes (A-D, F-H and J-K) and 66 circulating recombinant forms (CRFs), seen in 3 or more epidemiologically unrelated cases (**Figure 1.3**)[24]. Global dispersion patterns of the different HIV-1 subtypes are likely to recapitulate human migration routes and historical geographic links. Differing prominent subtypes result from founder effects when a local epidemic forms following a limited number of initial infections[25].

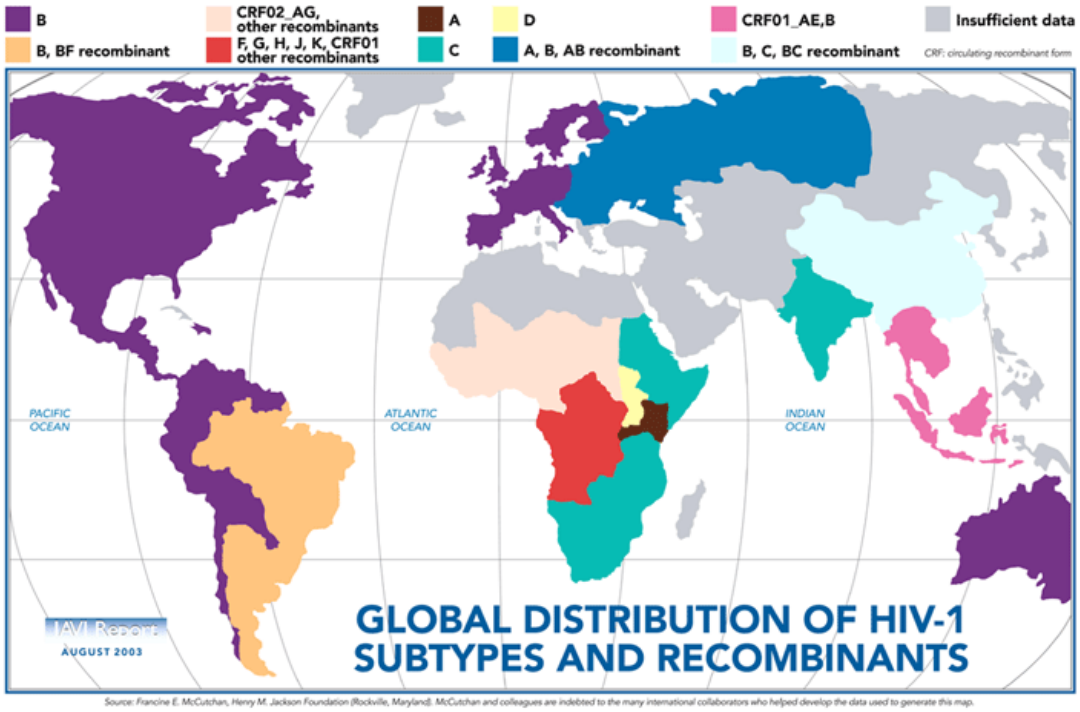



Figure 1.3: Global distribution of HIV-1 M subtypes.
Regions are coloured according to the most prevalent HIV-1 group M subtype or CRFs (adapted from [26])

HIV-2 also has a simian origin and sequence analysis shows that it arose from a zoonotic transmission of SIVsmm from the West African sooty mangabey (*Cercocebus atys atys*)[27,28]. In contrast to chimpanzees infected with SIVcpz, SIVsmm is non-pathogenic in sooty mangabeys, suggesting that they are the natural reservoir for the virus[29] . In a study of free-ranging sooty mangabeys in the Taï Forest, Côte d'Ivoire, SIVsmm prevalence was found to be around 59% and phylogenetic analysis showed evidence of both vertical and horizontal transmission of SIVsmm[30]. SIVsmm has entered human populations on at least 8 different occasions, giving rise to the HIV-2 lineages A to H[31]. Only two of these lineages, A and B, are endemic. HIV-2 group A infections are found throughout West Africa and group B infections are mainly localized in Côte d'Ivoire[32,33]. The exact number of HIV-2 group A and B infections remains unknown but it is estimated to be in the region of 1-2 million cases[6]. The other HIV-2 groups have almost all been seen in only one individual, suggesting that the transmitted virus has failed to spread beyond the initial patient[34]. An exception to this is group F, where a second patient was identified and showed symptoms of HIV-2 disease progression[35]. Of the single-patient HIV-2 groups, C, G and H cluster with SIVsmm strains from Côte d'Ivoire, group D clusters with an SIVsmm strain isolated in Liberia and groups E and F are closely related to SIVsmm strains found in Sierra Leone[34,30]. In all cases, the HIV-2 group and closest SIVsmm sequence were identified in the same country, suggesting that these groups represent 'dead-end' transmissions that failed to spread far past the initial human host. Additionally, an HIV-2 strain was identified that did not fall into any of the previously identified HIV-2 groups. A lack of whole genome sequences for groups C, D, E and F means this has not yet been formally classified as a new group. The isolation of this virus from an 8 year old boy may suggest on-going zoonotic transmission of diverse strains of SIVsmm in West Africa[36].

An HIV-2 recombinant virus, CRF01_AB, has also been described. CRF01_AB has genomic fragments from both HIV-2 group A and HIV-2 group B viruses. CRF01_AB has been identified in 3 individuals in Japan, two of whom are African

(a Nigerian and a Ghanaian)[37]. A single isolate, 7312A, with the same AB recombination pattern was described in 1994 in a patient from Côte d'Ivoire, implying that HIV-2 CRF01_AB may be more widespread than current sampling resolution indicates[38].

Although there is strong phylogenetic support for the SIVsmm origin of HIV-2 group A[28,39], the exact source and location of the zoonosis remains unclear. Molecular clock calibration using *env* sequences of known sampling date shows HIV-2 group A originated around 1938 (95% Bayesian credible intervals 1928-1947)[40]. Standard modelling, which assumes equal transition rates to and from West Africa, shows the most probable root of the HIV-2 group A epidemic to be either Caió, a rural village in Guinea-Bissau with a high HIV-2 prevalence, or Portugal. The identification of Portugal as the origin of HIV-2 group A is not surprising as historical links between Portugal and Guinea-Bissau lead to shared viral diversity. However, the lack of sooty mangabeys in Portugal means this is obviously an incorrect conclusion, caused by an over-representation of HIV-2 sequences from Portugal in the analysis. Analysis conducted using a location prior that favours migration of HIV-2 A out of Africa, protecting against non-African sampling bias, places Caió at the root for *env* and splits the root between Caió and Côte d'Ivoire for *pol*. This shows that oversampling from Caió may obfuscate the estimation of the spatial root of HIV-2 A. Further sampling of HIV-2 A sequences will allow more accurate identification of the geographic root of the epidemic, however, it is clear that Guinea-Bissau, Senegal and Côte d'Ivoire acted as key hubs early in the epidemic. In addition to the differences in sampling resolution in West Africa, the extinction of sooty mangabeys in Guinea-Bissau and Senegal means it is not possible to examine whether SIVsmm in Guinea-Bissau is closer to HIV-2 A than strains from Côte d'Ivoire[25].

## 1.3 The structure of the HIV-2 virion

HIV-2 is a retrovirus of the genus *Lentivirus* that belongs to the family *Retroviridae*[41]. Lentiviruses are named for the slow rate of disease progression following infection and are capable of infecting non-dividing cells. Retroviruses

are a family of Ribonucleic acid (RNA) viruses and their name reflects the retrograde flow of information they perform, transcribing RNA into Deoxyribonucleic acid (DNA) during integration into the host genome.

Mature HIV virions are spherical viruses, approximately 100-150nm in diameter (**Figure 1.4**).

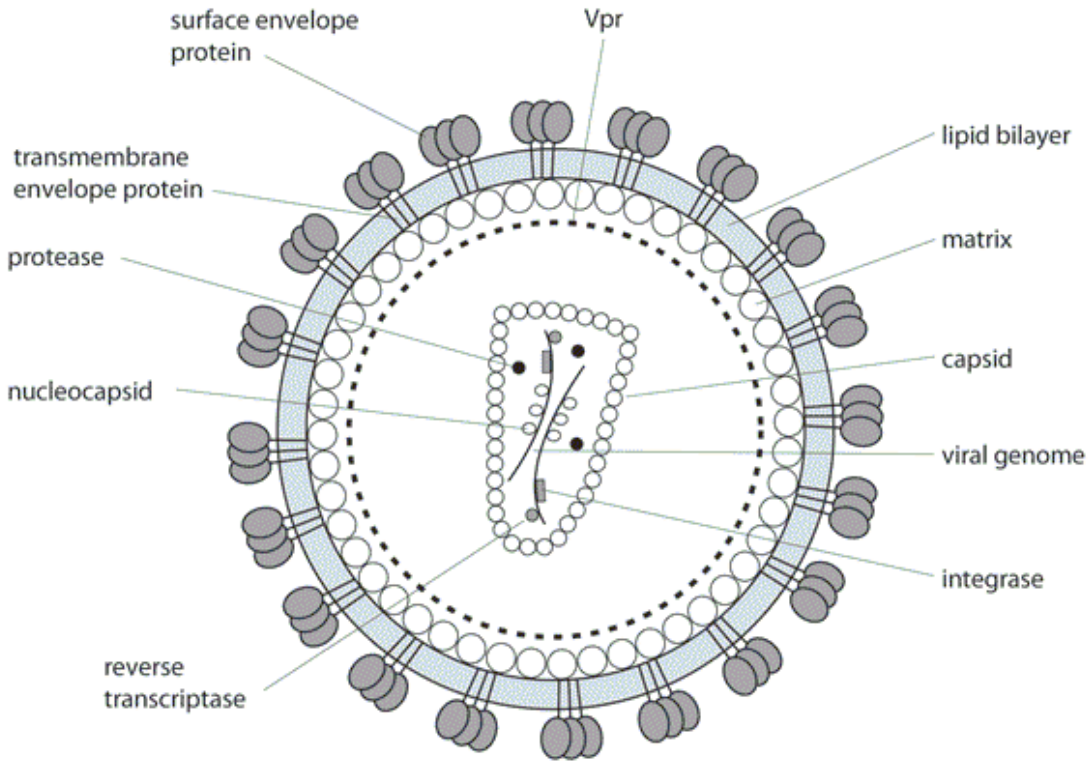



Figure 1.4: Structure of the mature HIV virion.
Schematic diagram of the HIV virion, showing the location and arrangement of the major structural proteins (adapted from [42])

At the core of the virion is the viral RNA genome. Genomic stability is maintained by association with the nucleocapsid (NC) protein p7[24]. In addition to providing genomic stability, NC also aids packaging of genomic RNA into nascent virions. The viral genome is contained within the capsid, a conical structure composed of the viral protein p26. Alongside the viral genome, the capsid also contains several viral proteins critical for replication such as reverse transcriptase (RT), integrase (IN) and protease (PR). The matrix, composed of the viral protein p17, maintains structural stability and integrity of the virion. The viral accessory proteins vpr, vpx, vif and nef are all contained within the matrix in addition to several host cell proteins. The matrix is surrounded by a host-derived

phospholipid bilayer, incorporated during budding and associated with myristoylated residues in the matrix. The surface of the virion contains membrane bound host proteins such as Major Histocompatibility Complex (MHC) molecules and heterotrimeric complexes of the viral envelope proteins gp120 and gp41, anchored by the trans-membrane domain of gp41[43,44]. Gp120 trimers form spikes, which extend from the surface of the virion. Mature HIV-1 virions are estimated to have 4-35 gp120 spikes on the surface: the number on HIV-2 is still unknown, however, it is assumed to be similar[45]. In addition to the ability of *env* to accumulate mutations rapidly in the variable (V1-V3) regions, abrogating binding of neutralising antibodies (Nab) to the virion, gp120 trimers also facilitate the acquisition of sugar moieties, known as a glycan shield, which is dynamic and is thought to provide steric hindrance to Nab binding[46].

## 1.4 The HIV-2 genome

HIV-2 has two copies of a positive sense single-stranded RNA (ssRNA) genome per virion. The genome is approximately 10kb in length and encodes 9 genes, all within 3 open reading frames (ORFs)[47]. Flanking the ORFs are two long terminal repeats (LTRs) containing *cis*-acting elements such as transcriptional regulatory elements, RNA processing signals and integration and packaging sites. In spite of the different origins of HIV-1 and HIV-2, the genomes of the two viruses share a similar genetic architecture and relatively high levels of sequence identity at the amino acid level (**Figure 1.5**)[48]. Homology is approximately 60% in gag and pol and around 40% in env[4]. The major difference in architecture is the presence of the accessory gene *vpx* in HIV-2, a gene that is specific to the HIV-2/SIVsmm lineage[49]. In contrast, the sequence homology between HIV-2 and SIVsmm is 75-90%, reflecting the origin of HIV-2 within the radiation of SIVsmm strains[39].

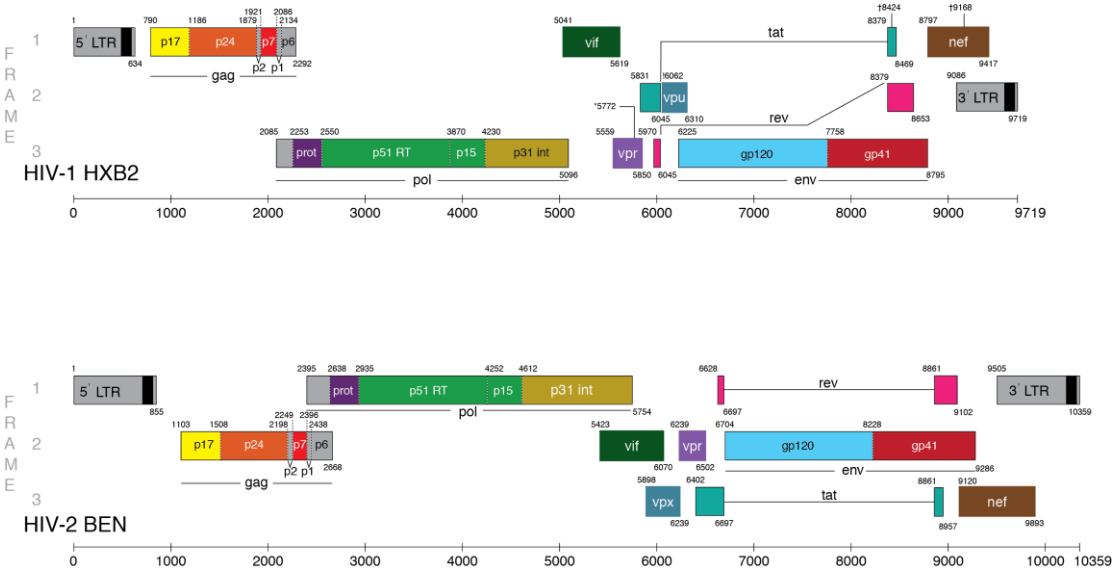



Figure 1.5: Genomes of HIV-1 and HIV-2.
ORFs are shown as rectangles and coloured according to the mature proteins produced. Gene positions are annotated by the location of A in the ATG start codon relative to the distal end of the 5′ LTR (adapted from [24])

The major structural genes of the HIV-2 genome are Group Specific Antigens (*gag*), Polymerase (*pol*) and Envelope (*env*). *Gag* is expressed as a 55 kDa poly-protein precursor which is cleaved to form p17 (Matrix), p26 (Capsid) and p7 (Nucleocapsid)[50]. *Pol* is also expressed as a poly-protein precursor ($Pr160^{Gag\text{-}Pol}$) which is cleaved to form protease (PR), integrase (IN) and reverse transcriptase (RT)[51]. *Env* is expressed as the gp160 precursor and is processed to produce heterotrimeric complexes of the external glycoprotein gp120 and the trans-membrane glycoprotein gp41[52].

Alongside the structural genes, the HIV-2 genome also contains two essential regulators of viral expression: transactivator of transcription (*tat*) and regulator of expression of virion proteins (*rev*). Tat is found in two forms; the 72 amino acid minor form (Tat-1) and the 86 amino acid major form (Tat-2). Tat is constitutively expressed in persistently infected cells, even in the absence of active viral production, and acts by binding to the trans-activation response (TAR) element in the LTR, promoting elongation of transcription through the AATAAA polyadenylation signal in the 5′ LTR. This ensures production of full-length transcripts of the viral genome[53]. Rev is a 19kDa phosphoprotein that shuttles rapidly between the nucleus and cytoplasm of host cells. The role of rev is to bind

to the rev responsive element (RRE) in *env*, promoting export of unspliced messenger RNAs (mRNAs) from the nucleus, allowing post-transcriptional processing[54].

HIV-2 also contains 4 accessory genes. These are: viral infectivity factor (*vif*), viral protein R (*vpr*), viral protein X (*vpx*) and negative factor (*nef*). The roles of these genes are summarised below and will be discussed in more detail later in this chapter.

Table 1.1: HIV-2 accessory genes.

| Gene | Function | Reference |
|---|---|---|
| *Vif* | Counteracts APOBEC3G | 55 |
| *Vpr* | Induces G2 cell cycle arrest and apoptosis | 56 |
| *Vpx* | Counteracts SAMHD1 | 57 |
| | Counteracts APOBEC3A | 58 |
| | Counteracts IRF5 | 59 |
| | Promotes nuclear import of pre-integration complex (PIC) | 56 |
| *Nef* | Down regulates CD4 expression | 60 |
| | Down regulates HLA expression | 61 |
| | Induces IL-2 production | 62 |

A summary of the major roles of the HIV-2 accessory proteins

## 1.5 The HIV-2 life cycle

Like other retroviruses, the lifecycle of HIV-2 has 9 distinct phases, each directed by proteins encoded in the viral genome and often involving the hijack of processes of the host cell. These steps are receptor binding, entry into the cell and un-coating, reverse transcription of the viral genome, nuclear import, integration into the host genome, transcription and translation of virally encoded proteins, assembly of nascent virions, budding and finally maturation (**Figure 1.6**).

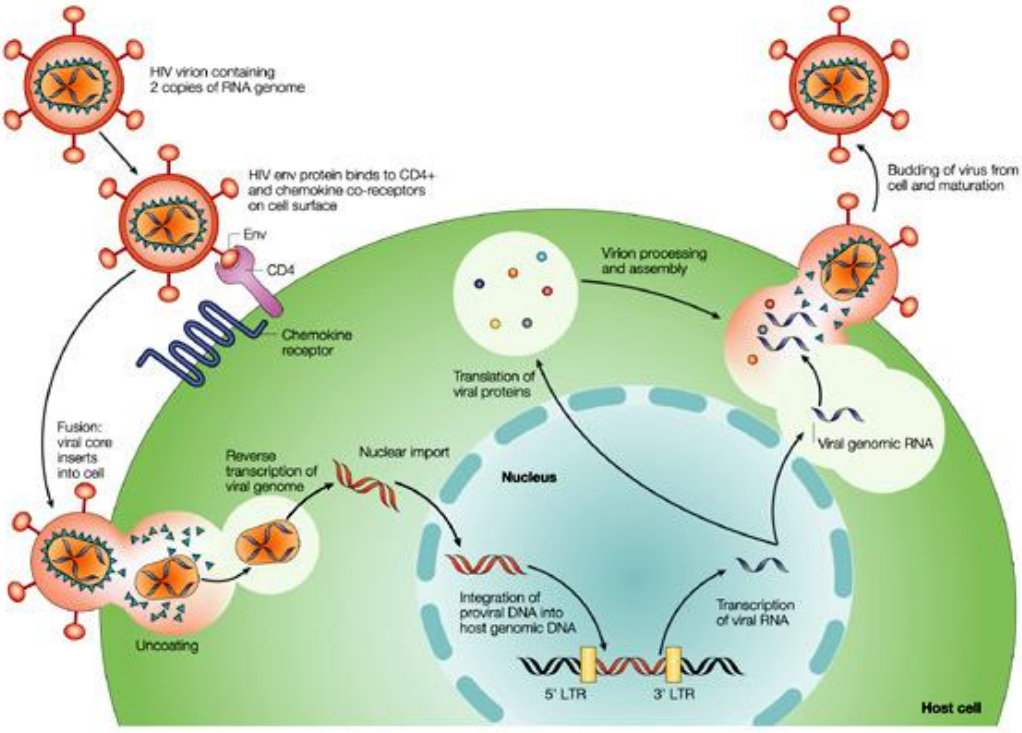



Figure 1.6: Lifecycle of HIV.
Schematic diagram summarising the lifecycle of HIV. Viral RNA is shown in dark blue, viral cDNA in red and host DNA in black (adapted from [63])

Like HIV-1, the major targets of HIV-2 are CD4+ T cells and macrophages. The majority of free virions in the blood interact with the CD4 receptor on the surface of target cells via the trimeric viral envelope spike (composed of 3 subunits of gp120) in the presence of a co-receptor. Occasionally non-CD4 usage has been reported for HIV-2[64]. HIV-2 is able to utilise a broad range of co-receptors including GPR15, CXCR6, CCR1, CCR2b, CCR3, CCR4, CCR8, CCR5 and CXCR4[65],[66],[67]. Co-receptor usage is referred to as the tropism of the virus and viruses that use CXCR4 are referred to as X4 whereas those using CCR5 are R5. In HIV-1 infection, CCR5 tropism is thought to be a fundamental property of transmitted-founder (TF) viruses, and a switch to X4 tropism occurs later in infection through gradual mutations in *env*[68]. This is supported by the observation that a naturally occurring 32bp deletion in the CCR5 chemokine receptor (CCR5Δ32) that prevents cell surface expression is associated with resistance to HIV-1 infection[69]. The low frequency of CCR5Δ32 in African populations means that the effect of this mutation on HIV-2 susceptibility *in vivo* cannot be verified[70]. Studies have shown that X4 tropic and R5/X4 dual tropic viruses are much less

commonly seen in HIV-2 infection[66],[71]. HIV-2 isolates from aviraemic individuals are able to use a broad range of co-receptors and therefore it seems that a broad range of co-receptor usage is not associated with pathogenicity for HIV-2.

CD4 binding causes a dissociation of the gp120 surface (SU) and gp41 transmembrane (TM) domains of env, bringing the fusion peptide of the TM domain into contact with the cell membrane[72]. Once the virion has fused to the cell membrane, the viral core enters the cytoplasm of the host cell where a step-by-step un-coating occurs. HIV un-coating is not well understood but seems to involve phosphorylation of the matrix: a recent study has shown that the recruitment of the host proteins CPSF6 and several cyclophilins by HIV-1 during un-coating and nuclear entry is important for the evasion of innate immune recognition[73],[74]. During the un-coating of the viral capsid, reverse transcription of the viral RNA genome begins. Un-coating leads to the formation of a reverse transcription complex which contains the virally encoded reverse transcription enzyme, a transfer RNA ($tRNA^{Lys}$) primer, NC and two copies of the ssRNA genome[75]. HIV has an extremely fast rate of mutation (approx. $4\text{-}9x10^{-3}$ subs/site/year), important for rapid evasion of host immune defences, and variation is generated during reverse transcription by two mechanisms[76],[63]. HIV-2 RT lacks the proofreading ability seen in many other polymerases and therefore introduces a large number of mutations during the replication cycle[77]. Additionally the two copies of the ssRNA genome undergo non-covalent dimerization in the virion particle[78]. Genome dimerization in HIV-2 is due to two sequence motifs in the genome, a self-complimentary sequence in the 5' end of the primer-binding site and SL1, a stem-loop structural element, homologous to SL1 in HIV-1[79]. HIV-2 RNA has been shown to transition from a loose to tight dimer following disruption of interactions with NC[80]. Genome dimerization is important as the close proximity of the ssRNA molecules allows template switching by RT, generating recombinant progeny that contain genomic fragments from both 'parental' genomes[81]. Reverse transcription takes place in the cytoplasm of the host cell and efficiency is increased in HIV-2 infection due to the presence of the accessory protein vpx. Vpx antagonises the

host restriction factor SAMHD1, targeting it for degradation and therefore rendering the target cell permissive to HIV-2 infection[82]. This will be discussed in more detail later in this chapter.

The structures of the 5′ and 3′ LTR of the HIV-2 genome are semi-palindromic and each consist of 3 regions, untranslated 3′ (U3), repeat (R) and untranslated 5′ (U5) in proviral DNA and R and U5 only (5′ end) and U3 and R only (3′ end) in the ssRNA genome[24]. Reverse transcription begins at the cellular tRNA which binds just downstream of the U5 region in the 5′ LTR. Reverse transcription then proceeds in a 3′ → 5′ direction, synthesising a negative sense DNA strand of the U5 and R and creating a short RNA-DNA duplex[83], [84]. RNase H then degrades the RNA in this duplex, leaving a short DNA fragment that is complimentary to R in the 3′ end of the genome. This small DNA molecule acts as a bridge to transfer minus strand stop DNA to the 3′ end of the RNA genome in a process known as first strand transfer. Reverse transcription then proceeds in a 3′ → 5′ direction from the short DNA fragment, generating a complete negative sense DNA strand (minus the 5′ LTR). RNase H degrades the majority of the viral RNA genome during reverse transcription. The polypurine tract (PPT) just upstream of the 3′ LTR is not degraded and acts as an RT binding site to allow 3′ → 5′ positive strand synthesis which terminates at a modified alanine in the tRNA binding site. Another short RNA-DNA duplex is created and following RNA degradation, the resulting DNA is able to act as a plus strand stop DNA, binding to the complementary tRNA binding site in the 5′ LTR in the negative sense DNA strand. The positive sense DNA copy of the genome is then synthesised 3′ → 5′ and the negative sense strand is elongated to incorporate the plus strand stop DNA. The final result is a double stranded DNA copy of the viral RNA genome which is longer than the original as both 3′ and 5′ LTR regions now contain the U3-R-U5 strong stop sequence **(Figure 1.7)**[83].

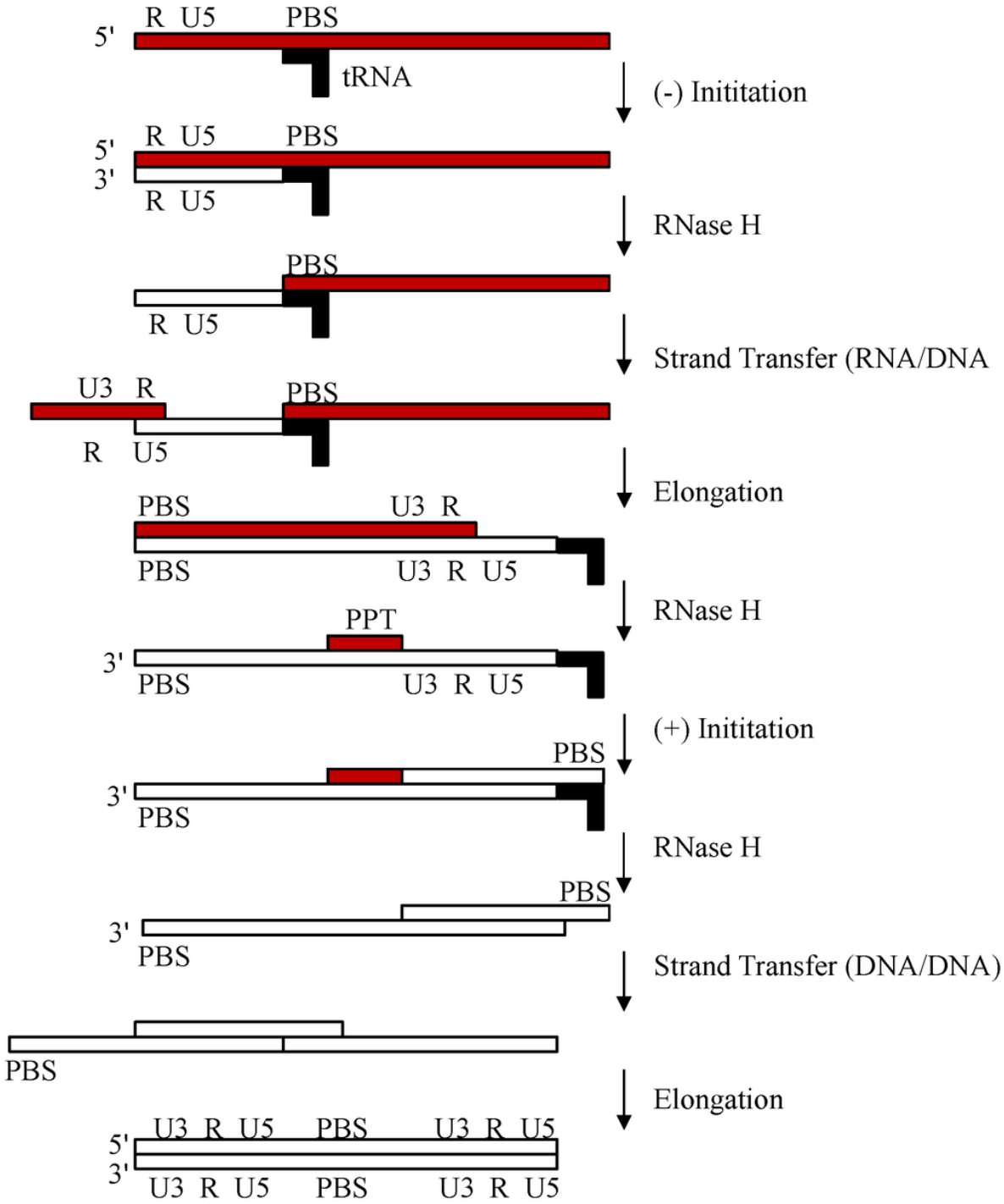



Figure 1.7: Reverse transcription of the HIV-2 genome.
Diagram showing the steps involved in reverse transcription of HIV-2 RNA into double-stranded cDNA. Viral RNA template is shown in red, host-derived tRNA molecules in black and newly synthesis viral cDNA in white (adapted from [85])

Following reverse transcription, the double stranded cDNA associates with uncoated viral core proteins and several host proteins to form a pre-integration complex (PIC). The PIC is a compacted molecule, allowing viral DNA to transverse through the nuclear pore. The viral protein integrase (IN) mediates integration of the viral cDNA into the host genome by removing a dinucleotide from both 3'

ends, exposing hydroxyl groups, which are used by IN to cut the host genomic DNA and allow ligation of viral cDNA. Integrated viral DNA is known as provirus. Integration can occur at many sites in the host genome, although sites of active transcription are usually favoured.

Once integrated, the transcriptional state of the proviral DNA is influenced by the exact genomic position and factors such as the epigenetic environment. The 5' LTR of the provirus contains a transcriptional promoter and *cis* elements needed for recruitment of host transcription factors such as NFκB to allow transcription by the host polymerase Pol II. Immediately following integration, there is a low level of proviral expression, even if integrated into a highly active locus. Due to the block in export of un-spliced transcripts, only the 'early' viral proteins tat, rev and nef, which are derived from fully spliced transcripts, are produced. Tat is a Pol II trans-activator that binds TAR in the 3' end of nascent viral mRNA. Binding of tat to TAR enhances transcription, leading to an up-regulation in expression of viral mRNA. Additionally, binding of Rev to the rev response element (RRE) mediates nuclear export of un-spliced and partially spliced transcripts encoding the 'late' viral proteins gag, pol, env, vpx, vpu and vif.

Transcription of the proviral DNA is initiated in the U3 regions of the LTR and elongation continues until the polyadenylation signal. Initially a whole length mRNA is transcribed. A proportion of these full-length transcripts are directly exported from the nucleus to be packaged into progeny virions. The remainder are spliced in the nucleus. *Gag* and *pol* are transcribed as a single mRNA transcript. The vast majority of transcripts (>90%) are used for translation of the $Pr55^{Gag}$ poly-protein precursor. A -1 ribosomal frame shift at the poly T slippage site leads to the translation of the $Pr160^{Gag\text{-}pol}$ precursor from the remaining transcripts. The $Pr55^{Gag}$ poly-protein precursors are targeted to the host cell membrane, which promotes virion assembly at the host cell membrane, leading to an accumulation of processed trimeric env proteins. Alongside the gag and pol precursors, 2 copies of the ssRNA genome and accessory proteins are packaged into each nascent virion. Budding is achieved by hijack of the host cellular export machinery. Following

budding of the immature virion, $Pr55^{Gag}$ is cleaved into p17, p26, p7 and p6 as well as the spacer proteins p1 and p2 by the virally encoded protease (PR). The cleavage of $Pr55^{Gag}$ promotes structural re-arrangements in the maturing virion, leading to an increase in structural stability and the correct conformation to allow infection of new target cells.

## 1.6 The natural history of HIV-2 infection

Like HIV-1, HIV-2 is an AIDS-causing virus. In contrast to HIV-1, where time of progression to AIDS following infection in Caucasian populations follows a normal distribution with a mean of 11 years for more than 99% of patients in the absence of treatment, HIV-2 disease progression appears to be bimodal and the majority of HIV-2 infected individuals do not show immunosuppression[86]. A longitudinal study of HIV-2 infected individuals from Caió showed that survival was highly correlated with HIV-2 viral load (VL). Patients with an undetectable VL had a similar survival probability to HIV uninfected individuals whereas those with a VL of >10,000 copies/mL showed a marked increase in mortality in a follow-up study that spanned almost 2 decades (**Figure 1.8**).

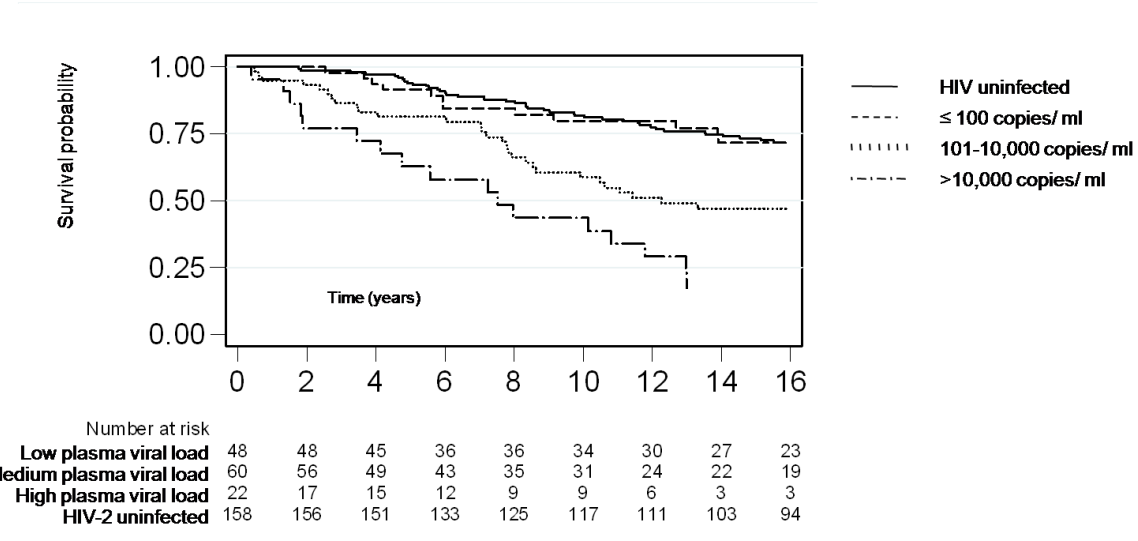



Figure 1.8: Summary of HIV-2 associated mortality.
Kaplan-Meier plot showing the relative survival probability of HIV-2 patients with undetectable plasma VL, viraemic HIV-2 patients and HIV uninfected individuals (adapted from [87])

The minority of patients (~20% in Caió) that behave as rapid progressors develop immunodeficiency that is characterised by high VL and a reduction in CD4+ T cell counts[88]. The clinical features of AIDS following HIV-2 infection are indistinguishable from those caused by HIV-1, apart from the unexplained reduction in cases of Kaposi's sarcoma[89]. Additionally, presentation with AIDS and death tends to occur at higher CD4+ T cell counts following HIV-2 infection when compared with HIV-1[90]. Overall there is a two-fold mortality risk associated with HIV-2 infection when compared to matched uninfected individuals, which is much lower than the 10-fold increase associated with HIV-1[91].

The proportion of patients who are Long-Term Non-Progressors (LTNPs) varies depending on which cohort is examined. In Caió, 37% of subjects recruited in 1989 maintained an undetectable VL for over a decade[87]. In contrast, a French follow up study on the ANRS CO5 clinical cohort reported only 6.1% of patients behaving as LTNPs (as defined by a follow up time of at least 8 years with a CD4+ count of >500 cells/μl). When a VL of <500 copies/ml for at least 10 years was used as the definition, 9.1% of the cohort were LTNPs[92]. The observed differences in the proportion of LTNPs between these two cohorts are relatively easy to explain. Caió is a community cohort, established following regular sero-surveys of the entire adult population, and therefore there is no bias towards recruitment of patients who present with symptoms or are receiving treatment for immunosuppression. However, there is likely to be a lack of rapid progressors who may have become infected then lost to AIDS between sero-surveys. The converse is true for the ANRS CO5 clinical cohort, which is biased towards patients receiving treatment for HIV-2 induced illness. The small number of HIV-2 cohorts with sufficient follow-up time means the true number of LTNPs has yet to be established, although it is significantly higher than for HIV-1. A better estimation of the proportion of LTNPs would establish whether the differences between the ANRS CO5 and Caió cohorts were due simply to cohort recruitment biases or whether there was also an effect of host or viral genetic factors.

In contrast to HIV-1 LTNPs (with the exception of extremely rare elite controllers), HIV-2 LTNPs have undetectable plasma VLs[93,94]. In Caió, many LTNPs maintain this phenotype over at least two decades and have no increased mortality risk when compared to matched HIV uninfected individuals[87,95]. Additionally, HIV-2 patients with high CD4+ counts (>500 cells/μl or >28% total lymphocyte count) have a significant survival advantage when compared to HIV-1 patients matched for disease stage[96]. Plasma VL is the most reliable marker of disease progression in HIV-2 and when HIV-1 and HIV-2 patients are matched for baseline plasma VL, the median annual rate in CD4+ T cell decline is equivalent[97]. When HIV-1 and HIV-2 patients are matched for disease stage, HIV-2 patients have a significantly lower plasma VL whereas the proviral loads are similar in both HIV-1 and HIV-2[98,99,100]. As viral shedding is closely related to plasma VL, this may explain the transmission patterns of HIV-2, where the levels of vertical and horizontal transmission are both significantly lower than for HIV-1.

## 1.7 The epidemiology of HIV-2

In contrast to the HIV-1 M global pandemic, HIV-2 A infections have remained largely confined to West Africa (**Figure 1.9**). Where global dispersion of the virus is seen, it tends to follow colonial routes and historical migration patterns. The highest prevalence outside West Africa is seen in Portugal and epidemiological analysis links these cases to Guinea-Bissau and Cape Verde. Clinical cohorts are also established in the UK and France, with French cases appearing to be derived from strains seen in Côte d'Ivoire and Senegal. There is strong support for a link between the Portuguese and British epidemics. Additionally HIV-2 has been reported in Mozambique, Angola, India and Brazil which all have ex-colonial links to Portugal[101,102]. HIV-2 has also been reported in Japan; however, the majority of infections are in individuals of African origin[103]. HIV-2 reached peak prevalence in Guinea-Bissau in the late 1980s when a community study of 100 randomly selected houses showed prevalence rates of 8.9% in 1987 and 10.1% in 1989[104]. Prevalence was highest amongst women in the 45-60 year age group, with rates of infection of close to 20%. Subsequent studies have all shown a decrease in the prevalence of

HIV-2 whilst the cases of HIV-1 in West Africa continue to rise[105]. A prospective study of female sex workers (FSW) in Senegal showed HIV-2 decreasing from 8% in 1985 to 5.5% in 2003 whilst HIV-1 prevalence increased from 1% to 13.8%[106]. Similar patterns of prevalence have been seen in Caió, Guinea-Bissau. An initial sero-survey in 1988 showed HIV-2 infection in 8% of the population, a figure that had dropped to 4.7% by 2007[107]. HIV-1 infections rose from <1% in 1988 to 3.6% in 2007. Modelling of the HIV-2 epidemic in Caió predicts that prevalence will fall below 0.1% in 2053 with no new infections after 2047 (95% C.I. 2029-2083)[108]. Ultimately HIV-2 is expected to become extinct in Caió before the end of the century, in 2071 (95% C.I. 2054-2106).

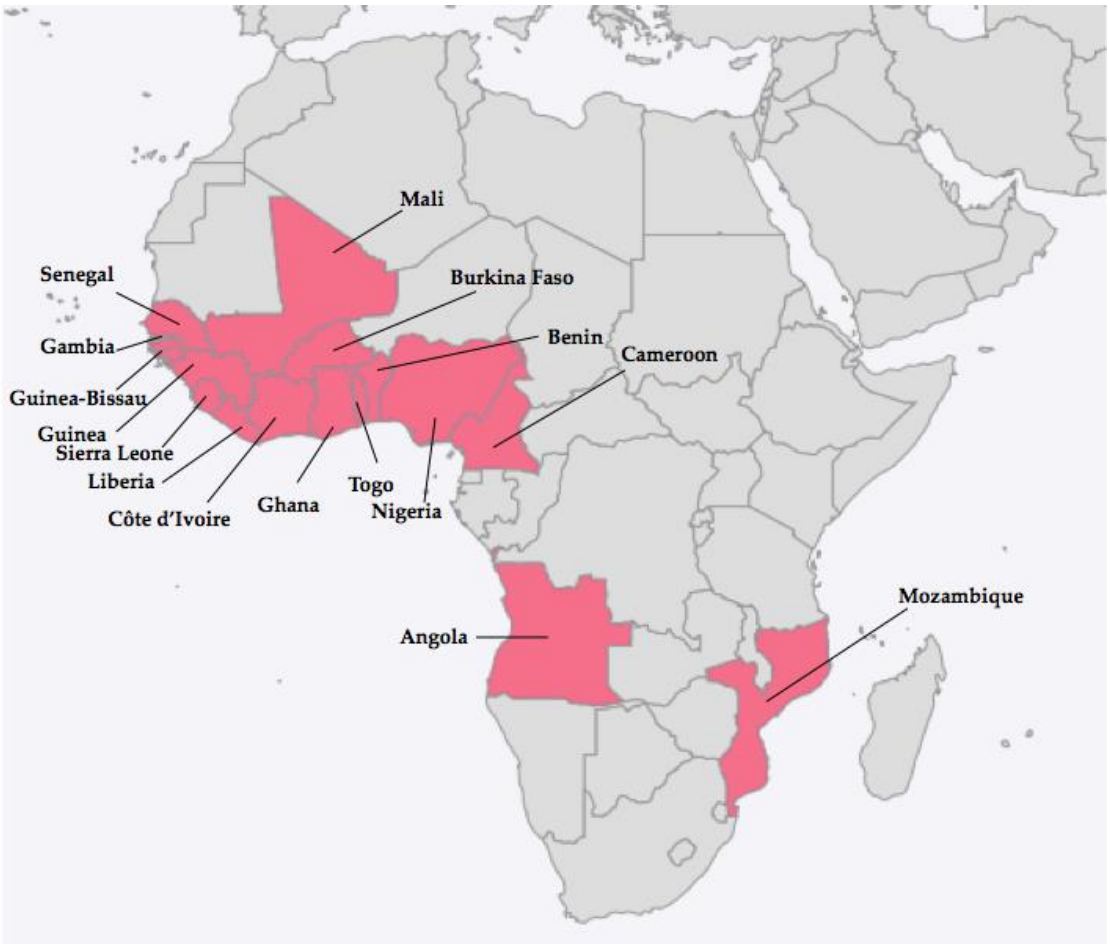



Figure 1.9: The West African HIV-2 epidemic.
African countries with reported cases of HIV-2 infection and transmission are shown in pink.

Whilst the decline in HIV-2 can be easily explained due to the low transmissibility of HIV-2 compared to HIV-1, what seems more surprising is that the epidemic ever reached such high levels in the first place. Vertical transmission

of HIV-2 in breast-feeding mothers who are not receiving ART is less than 4% compared to a risk of 24.4% for HIV-1[109]. Similarly, the risk of sexual transmission is highly correlated with plasma VL, which tends to be significantly lower in HIV-2 patients when compared with matched HIV-1 cohorts[110]. Reduced risk of sexual transmission is also explained by lower levels of HIV-2 shedding in genital secretions compared to HIV-1 and the fact that HIV-2 levels in semen have been shown to be closely correlated with plasma VL[111,112]. Skyline analysis suggests that following the initial zoonosis, HIV-2 underwent a period of low endemicity, with a relatively constant population size, followed by an exponential growth in infections[28]. The transition from constant population size to exponential growth occurred between 1955-1970, which overlaps with the dates of the war of independence in Guinea-Bissau (1963-1974)[113]. It has been hypothesised that the displacement of people and increase in prostitution associated with the war aided the spread of HIV-2 and it certainly seems likely that Portuguese troops who fought in Guinea-Bissau facilitated the transmission of HIV-2 to Portugal. However, treatment regimes beginning in the 1940s, particularly those against tuberculosis and trypanosomiasis, may have also contributed to an inadvertent iatrogenic spread of HIV-2[114]. There is a high prevalence of HIV-2 in women born before 1960, which could be due to several factors. These individuals are likely to have reached sexual maturity during the war, putting them at an increased risk, and similarly may have received treatment with contaminated syringes. However, these factors do not explain the gender bias towards women. A clinic-based study in the Gambia showed an increased survival time for HIV-2 infected women when compared to HIV-2 infected men, raising the possibility the bias is due to an over-representation of HIV-2 widows[115]. It is also possible that age-related hormonal changes affecting the vaginal mucosa lead to a higher susceptibility of infection in post-menopausal women[116]. Another important factor in the decrease in HIV-2 prevalence in West Africa is the emergence of the HIV-1 epidemic. Simulations of heterosexual transmission in rural Guinea-Bissau estimate that approximately 30%

of the decrease in HIV-2 cases is due to competitive elimination by HIV-1, although HIV-2 is still more prevalent than HIV-1 in Caió[117].

There are still cases of new HIV-2 infections in West Africa, even though the overall prevalence is decreasing. The most likely source of new cases is heterosexual transmission: however, not much is known about the on-going spread of HIV-2 in West Africa and many national HIV testing strategies fail to differentiate between HIV-1 and HIV-2 infection. Lack of access to clinics and long periods of asymptomatic infection may also mask the true prevalence, particularly in rural communities.

The history of the West African HIV-2 epidemic is markedly different to that of the global HIV-1 pandemic. Whereas HIV-1 has a global distribution and the number of infections continues to rise, HIV-2 has remained largely restricted to West Africa; following a period of exponential growth in the mid to late 20$^{th}$ century, prevalence is now falling at such a rate that extinction is predicted before the end of this century. This is undoubtedly a product of the differing pathogenicity of the two viruses but it also highlights how informative HIV-2 may be for the identification and exploitation of factors that could be implemented to curb the HIV-1 global pandemic.

## 1.8 HIV-2 and HIV-1 dual infection

Co-infection or dual infection of HIV-1 and HIV-2 (HIV-D) in the same individual is a relatively rare but highly informative event. Early analysis of dual infected individuals appeared to show that, in a cohort of HIV-2 positive FSW, HIV-2 infection conferred a protective effect of approximately 70% against subsequent sero-conversion to HIV-1 when compared to HIV-2 negative individuals. However, subsequent studies failed to repeat this finding and a later sero-prevalence study showed a significantly higher risk of HIV-1 acquisition for HIV-2 infected patients compared to HIV un-infected individuals when matched for age and risk factors (relative risk ratio >3)[118],[119],[120]. Modelling of the HIV-2 epidemic in Guinea-Bissau estimates that approximately 30% of the decrease in HIV-2 prevalence can be attributed to competitive elimination by HIV-1 infection,

demonstrating the absence of a globally protective effect of HIV-2 infection in the context of HIV-1 associated mortality[117]. In recent years, however, the debate over the effect of dual infection on HIV disease progression has been re-ignited. A longitudinal study of HIV-1 and HIV-D infected individuals in the Guinea-Bissau police cohort suggested a inhibitory effect of HIV-2 infection on disease progression in the context of subsequent HIV-1 co-infection[121]. Individuals who acquired HIV-1 after HIV-2 (2→D) showed a longer progression time to AIDS (129 months compared to 68 months for HIV-1 infection only). In addition to a significantly longer progression time, 2→D individuals also showed higher CD4+ T cell percentages, a reduction in annual CD8+ T cell increase and lower HIV-1 genetic diversity[122]. Additionally, several *in vitro* studies have shown an effect of HIV-2 infection on HIV-1 infectivity and replication[123],[124]. However, a meta-analysis of 6 studies on HIV-D individuals showed no difference between time to death following HIV-1 infection for HIV-1 and HIV-D infected patients[125]. Analysis of 259 HIV-1-seroincident cases taken from the Guinea-Bissau police cohort, using time to death as the measurement of HIV-1 disease progression showed that individuals who acquired HIV-2 prior to HIV-1 showed a 42% longer time to mortality when compared to HIV-1 single-infected individuals[122]. Therefore, there is evidence supporting a protective effect of HIV-2 infection on disease progression following subsequent HIV-1 infection. However, the mechanisms behind this protective effect have yet to be shown and further sampling on well-defined cohorts with known HIV-1 and HIV-2 infection dates is needed to expand the knowledge of this phenomenon.

## 1.9 Correlates of immunity in HIV-2 infection

### 1.9.1 Immune activation in HIV-2 infection

The role of immune activation in HIV-induced disease was initially postulated following the observation that non-pathogenic SIVsmm infection leads to little or no immune activation in sooty mangabeys, even in the presence of high levels of plasma viraemia[126]. In contrast, chronic immune activation, as measured by the up-

regulation of CD38 and HLA-DR on CD4+ T cells, shows a significant correlation with disease progression in HIV-1 infected patients[127]. Similarly, in HIV-2 infection, elevated levels of the soluble immune activation marker $\beta_2$-microglobulin is a strong diagnostic marker of AIDS-related death, showing that chronic immune activation is also a feature of HIV-2 induced AIDS[128]. Although immune activation is a characteristic feature of both HIV-1 and HIV-2 pathogenesis, overall levels of immune activation are reduced in HIV-2 infection relative to those seen in HIV-1 infection. Studies of immune activation markers and plasma VL levels in Caió showed a direct correlation between the two, with the majority of patients maintaining an undetectable VL also showing immune activation levels similar to uninfected individuals[129]. However, the identification of a few aviraemic patients showing a significant level of immune activation does suggest that there is another method of immune activation that is not driven by HIV-2 viraemia levels. HIV-2 and HIV-1 levels of immune activation have also been shown to be similar in the 5-6 years prior to death, suggesting persistent immune activation is a hallmark of HIV-2 induced AIDS[130].

The differential levels of immune activation in HIV-1 and HIV-2 are likely to be a factor of the ability of HIV-2/SIVsmm nef protein to down-regulate expression of the T-cell receptor (TCR)-CD3 complex on infected cells. HIV-1 nef completely lacks the ability to mediate suppression of T-cell activation via down-regulation of the TCR-CD3 complex, leading to an increased responsiveness of infected T-cells to activation in the presence of viraemia[131]. An association between immune activation and the ability of patient-specific nef variants to down-regulate TCR-CD3 was shown in Caió but no association was seen with clinical outcome[132]. When only viraemic patients were included in the analysis, a significant correlation was seen between TCR-CD3 down-regulation and preserved CD4+ T cell counts[133]. Therefore it seems that the ability of HIV-2 nef to down-regulate TCR-CD3 is an important factor in protection against chronic immune activation but the lack of a strong association between down-regulation and survival implies

many more factors must be involved in determining clinical outcome following HIV-2 infection.

Elevated immune activation is thought to drive increased T-cell turnover in HIV-1 infections, which, when paired with decreased thymic activity, leads to an inability to replace T-cells lost through apoptosis. In HIV-2, T-cell apoptosis is reduced, compared with HIV-1 infection, particularly in LTNPs, and thymic function may be enhanced in HIV-2 infected patients, suggesting a better ability to control T-cell homeostasis[134]. However, this effect is only seen in aviraemic individuals and HIV-2 patients with high viral loads have been shown to have similar, or even higher, rates of T-cell turnover and immune activation to HIV-1 patients[135].

### 1.9.2 Neutralising antibodies

One of the major aims of an HIV vaccine would be the ability to generate broadly neutralising antibodies (Nabs) against the envelope gp120 trimer. There is evidence that Nabs maybe enhanced in HIV-2 infection but this observation has yet to be linked to clinical outcome[136]. Differential neutralisation sensitivity has been observed between HIV-2 isolates, which may be related to the ability to enter cells independently of CD4[64]. A small study showed a lower neutralising titre in AIDS patients and longitudinal studies have shown a lower level of neutralisation escape than is seen in HIV-1[137]. The lack of escape could be due to higher proviral reservoirs, lower replication rates in HIV-2 or negative selection pressure acting on *env*[138],[139]. HIV-2 also seems to elicit a broader Nab response but with lower potency than HIV-1, which is more likely to be related to a higher level of conservation in *env* than a better preservation of memory B-cells[140],[141].

### 1.9.3 T-Cell responses

The critical role of CD8+ T cells in controlling viraemia in acute HIV-1 infection is well described[142],[143]. CD8+ depleted rhesus macaques challenged with SIVmac showed a dramatic increase in viraemia and associated disease progression when compared to animals with an intact T cell repertoire, showing the importance of CD8+ T cells in controlling HIV/SIV replication[144]. Studies of HIV-1 elite controllers

(patients who maintain an undetectable VL in the absence of ART) have shown enrichment for the Human leucocyte antigen (HLA) class I alleles HLA-B*57 and HLA-B*27, demonstrating the role of differential HLA restriction in the control of viral replication[145,146]. HLA-associated viral escape mutations that reduce or abrogate HLA binding of specific viral epitopes provide evidence of strong selection pressure exerted on HIV-1 by the cytotoxic T lymphocyte (CTL) response[147,148]. HIV-1 specific CD8+ T cell quality has also been shown to have a role in disease progression. CD8+ T cells from patients with advanced disease secrete interferon-γ (IFN-γ) alone whereas CD8+ T cells from elite controllers are more likely to secrete multiple soluble factors (polyfunctional responses) in response to antigen stimulation and maintain a higher proliferative capacity[149,150].

The magnitude, frequency and quality of HIV-specific CD8+ T cell responses have been shown to vary between HIV-1 and HIV-2 infection. HIV-2 LTNPs exhibit strong polyfunctional T-helper responses and CTL responses against HIV-2, and gag appears to be the most immunogenic HIV-2 protein[151]. The IFN-γ response to gag was also significantly greater in patients with low plasma VL. The magnitude of gag-specific responses was inversely correlated to plasma VL and targeted to the major homology region of the HIV-2 capsid protein, p26, which appears to contain six immunodominant epitopes. The differentiation phenotype of HIV-2 specific T cells is also thought to have an impact on function and viral control. CD8+ T cells targeting an HLA-B*35 restricted HIV-2 capsid epitope showed a high level of functional avidity and appeared to be at an earlier stage of differentiation than HIV-1-specific CD8+ T-cells[152]. A recent study of HIV-2 LTNPs and progressors showed that HIV-2 control was strongly associated with an avid and polyfunctional gag-specific CD8+ T cell response and that these cells are at an earlier stage of differentiation than Cytomegalovirus (CMV)-specific CD8+ T cells[153]. Gag-specific T cells also showed low levels of activation and proliferation, a feature that is in contrast to CD8+ T cells associated with HIV-1 control[154].

## 1.10 Viral genetic factors involved in disease progression

### 1.10.1 *Vpx* and SAMHD1

Vpx is a 12-16 kDa accessory protein encoded by viruses of the HIV-2/SIVsmm and SIVrcm/SIVmnd-2 lineages[155,156,157]. *Vpx* is homologous to *vpr*, a gene encoding an accessory protein that is found in most primate lentiviral lineages (including HIV-1 and HIV-2)[158,159]. Early work showed that *vpx* was not necessary for HIV-2 infection of immortalised lymphocyte cell lines such as HUT78, CEM and SupT1 or the monocyte derived cell lines HL60 and U937[160,161,162]. In contrast, when *vpx* mutant viruses were grown in monocyte derived macrophages (MDMs), a significant replication deficiency was seen[163,164]. *Vpx* knock-out mutants also show reduced replication in activated peripheral blood mononuclear cells (PBMCs) and primary T cells[165,166]. *Vpx* is critical for HIV-2 replication in non-dividing or slow-dividing cell lines but appears to be dispensable in rapidly dividing lines. Vpx is packaged into virions via interaction of the third helical domain of vpx with the p6 domain of gag[167]. Vpx proteins are packaged into the virion at much higher levels than vpr, a typical mature HIV-2 virion contains approximately 4000 copies of vpx whereas the typical HIV-1 virion contains only 14-18 copies of vpr[168]. The association of vpx with mature viral cores at high copy number also suggests vpx is most likely to be involved with the early stages of the viral life cycle. Following infection, vpx localises to the nucleus of the host cell via an 8 amino acid nuclear localisation signal (NLS) in a process that appears to be modulated by vpx phosphorylation[169,170,171].

The host restriction factor SAMHD1 was identified as the target of vpx antagonism using tandem affinity purification and binding partner pull down assays[57,82]. SAMHD1 is a protein consisting of a sterile alpha motif (SAM) domain, an HD/COG1078 domain and a largely uncharacterised C-terminal region[172,173]. SAMHD1 is a GTP/dGTP dependent dNTP triphosphohyrolase[174]. Binding of GTP or dGTP to the allosteric domain of SAMHD1 results in a conformational change that allows recruitment of canonical and non-canonical dNTPs to the active site[175].

In addition to hydrolysing dNTPs, SAMHD1 is also able to act as a single stranded nucleic acid exonuclease. Hydrolysis of dNTPs by SAMHD1 results in a reduction in the concentration of free dNTPs from 1-15μM in dividing cells to 20-70nM in non-dividing cells[176]. The reduction in dNTP concentration is thought to be a major block to retroviral reverse transcription in myeloid cells and resting T cells, making these cell types resistant to retroviral infection[177,178,179]. SAMHD1 also acts to block the replication of DNA viruses such as vaccinia and herpes simplex viruses as well as retroelements[180,181,182].

Vpx targets SAMHD1 for degradation by the proteasome via recruitment of the $CRL4^{DCAF1}$ E3 ubiquitin ligase complex (**Figure 1.10**). $CRL4^{DCAF1}$ is a complex of a RING finger H2 protein, Cullin 4 (CUL4), DDB1 and DCAF1[183,184]. The crystal structure of SIVsmm vpx showed a three-helical bundle stabilised by a zinc finger motif[185]. DCAF1 is a substrate-adaptor protein that interacts with vpx at three sites. DCAF1 is a disc-shaped protein which vpx binds tightly around. The interaction between vpx and DCAF1 creates a new binding pocket in DCAF1, allowing recruitment of SAMHD1 through the C-terminal domain. Mutagenesis studies have shown that 31 residues of the disordered C-terminal domain of SAMHD1 are necessary and sufficient for binding to the vpx-DCAF1 complex[183].

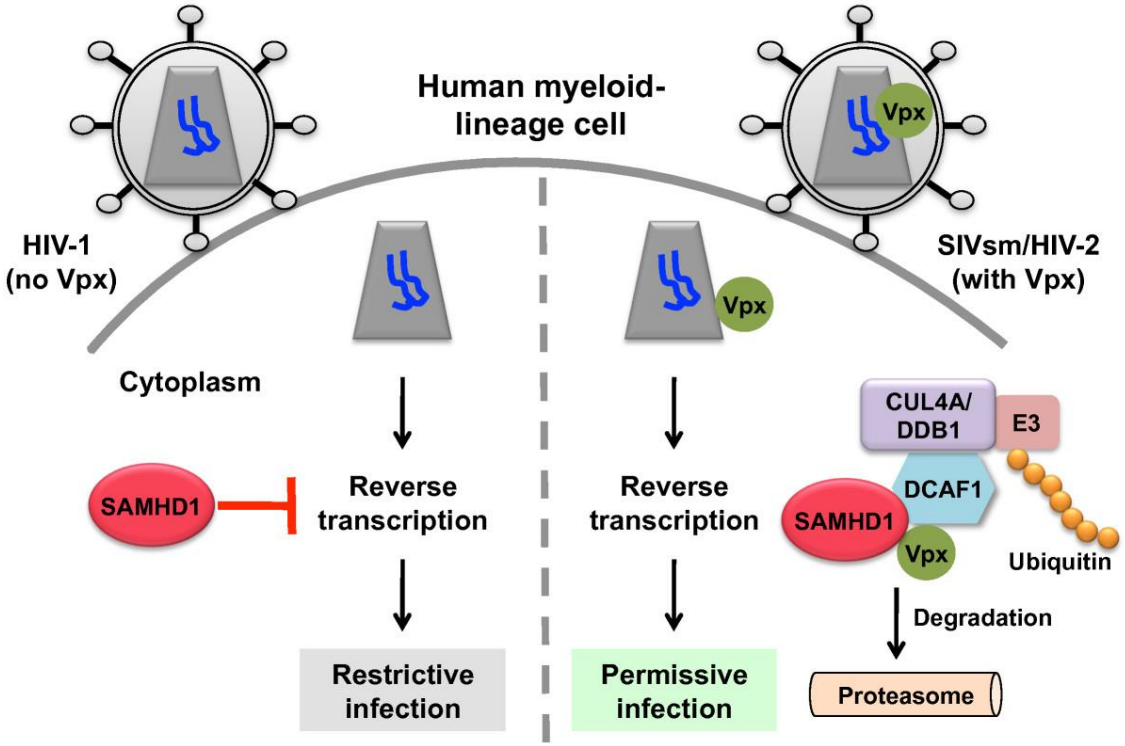



Figure 1.10: Illustration of the interaction between SAMHD1 and vpx. Schematic representation of the effect on a human myeloid lineage cell following infection with *vpx* +/- viruses (adapted from [186])

A naturally occurring SAMHD1 mutation that leads to knock out via a premature stop codon is a cause of Aicardi-Goutières Syndrome (AGS)[187]. AGS resembles congenital viral infection and is characterised by high levels of type 1 interferon (IFN)[188]. Resting CD4+ T cells derived from AGS patients permit higher levels of HIV-1 replication than those from healthy controls, showing that naturally occurring mutations in SAMHD1 can mimic the effect seen with the presence of *vpx*[189]. Similarly, small hairpin RNA (shRNA) knock down of SAMHD1 in THP-1 cells showed a 12-fold increase in HIV-1 infection, recapitulating the effects seen by infection with a *vpx*-containing virus[57]. Addition of exogenous deoxyribonucleosides (dN) has also been shown to increase accumulation of HIV-1 DNA in resting T cells, a cell type that is normally resistant to infection[190].

Restriction of HIV-1 infection by SAMHD1 appears to be related to the life cycle of the target cells, with cycling cells being more permissive than non-cycling or resting cells. However, the levels of SAMHD1 expression in MDMs, resting CD4+ and activated CD4+ T cells are similar[191]. SAMHD1 interacts with the cell cycle regulator cyclin-dependent kinase 1 (CDK1), which phosphorylates residue T592

of SAMHD1, leading to the prevention of lentiviral restriction[192,193]. CDK1 is inactive in resting cells. A phosphorylation defective mutant of SAMHD1 has been shown to be antiviral in both dividing and resting U937 cells (which lack endogenous SAMHD1) and the phosphomimetic SAMHD1 mutant T592E is unable to restrict HIV-1 infection in any cell line[194]. Although it seems likely that the antiviral activity of SAMHD1 against HIV-1 is closely linked to the cell cycle, the ability of phosphorylated and un-phosphorylated SAMHD1 to hydrolyse dNTPs does not appear to vary *in vitro*. Lack of SAMHD1 restriction in cycling cells may also be due to the extremely high rate of *de novo* dNTP synthesis in cycling cells, overwhelming SAMHD1 dNTP hydrolysis[195]. Recent work has suggested that the RNase activity of SAMHD1 is implicated in HIV restriction[196]. A naturally occurring SAMHD1 mutation Q548A, which abrogates RNase activity but not dNTP hydrolysis activity, causes a loss of HIV restriction *in vitro*. Additionally, a mutation that knocks out dNTP hydrolysis but not RNase activity was shown to have no effect on HIV restriction. In light of these findings it seems that SAMHD1 restriction of HIV-1 is a complex interplay of RNase activity and dNTP hydrolysis and the importance of each function may be related to the cellular environment. However, the picture of SAMHD1 restriction of HIV-1 remains unclear and further work is needed to elucidate the contribution of each enzymatic process to overall restriction in both resting and cycling cells[197].

In addition to antagonising SAMHD1, *vpx* also interacts with other host cell proteins. Members of the apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like 3 (APOBEC3) family are involved in the restriction of viral infection and HIV-2/SIVsmm vpx has been shown, via co-immunoprecipitation, to interact with APOBEC3A[198]. Association of HIV-2 vpx with APOBEC3A results in decreased stability and SIVsmm replication has been shown to lead to decreased APOBEC3A expression levels in monocytes. The exact mechanisms behind these observations and the contributions that APOBEC3A makes to the restriction of HIV-1 infection in myeloid cells have been hard to elucidate, not least because of the strong block on myeloid cell infection from SAMHD1. Vpx has also been

shown to interact with interferon regulatory factor 5 (IRF5)[59]. IRFs are transcription factors that are implicated in immune responses, cell growth regulation and haematopoietic differentiation. Vpx interacts with IRF5 and inhibits transactivation through decreased promoter binding. This leads to a reduction in the expression of IRF5-induced genes such as IL6, TNF$\alpha$ and IL12p40. Although only recently described, the interaction between vpx and IRF5 could explain the effects of *vpx* on the innate immune system.

One of the most striking questions surrounding *vpx* is the absence of the gene in the majority of SIVs, most notably in the SIVcpz/HIV-1 lineage. The ability to antagonise SAMHD1 is clearly not necessary for HIV-1 to establish a successful infection. The differences in pathogenicity and epidemiology between HIV-1 and HIV-2 may suggest that the lack of SAMHD1 antagonism and infection of dendritic cells (DCs) is beneficial, leading to an avoidance of DC activation and therefore the prevention of some innate immune responses[199,200]. It also remains possible that SAMHD1 antagonism is simply not needed due to the high affinity of HIV-1 RT for dNTPs when compared to other retroviruses, such as Murine leukaemia virus (MLV)[201]. However, HIV-1 is unable to infect non-dividing cells *in vitro* so there is still a block caused by SAMHD1, and until the affinity of HIV-2 RT for dNTPs is known this remains an unanswerable question.

*Vpx* is located in the HIV-2 genome between *pol* and *env* and overlaps with two other accessory genes, *vpr* and *vif*. HIV-1 contains a *vpu* gene in this position but retains *vpr* and *vif*. *Vpx* and *vpr* are paralogues; leading to the hypothesis that *vpx* is the result of a duplication event of *vpr*. Whether or not this duplication event occurred after SIV speciation, leading to the presence of *vpx* in certain SIV lineages, or whether *vpx* was present in the SIV ancestor and subsequently lost in SIVcpz/HIV-1 remains unclear. The detection of positive selection signals in Old World Monkey (OWMs) SAMHD1 but not in human, gorilla or chimpanzee SAMHD1 may suggest that *vpx* is the driver of this evolutionary pressure and the loss of *vpx* from SIVcpz/HIV-1 explains the lack of signals of selection[202,203,204]. Although the date of HIV-1 introduction can be accurately estimated to the start of

the 20th century, the age of the SIVs is less clear. The lack of pathogenicity of SIVs in their natural hosts suggests an ancient origin. Four species of the *Chlorocebus* or African green monkey (AGM) are infected with distinct strains of SIVagm and the similarity of the topologies between AGM and SIVagm phylogenies may be evidence of co-speciation between host and virus, also implying an ancient origin for the SIVs[205]. This assumption has been challenged by the finding that the mitochondrial AGM and SIVagm phylogenies have incongruous topologies, supporting the hypothesis of 'preferential host-switching' at the borders of AGM habitats as the model of SIVagm diversification, rather than host and virus co-speciation[206]. The SIVs may be much younger than previously assumed and therefore it is unlikely that *vpx* is the driver of SAMHD1 evolution.

The ability of *vpx* to antagonise host SAMHD1 is conserved amongst different SIVs[207]. It has also been shown in certain SIVs that *vpr* is able to antagonise SAMHD1. Overlaying SAMHD1 degradation onto a phylogeny of SIV/HIV *vpx* and *vpr* sequences showed monophyletic clustering of *vpx* and *vpr* genes that are able to degrade SAMHD1[202]. The most parsimonious explanation for this observation is that *vpr* gained the ability to degrade SAMHD1 (in addition to the conserved function of cell-cycle arrest) sometime after the initial radiation of SIVs and the neofunctionalisation of *vpr* led to a selection pressure which favoured *vpr* duplication and evolution of the specialised SAMHD1 antagonist *vpx*. An exception to this is the presence of *vpx* in SIVrcm, which is one of the ancestral strains of the recombinant SIVcpz[13]. If *vpx* was acquired only once during SIV evolution and never lost in SIVcpz/HIV-1 then SIVrcm must have acquired *vpx* through recombination with SIVsmm. The absence of recombination marks around *vpx* in SIVrcm suggests *vpx* was present in the ancestor of SIVsmm and SIVrcm rather than acquired later through recombination. *Vpx* appears to have a slower evolutionary rate than *vpr* and this may be due to the overlap with the apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like 3G (APOBEC3G) antagonist *vif*[49]. The consequences of a putative deletion of *vpx* in SIVcpz are more readily explained when looking at the functionality of *vif* as an

APOBEC3G antagonist[208]. APOBEC3G is a potent host restriction factor and is targeted for degradation by *vif*[209]. For most SIVs, APOBEC3G antagonism is host-specific. A notable exception to this is SIVsmm *vif* which is able to antagonise human APOBEC3G[210]. SIVrcm *vif* is only moderately active against chimpanzee APOBEC3G and completely inactive against human APOBEC3G. APOBEC3G antagonism is likely to be one of the major barriers in cross-species transmission of SIVs. Deletion of *vpx* from SIVcpz leads to the loss of the stop codon from *vif*, rendering the protein inactive. Therefore there is a strong selection pressure that leads to a phenomenon known as 'overprinting'. Overprinting is any genome modification that leads to the creation of a second ORF[211]. In the case of SIVcpz, there is a second stop codon that is out of frame in *vpr*. Any insertion that causes a frame shift and brings this stop codon into frame is favourable. In SIVcpz overprinting led to a 60-75bp insertion. The origins of this sequence are unknown but it contains a SIVcpz/HIV-1 unique 'cullin box' motif that is critical for APOBEC3G antagonism[212]. The lack of human APOBEC3G antagonism by SIVrcm seems to be a larger barrier to zoonosis than SAMHD1 activity and so *vpx* deletion, overprinting of *vif* and the creation of the cullin-box motif may have been critical adaptations of SIVrcm in chimpanzees that allowed the virus to cross successfully into humans.

Elucidation of the crystal structure of SIVsmm vpx (as a ternary complex with $CRL4^{DCAF1}$ and SAMHD1) highlights residues involved in interactions[185]. Many of these residues have been mutated *in vitro* and the positions and phenotypes of these residues are shown (**Table 1.2**).

Table 1.2: Sites of interaction between vpx and DCAF1/SAMHD1.

| Position | HIV-2 | Region | Interaction | Mutant | Phenotype | Reference |
|---|---|---|---|---|---|---|
| E6 | E | Nt | DCAF1 | | | |
| I8 | V | Nt | DCAF1 | | | |
| P9 | P | Nt | DCAF1 | | | |
| N12 | N | Nt | DCAF1 | N12A | Partial SHD1 degradation | [213,183] |
| | | | | | No HIV-1 infection in MDMs | |
| S13 | S | Nt | DCAF1 | S13A | SAMHD1 degradation | [214,213,215] |
| | | | | | Rescue of HIV-1 in MDMs | |
| W24 | W | α1 | DCAF1 | W24A | No infection rescue | [216] |
| T28 | T | α1 | DCAF1 | T28A | No rescue of SIVmac infection | [214] |
| | | | | | Partial rescue of HIV-1 in MDDCs | |
| I32 | I | α1 | DCAF1 | I32S | No infection rescue | [216] |
| | | | | | No interaction with DCAF1 | |
| | | | | | No SAMHD1 degradation | |
| A35 | A | α1 | DCAF1 | | | |
| L44 | L | α2 | DCAF1 | | | |
| L48 | V | α2 | DCAF1 | | | |
| R51 | R | α2 | DCAF1 | | | |
| W56 | W | α2 | DCAF1 | | | |
| Y66 | Y | α3 | DCAF1 | Y66A | Y66, 69,71A | [214,217,215] |
| Y69 | Y | α3 | DCAF1 | Y69A | No infection rescue | [214,217,215] |
| | | | | | No interaction with DCAF1 | |
| | | | | | No SAMHD1 degradation | |
| R70 | R | α3 | DCAF1 | | | |
| L74 | L | α3 | DCAF1 | | | |
| Q76 | Q | α3 | DCAF1 | Q76A/R | No infection rescue | [218,82,215] |
| | | | | | No interaction with DCAF1 | [219,57,220] |
| | | | | | No SAMHD1 degradation | [216] |
| K77 | K | α3 | DCAF1 | K77A | No infection rescue | [219] |
| | | | | | No interaction with DCAF1 | |
| A78 | A | α3 | DCAF1 | | | |
| F80 | F | α3 | DCAF1 | F80A | No infection rescue | [218,57,220] |
| | | | | | No interaction with DCAF1 | |
| | | | | | No SAMHD1 degradation | |
| M81 | I | α3 | DCAF1 | | | |
| K84 | K | α3 | DCAF1 | | | |
| K85 | R | α3 | DCAF1 | | | |
| G86 | G | α3 | DCAF1 | GC86, 87A | No rescue of SIVmac in MDCCs | [214] |
| | | | | | Partial HIV-1 infection rescue | |

| | | | | | | |
|---|---|---|---|---|---|---|
| G14 | G | Nt | SAMHD1 | | | |
| E15 | E | Nt | SAMHD1 | E15A | No rescue of HIV-1 in MDMs<br>Partial SAMHD1 degradation | [213,183] |
| E16 | E | Nt | SAMHD1 | E16A | No rescue of HIV-1 in MDMs<br>Partial SAMHD1 degradation | [213,183] |
| I18 | I | α1 | SAMHD1 | | | |
| G19 | G | α1 | SAMHD1 | | | |
| A21 | A | α1 | SAMHD1 | | | |
| F22 | F | α1 | SAMHD1 | | | |
| W24 | W | α1 | SAMHD1 | W24A | No infection rescue | [216] |
| L25 | L | α1 | SAMHD1 | | | |
| M62 | M | α2-α3 | SAMHD1 | | | |
| S63 | S | α2-α3 | SAMHD1 | S63, 65A | Infection rescue in MDDCs with SIVmac<br>Infection rescue in MDDCs with HIV-1 | [214] |
| Y66 | Y | α3 | SAMHD1 | Y66A | Y66, 69,71A | [214,217,215] |
| Y69 | Y | α3 | SAMHD1 | Y69A | No infection rescue<br>No interaction with DCAF1<br>No SAMHD1 degradation | [214,217,215] |

Position shows consensus amino acid in SIVsmm *vpx* and location of the site of interaction. HIV-2 consensus residues at each position are also shown and differences between HIV-2 and SIVsmm are highlighted in red. Mutations shown to have an effect on infectivity *in vitro* are listed alongside phenotype of the mutation.

The C-terminus of vpx contains a highly conserved poly-proline motif (PPM) consisting of seven consecutive prolines (positions 103-109). The PPM has been shown to be critical for efficient translation of *vpx*[221]. Mutations of the PPM *in vitro* have shown that mutation of a single proline to alanine has no effect apart from at position 106, where a moderate reduction was seen in vpx protein[222]. There was no change in mRNA levels and addition of the protease inhibitor MG-132 did not affect this observation, suggesting the reduction is at the translation level. Mutation of multiple prolines to alanine has a more significant effect and mutations of P103-106 and p106-109 to alanine results in a minimal expression of vpx.

A study of *vpx* sequences from viral isolates derived from progressors and LTNPs showed there was no correlation between the ability of *vpx* to degrade

SAMHD1 and clinical outcome[223]. This study only identified one mutation, K68M, in a viraemic patient, which showed reduced SAMHD1 antagonism *in vitro*. A neighbouring mutation, Y69F, was also identified but this showed no effect on SAMHD1 degradation and for all alleles tested, promotion of SAMHD1 degradation also promoted macrophage infection. An artificial mutation in HIV-2 ROD (E15G) impaired the ability to degrade SAMHD1. This is in line with previous observations that have shown HIV-2 ROD vpx is only active against human SAMHD1, whereas vpx from the HIV-2 isolate 7312a is active against human and OWM SAMHD1[202].

### 1.10.2 HIV-2 capsid poly-proline motif

A study of capsid diversity in HIV-2 singly infected and ART-naïve patients from Caiò, showed that three positions in the p26 sequence (positions 119, 159 and 178) showed considerable variation when compared to the HIV-2 ROD reference strain[224]. In ROD, a proline residue occupies all 3 positions whereas in the sequences generated from patients, 35-70% had an alternative residue at these positions. A clear inverse correlation was seen between average plasma VL and the number of prolines at these three positions and a second positive correlation was observed between the number of proline residues and the susceptibility to TRIM5$\alpha$ restriction. Phylogenetic analyses of the patient-derived *env* sequences failed to show differential clustering of PPP and non-PPP viruses, suggesting that there may have been multiple selections for and multiple occurrence of the PPP motif in Caiò, rather than a single PPP founder virus. Clusters containing exclusively PPP viruses were present, suggesting that the PPP motif might lead to a more transmissible virus; however the exact method by which this increased transmissibility might be achieved remains unclear and the presence of PPP clusters due to a founder effect has not been formally excluded. These results show the presence or absence of proline at positions 119, 159 and 178 in p26 has a significant impact on TRIM5$\alpha$ restriction and viral replication but is not able to explain fully the differences seen in HIV-2 disease progression between LTNP and

progressors, as viruses containing PPP and non-PPP motifs cluster together in a phylogeny created from non p26 capsid sequences.

### 1.10.3 Viral diversity and evolution

The genetic heterogeneity and rapid evolution of HIV-1 are fundamental barriers to the development of an effective vaccine[225]. The rapid rate of HIV-1 evolution results from a quick life cycle, an error-prone reverse transcriptase and recombination during reverse transcription that generates diverse progeny viruses[226],[227]. Factors that exert selective pressure on HIV-1, leading to directed divergence, include ART and host immune responses. Following infection with a single TF virus, the viral quasi-species quickly diversifies within a patient and high diversity is often a feature of AIDS[228],[229].

In contrast, knowledge of the diversity and evolution of HIV-2 following infection is limited. This is due in part to the limited number of HIV-2 sequences that are available. As of 2014 there were 29 full-length HIV-2 genomes available in addition to multiple partial sequences (**Figure 1.11**).

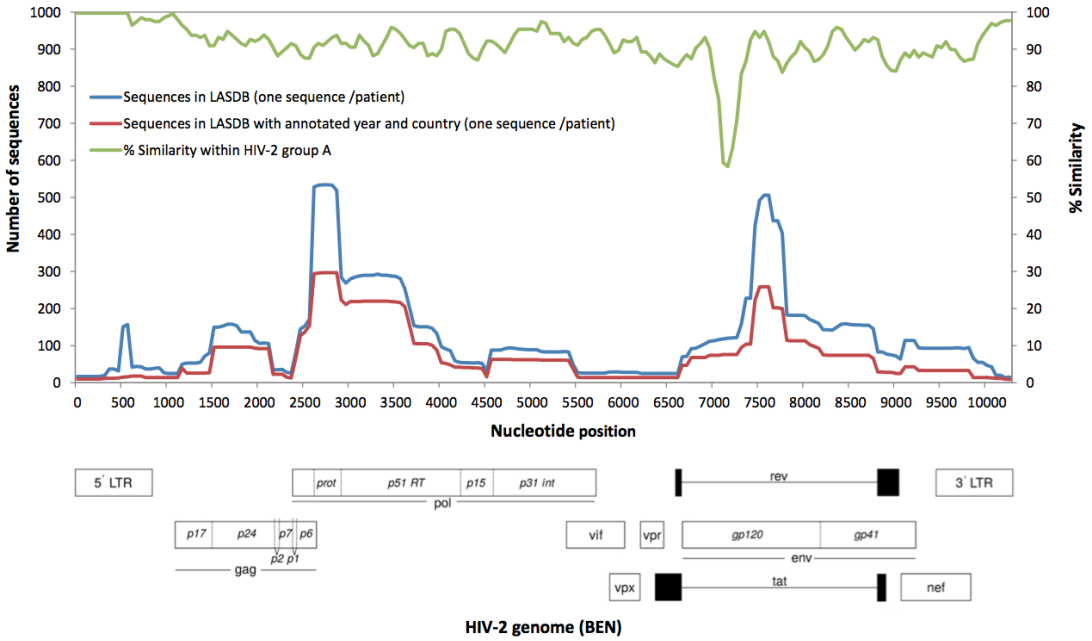



Figure 1.11: Current HIV-2 sequence coverage.
Plot shows the number of sequences in the Los Alamos National Laboratory (LANL) HIV database relative to genomic position (reproduced with kind permission from J. Esbjörnsson)

As can be seen from the LANL database plot, the majority of sequences have been generated from the structural genes *gag, env* and *pol* with much lower sequencing coverage of the accessory genes. Previous studies of HIV-2 evolution and diversity have been focussed on single genes. A longitudinal analysis of the C2-C3 region of *env* from 4 individuals showed a low non-synonymous to synonymous (dN/dS) ratio in 3 of the samples, indicative of negative selection pressures acting on this region[230,231]. Another study showed a higher evolutionary rate in the gp125 and V3 regions of *env* for HIV-2 when compared with disease-stage matched HIV-1 patients[76]. This result was most evident when only patients with advanced disease were included in the analysis and was caused by a higher rate of synonymous mutations in the advanced patients when compared to LTNPs. The authors postulate that this could be due to a shorter generation time of higher error rate for HIV-2 RT. Although the error rate of HIV-2 RT is unknown, a study involving HIV-1 and HIV-2 matched patients showed that HIV-2 probably has a lower replication rate[232]. Replication rate was estimated by comparing the levels of *gag* and *tat* mRNA. *Gag* is an unspliced mRNA that is constitutively expressed in all host cells containing proviral DNA whereas *tat* mRNA is only seen in recently infected or virus producing cells. HIV-2 infected patients showed similar levels of *gag* mRNA as would be expected but significantly reduced levels of *tat* mRNA when compared to matched HIV-1 patients, implying a lower rate of replication for HIV-2. It is not clear which factors contribute to a reduced replication rate in HIV-2 but it could be due to lower levels of transcriptional activation or reduced HIV-2 RNA stability. The presence of *tat* mRNA in all patients, even in those with undetectable plasma VL showed that there is on-going HIV-2 replication, even in the absence of measurable viraemia. The emergence of ART-resistance mutations in HIV-2 patients also implies that there is on-going viral replication after successful ART has lowered VL below the limit of detection[233]. In HIV-D patients, HIV-1 genetic diversity is positively correlated with VL and is significantly reduced at the same time point in infection when HIV-D and HIV-1 monoinfected individuals are compared[122]. This observation suggests a link between HIV-1

genetic diversity and disease progression in the context of HIV-D infection but the reasons behind this correlation and the contribution of HIV-D infection to HIV-1 and HIV-2 evolution and HIV-2 genetic diversity have yet to be fully explained.

## 1.11 Host genetic factors involved in HIV-2 disease progression

Interaction between pathogens and their hosts is one of the major drivers in evolution. Both pathogen and host are constantly attempting to thwart the other, leading to a fitness gain for the successful party and therefore a strong selective pressure on mutations that confer resistance to harmful pathogens or allow evasion of host defences. This theory is normally referred to as the 'Red Queen Hypothesis' or an evolutionary arms race, in which adaptations that favour one party are compensated by adaptations in the other[234].

*"Now, here, you see, it takes all the running you can do, to keep in the same place."*
Lewis Carroll, "Through the Looking-Glass and What Alice Found There" (1960)

Although HIV is a recent human pathogen on an evolutionary scale, humans are not without defences that block or slow the spread of retroviral infection and host immune genes and restriction factors are the major source of selection pressure on the HIV genome.

### 1.11.1 Human leucocyte antigens

Human Leucocyte Antigen (HLA) class I and class II genes are located in the human major histocompatibility complex (MHC) on the short arm of chromosome 6[235]. The MHC contains the most polymorphic genes in the human genome and HLA molecules are one of the fundamental determinants of adaptive immune responses. HLA class I contains 3 loci; HLA –A, -B and –C. HLA class I molecules act by binding antigenic epitopes derived from intracellular pathogens, presenting them on the surface of infected cells and initiating a cytotoxic T cell response (commonly referred to as a CTL response)[236]. CD8+ T cells are able to kill infected

cells through recognition of viral peptides presented on the cell surface through restriction by HLA class I molecules.

Differential HLA class I alleles have been implicated in control of HIV-1 and have been shown to be a major determinant in viral set point following the acute stage of HIV-1 infection[237]. HIV-1 evolution is driven to a large extent by the frequencies of HLA class I alleles in host populations[238]. A genome wide association study (GWAS) analysing 1,384,048 single-nucleotide polymorphisms (SNPs) from 974 HIV-1 controllers and 2648 HIV-1 progressors identified >300 SNPs that reached genome-wide significance in the MHC and none that reached significance elsewhere in the genome[146]. HLA alleles have been associated with both good and poor prognosis following HIV-1 infection[239]. HLA alleles HLA-B*B5701, HLA-B*5703, HLA-B*5801, HLA-B*27 and HLA-B*51 are associated with lower viral set points and slower progression to AIDS. This is likely to be due to the fact that these alleles target and bind to epitopes located in conserved regions of the viral protein gag. Single mutations occurring in conserved regions of the structural protein gag that reduce or abrogate HLA binding or hinder recognition by the T-Cell receptor (known as 'escape mutations') result in a fitness cost and are therefore not favoured[240]. Additionally, gag is the earliest and most abundant viral protein, and presentation of gag epitopes can occur before viral integration and replication[241]. Conversely, some subtypes of the HLA-B*35 allele are associated with poor control of viral replication and rapid progression to AIDS, in a co-dominant manner, where homozygotes progress more rapidly than heterozygotes[242].

Associations between *HLA* class I alleles and HIV-2 disease progression have also been shown. An early study using the presence of HIV-2 gag p26-specific antibody as a proxy for HIV-2 disease progression showed an association between HLA-B*35 and rapid disease progression, an observation which is similar to the effect of HLA-B*35 in HIV-1 infection[243]. A larger study on the Caió community cohort showed an association between the common HLA-B*1503 allele, higher HIV-2 viral loads and lower CD4+ T-Cell counts, suggesting this allele might be associated with lack of viral control[244]. Interestingly, the other common HLA-B*15

allotype HLA-B*1510, showed no association with HIV-2 disease progression. Additionally, no association was seen with HLA alleles associated with HIV-1 progression, including HLA-B*5703 and HLA-B*35.

### 1.11.2 Killer-cell immunoglobulin-like receptors

Natural Killer (NK) cells are involved in the innate immune response and are able to recognise and kill aberrant target cells without prior exposure to the pathogen[245]. Therefore, NK cells are likely to play an important role in the control of acute HIV infection prior to the CTL response, which lowers peak viraemia to the viral set-point. NK cell activity broadly encompasses two factors, tolerance to self and recognition of non-self infected target cells. During NK cell maturation (or licensing), activity is tuned by a large number of activating and inhibitor receptors. Inhibitory receptors allow NK cells to recognise normal HLA expression levels and activating receptors educate NK cells to recognise cells expressing an abnormal level of HLA[246]. The most widely studied of these are the killer-cell immunoglobulin-like receptors (KIR) genes. KIR genes are arranged in a cluster on chromosome 19 and like the MHC, show a high level of polymorphism[247]. NK cell recognition of HLA-B is modulated by the inhibitory KIR3DL1. The KIR3D locus can contain the inhibitory KIR3DL1 and/or the activating KIR3DS1[248]. These two alleles are highly homologous despite their divergent functions and share approximately 97% amino acid sequence identity in the extracellular domain. Several studies have shown associations between compound HLA-B/KIR3D genotypes and HIV control[249,250]. The KIR3DS1/HLA-Bw4-80I genotype is associated with lower viral set-point and protection against opportunistic infection in HIV-1. This is probably due to an epistatic interaction between KIR3DS1 and the Bw4-80I ligands on HIV infected cells[251,252]. A similar effect was shown for the inhibitory KIR3DL1 where a compound genotype with HLA-Bw4-80I was also protective against HIV-1 disease progression[253]. Analysis of KIR and HLA genotypes in Caió showed a much higher frequency of KIR3DS1 when compared with neighbouring populations. However, no association was seen between HLA/KIR compound genotypes and HIV-2 disease progression. There was an

over-representation of individuals with KIR2DS2 or KIR2DL2 and the corresponding HLA-C ligand in the uninfected individuals when compared with a matched HIV-2 infected group, but this did not reach statistical significance[244].

The identification of HLA/KIR compound genotypes involved in the control of HIV-1 suggests a role for HLA-B alleles in both the innate and adaptive immune responses against HIV. The associations with control for both KIR3DS1 and KIR3DL1 allotypes implies that efficient action of inhibitory and activating KIRs are beneficial for the host, allowing fine-tuning of NK recognition of HIV infected cells. The HIV accessory protein nef down-regulates cell-surface expression of HLA-A and HLA-B alleles, providing a mechanism whereby appropriately educated NK cells would be able to efficiently recognise HIV infected cells[61]. Whilst HLA/KIR compound genotypes have been shown to be associated with differing outcome following HIV-1 infection, the role of KIR in HIV-2 disease progression remains unclear, partly due to the small number of HIV-2 infected individuals available. The extent to which HIV-2 nef down regulates HLA A and B expression on the surface of infected cells is also not known.

### 1.11.3 Restriction factors

Restriction factors are a set of genes present in the host that are able to decrease significantly the infectivity of retroviruses. Differing from dependency factors, which are host cellular factors hijacked by the virus, true restriction factors are specialised species-specific defence proteins that act to block distinct stages in the retroviral life-cycle[254]. Four key factors are normally employed in the identification of HIV restriction factors. These are: (1) a significant and direct *in vitro* decrease in HIV infectivity in the presence of the restriction factor, (2) the presence of a corresponding antagonist in HIV which has evolved to counteract the actions of the host protein, (3) strong evidence of positive selection of the restriction factor when compared to close relatives, indicative of a fitness advantage driven by historical pathogen challenges, (4) a close link to the innate immune system. When these factors are considered, HIV-2 has four major restriction factors. These are: APOBEC3G that targets reverse transcription of the viral genome, BST2

(commonly referred to as Tetherin) which prevents immature virion budding, SAMHD1 which blocks reverse transcription and TRIM5$\alpha$ which leads to premature dissociation of the viral capsid. Restriction factors are the result of an evolutionary arms race, not with a single pathogen, but with changing evolutionary pressures from multiple pathogens; and many of the restriction factors that are involved in the block of HIV-2 infection are also implicated in HIV-1 restriction[255].

*APOBEC3G*

APOBEC3G is a member of the human apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like 3 (APOBEC3) family of genes and is a potent antiviral restriction factor[256]. APOBEC3G is a cytidine deaminase that is packaged into immature virions, resulting in a mis-reading of deoxyuridine as thymidine during reverse transcription upon infection of a target cell, leading to a guanine to alanine substitution. APOBEC3G-induced hyper-mutation causes a decrease in viral complementary DNA (cDNA) stability and induces lethal fragmentation of viral cDNA by the cellular DNA repair machinery[257]. APOBEC3G-induced mutations can also lead to the production of defective viral transcripts producing non-functional proteins.

Both HIV-1 and HIV-2 counteract APOBEC3G via interaction with the accessory protein vif[258],[259]. Vif prevents APOBEC3G incorporation into virions, possibly through the formation of a vif-APOBEC3G complex and also targets APOBEC3G for proteasomal degradation[260]. Although vif is present in both HIV-1 and HIV-2, the level of sequence similarity at the amino acid level is only 30%[261]. Recent work on the regions of APOBEC3G targeted by HIV-2 vif has shown that HIV-1 vif and HIV-2 vif interact with distinct and separate domains in APOBEC3G[55]. Whilst HIV-1 vif targets a short motif (FWDPDY) at position 128-130 in APOBEC3G, HIV-2 vif targets multiple, non-neighbouring residues found between positions 163-321. This is in line with a previous study which suggested that HIV-1 vif originated from the overprinting of vpr in SIVcpz, generating the HIV-1/SIVcpz specific 'cullin-box' motif which is required for APOBEC3G antagonism[208].

*BST2 (Tetherin)*

Tetherin is a type II transmembrane protein comprising a cytoplasmic N-terminal domain, a single transmembrane domain, an extracellular coiled-coil domain and a C-terminal anchor[262]. Nascent virions are tethered to the plasma membrane through interaction between envelope proteins and tetherin, preventing virion release and infection of new target cells[263]. Tetherin associated virions are subsequently internalised and undergo degradation in the lysosomes. All HIV/SIVs counteract tetherin, however, the viral protein used in the antagonism varies between different viruses.

Vpu is an HIV-1 specific membrane-spanning accessory protein involved in both the sequestration of tetherin into a perinuclear compartment and targeting of tetherin for proteasomal degradation[264]. Following an initial interaction between the antiparallel trans-membrane helices of vpu and tetherin, the relative importance of sequestration versus degradation remains unclear. However, both pathways depend on dynamin-2, a GTPase involved in the breaking of vesicle membranes[265]. The two mechanisms of vpu are not mutually exclusive and therefore it seems likely that vpu evolved this dual functionality to ensure efficient removal of tetherin from sites of viral release.

In the absence of vpu, HIV-2 uses the viral envelope glycoprotein gp41 to counteract tetherin[266]. The exact residues that contribute to gp41-tetherin interactions are not clearly defined but interaction appears to involve the extracellular domains of both gp41 and tetherin. Another critical factor is the conserved tyrosine-based endocytosis motif in the cytoplasmic tail of gp41. Unlike vpu, gp41 is unable to target tetherin for degradation and instead relies on the single function of clathrin-dependent internalisation to remove tetherin from sites of viral release[267].

The majority of SIVs, including SIVcpz and SIVsmm, use nef to counteract tetherin in their natural hosts[268]. Like HIV-2 env, nef-mediated tetherin antagonism involves a single cellular pathway, resulting in the removal of tetherin from the plasma membrane via AP-2 dependent endocytosis[269]. The divergent mechanisms

for tetherin antagonism employed by SIVs and the two HIVs are likely to be due to the fact that human tetherin lacks a 5 amino acid motif (G/D, DIWK) in the N-terminal cytoplasmic domain necessary for susceptibility to nef[270]. The potency of tetherin restriction on viral fitness is clearly demonstrated by the recent emergence of env and vpu mediated anti-tetherin strategies, which appear to have been gained by HIV-1 and HIV-2 following zoonotic transfer of SIVs, under strong selection pressure from nef-resistant tetherin[271].

*SAMHD1*

The interplay between SAMHD1 and *vpx* is covered in detail earlier in this chapter. Following the recent identification of SAMHD1 as the target for *vpx*, the evolutionary relationship between orthologous SAMHD1 genes through different primate species was investigated. SAMHD1 seems to have been subject to periods of strong positive selection pressure throughout the diversification of the primates, leading to species specificity and signals of positive selection in both the amino-terminal and carboxyl-terminal domains[203],[202]. This suggests that SAMHD1 is a true restriction factor, even in the absence of an apparent antagonist in HIV-1.

*TRIM5α*

TRIM5α is a restriction factor that is expressed by all primate species and consists of a Really Interesting New Gene (RING) domain, a B-Box 2 and a coiled-coil domain[272]. TRIM5α acts early in the viral life-cycle through interaction with the incoming viral capsid, resulting in destabilisation of the capsid (CA) structure, abortive reverse transcription and initiation of a signalling cascade promoting innate immune responses[273]. In contrast to the other restriction factors implicated in HIV-1 infection, human TRIM5α (hTRIM5α) contains a single amino acid substitution in the B30.2/SPRY domain that renders it almost completely inactive against HIV-1[274]. In contrast, CA proteins derived from multiple HIV-2 plasma samples have been shown to be highly susceptible to hTRIM5α[275]. Susceptibility did not vary between CA proteins derived from patients with differing clinical outcomes, suggesting that increased hTRIM5α susceptibility is a general feature of HIV-2 infection. Additionally, susceptibility was similar between HIV-2 groups A

and B and SIVsmm and therefore it seems likely that hTRIM5α restriction was not a major barrier to the cross-species transmission of SIVsmm. The susceptibilities of HIV-1 and HIV-2 CA to hTRIM5α may have a role in the differing pathogenicity of these two viruses; however, the lack of a TRIM5α antagonist in either HIV-1 or HIV-2 suggests that the selection pressure exerted by hTRIM5α is unlikely to be as potent as for other restriction factors.

## 1.12 Evolution of sequencing technologies

In contrast to current sequencing methods that focus on DNA sequencing, the first nucleic acid to be sequenced was a 77bp alanine tRNA taken from the budding yeast *Saccharomyces cerevisiae*, which was accomplished through resolution of fragments generated by digestion with multiple specific nucleases[276]. Whilst this method allowed sequencing of both tRNA and rRNA in quick succession, it required a very pure RNA input and was not broadly applicable to all nucleic acids, including DNA. In order to sequence DNA, an alternative method was sought and in 1975, a novel technique known as the 'Plus and Minus' technique was used to sequence the 5,386bp genome of the bacteriophage X174, generating the first whole genome sequence[277]. Shortly afterwards, in 1977, a more advanced method of sequencing was developed, known as the "dideoxy" chain termination or "Sanger" (named after Fredrick Sanger, who invented the technique) method of sequencing[278]. Chain termination sequencing involves the incorporation of fluorescently labelled chain terminating ddNTPs during DNA elongation by polymerase. The presence of both chain terminating and standard dNTPs in the reaction ensures that termination occurs randomly along the length of the DNA molecule, producing products of varying lengths, all ending with a fluorescently labelled ddNTP. Resolution of the products according to size allows the DNA sequence to be read from the majority nucleotide at each position, resulting in contiguous DNA sequences of up to 800bp. For 30 years, Sanger sequencing remained the most popular and widely used form of sequencing and was used during the Human Genome Project and to generate the first whole genome sequences for HIV-1 and HIV-2[279].

Many advances in our knowledge of HIV have been the direct result of the ability to sequence the genomes of HIV-1 and HIV-2. Refinement of library preparation methods, including the advent of molecular cloning into E. coli allowed an understanding of the vast genetic diversity of HIV-1, one of the main reasons behind the failure of multiple preventative vaccine trials[280]. Recently, advanced techniques such as single genome amplification (or limiting dilution sequencing) have allowed sequences to be generated from individual viruses, rather than from the population as a whole. Phylogenetic reconstruction of individual viral sequences from a single patient has allowed an understanding of the earliest events in HIV-1 transmission, namely that the majority of HIV-1 infections are seeded by a single transmitted founder (TF) virus[228]. The properties of these TF viruses are of great interest and still remain to be completely defined, however, the narrow bottleneck through which HIV must pass provides an explanation for the differing dominant HIV-1 subtypes in different countries and also provides new avenues for targeted interventions to prevent transmission.

Following the dominance of Sanger sequencing, the early part of the 21$^{st}$ century saw a revolution in sequencing technologies. These technologies are collectively known as 'next generation sequencing technologies'. Although there are many platforms available for next generation sequencing, they all work on the principle of repeat sequencing of the same DNA molecule[281]. In contrast to Sanger sequencing, the block on elongation imposed by the incorporation of a labelled ddNTP is reversible, allowing the DNA sequence to be read once during each round, followed by the reversion of the block and another sequencing cycle. This novel chemistry has led to a significant decrease in the cost of sequencing a single nucleotide, allowing large-scale human genome sequencing projects (such as the 1000 Genome Project)[282]. Although Sanger sequencing is still commonly employed for the study of HIV, many studies on HIV have been conducted using next generation sequencing technologies including the development of a pan HIV-1 NGS strategy and a sequence-independent method of DNA amplification[283],[284]. In addition, shotgun RNA sequencing has recently been proposed as a target

enrichment free method for sequencing RNA viruses, and has been demonstrated for HCV and norovirus[285]. The development of next generation protocols for HIV sequencing look set to even further increase our ability to resolve the subtle evolutionary dynamics of HIV, the understanding of which is a critical step in the development of a globally effective HIV vaccine[286].

## 1.13 Summary

The natural history of HIV-2 infection poses an intriguing question: How does a virus that shares a similar origin, genome and life cycle with HIV-1 cause AIDS in some patients whilst others remain aviraemic and asymptomatic in the absence of treatment? The mechanisms underlying the differing pathogenicities between HIV-1 and HIV-2 remain unclear, however, it is apparent that lessons learned from natural control of HIV-2 in LTNPs would be extremely beneficial in the field of HIV vaccine design and functional cure. Unlike HIV-1, which has a global distribution and increasing number of infections, HIV-2 has remained largely confined to West Africa and prevalence is now dropping at such a rate that extinction is predicted before the end of the century. Much of the current focus of HIV-1 research is directed towards finding a cure, either a sterilising cure (resulting in total elimination of HIV-1 from all compartments of the body and eradication of the proviral reservoir) or a functional cure (resulting in control of viral replication and maintenance of CD4+ T-cell counts in the absence of treatment)[277]. The recent re-bound of viraemia in the 'Mississippi baby', a child who had previously controlled viral replication following the cessation ART initiated during acute infection, shows how challenging either strategy is likely to be[287],[288]. Therefore, the knowledge gained from the factors underlying differential HIV-2 progression (particularly from the high proportion of LTNPs, whose clinical and virological status has analogies with a 'functional' cure) should not be overlooked and it is important to gain a clear picture of the genetics of HIV-2 infection before the virus disappears.

The data presented in this thesis were generated to address two gaps in the current knowledge of HIV-2. The data presented in Chapter 3 was gathered to

assess the diversity, evolution and selection acting on the HIV-2/SIVsmm accessory gene *vpx* and to look for alleles in *vpx* that may be associated with control of HIV-2. Sequences were generated from primary patient samples for the first time, allowing an accurate estimation of the nature of *vpx* in natural infection rather than following *in vitro* viral isolation and propagation. The data presented in Chapters 4 and 5 describe a novel application of shotgun RNA sequencing (RNA-Seq), allowing whole genome sequencing of HIV-2 without the need for prior target amplification. This approach is especially beneficial in the context of HIV-2, where the total sequence knowledge is much reduced when compared to HIV-1. Chapter 4 assesses the divergence of the HIV-2 isolate HIV-2 ROD from the published reference sequence following *in vitro* expansion and multiple rounds of replication. Many of the previous studies of HIV-2 have involved viruses isolated from patient plasma and therefore, divergence *in vitro* is an important factor that may have implications for the interpretation of results from these studies. Chapter 5 presents whole genome HIV-2 sequences, generated from primary patient samples without the need for prior target enrichment, reducing many of the biases commonly involved with the sequencing of RNA viruses. The power of this technique is only beginning to be appreciated and Chapter 5 demonstrates the biases involved, the technical challenges and ultimately allows the presentation of a low-bias picture of the genetic diversity over the whole genome of HIV-2 for the first time.

# Chapter 2: Materials and Methods

## 2.1 Caió Community Cohort

Study participants used in this project were recruited from the Caiò Community Cohort in rural Guinea Bissau. Caió is a small village with a population of approximately 10,000, which is located on the coast near the islands of Jeta and Pecixe (**Figure 2.1**).


Figure 2.1: Map of Guinea Bissau.
Location of Caió is highlighted with an orange circle.

Caió is a geographically isolated community and subsistence farming is the main source of employment, with cashew nuts forming the major crop. Additionally members of the community migrate to local towns for education and work, as well as frequently travelling to European countries, particularly to Portugal and France[289]. More than 95% of the inhabitants of Caió belong to the Manjako ethnic group, a genetically isolated tribe who differ from neighbouring tribes through their animistic belief system[244].

The Caió community HIV-2 cohort was initiated in 1989 as a joint project between the Medical Research Council (MRC) laboratories, The Gambia and the Bandim Health Project based in Bissau. An initial sero-survey was carried out following the observation of a high number of HIV-2 infected commercial sex workers (CSWs) from Caió attending a genitourinary medicine (GUM) clinic in Senegal[290]. There have been three sero-surveys of the general adult population, conducted in 1989-1991, 1996-1997 and 2006-2007[107]. A case-control cohort taken from the total adult population was established in 1991 and consists of age and sex matched HIV-2 cases and negative controls. In addition to the major sero-surveys, studies were carried out on the case-control cohort in 1996, 2003, 2006, 2008 and 2010, during which viral load and CD4+ T-cell subset analyses were performed on all available members[87]. All HIV infected patients in the cohort have access to free medical care and ART has been available in Caió since 2007.

The Caió community cohort represents a rare community based HIV-2 cohort and has allowed HIV-2 disease progression to be longitudinally followed, sometimes for decades, in the same individuals[291]. Community cohorts avoid clinic bias, where there is an over-representation of patients with advanced disease seeking medical treatment[292]. Conversely, there may be a deficit of patients with advanced disease in community cohorts, as they will be lost to follow up, either before or during the period of observation[293]. All patient plasma samples used in this study were collected in 2010 and genomic DNA samples were from the 2003 and 2010 case-control studies. All samples were stored at -80$^{o}$C and transported to Oxford shortly after collection.

This study used samples from 60 patients selected from the Caió community cohort in 2010 by Dr Thushan De Silva. Participants were chosen on the basis of the following criterion: HIV-2 mono-infected patients only, present in the village at the time of sero-survey, no participation in any trial for the preceding 3 years and the availability of at least one set of clinical measurements (viral load, CD4+ T-Cell count and CD4+ T-cell %) taken prior to 2010. The characteristics of the study sub-group are shown in **Table 2.1**.

Table 2.1: Characteristics of the 60 subjects in the study sub-group

| Characteristic | N (%) |
|---|---|
| Women | 40 (66) |
| Resident in central area of village | 45 (75) |
| Mean age at HIV-2 diagnosis in years (SD) | 49 (14) |
| Ever married at time of HIV-2 diagnosis | 49 (82) |
| Timing of infection: | |
| Known to be infected prior to 1989-91 survey | 23 (39) |
| Documented incident infection after 1991 | 17 (29) |
| No documented infection in 1989 or incident infection | 19 (32) |

Patients were stratified into two groups, progressors and non-progressors, based on viral load data collected in 2010. Patients with a viral load below the limit of detection (<100 copies/mL) were classified as non-progressors and patients with detectable viral loads over 500 copies/mL were classified as progressors or non-controllers. Alongside viral load data, absolute CD4+ T-cell counts (cells/μL); CD4+ T cell percentages (as a % of the total lymphocyte count) and years of follow up since initial diagnosis were also available for all patients. In order to verify the stratification, differences in the mean CD4%, absolute CD4+ T cell counts and years of follow up were assessed using an unpaired t-test implemented in R[294]. For patients with viral loads below the limit of detection, a value of 100 was attributed (**Table 2.2**).

Table 2.2: Markers of clinical progression in the study population.

| Characteristic | Total | Non-Progressors | Progressors | p value |
|---|---|---|---|---|
| n | 60 | 33 | 27 | |
| Viral Load | | | | |
| Mean | 22800 | 100 | 50540 | 0.046 |
| IQR | 100-811 | N/A | 700-15760 | |
| Absolute CD4+ | | | | |
| Mean | 555 | 660.5 | 439 | 0.0005 |
| IQR | 363-692 | 489-823 | 297-524 | |
| CD4% | | | | |
| Mean | 32.3 | 36.37 | 28.81 | 0.006 |
| IQR | 26.3-39.9 | 31.9-42.4 | 22.3-36.6 | |
| Years of Follow Up | | | | |
| Mean | 13.5 | 14.8 | 11.8 | 0.1 |
| IQR | 6-21 | 13-21 | 3-21 | |

Results of the non-paired t tests are shown for progressors and non-progressors. Mean and Interquartile Range (IQR) values are also shown for each partitioned group

For all measures of HIV-2 disease progression there was a significant difference at the 0.05 level between the two groups. Progressors had lower CD4% suggesting more advanced disease status. The number of years of follow up was not statistically different between the two groups, implying that the differences in markers of disease progression are not simply due to differences in observation times. Therefore, when stratified on 2010 viral loads, the two groups are also statistically distinct when other measures of disease progression are considered.

### 2.1.1 Patient data confidentiality and identification codes

All samples collected in 2010 were given a patient-specific code to ensure that the confidentiality of patient medical data was maintained. Samples were assigned codes prior to the start of this project and code allocation followed the standard protocols of the field site laboratory in Caió. The code is composed of two sections, an initial two letters identifying the researcher responsible for sample collection

(in this case 'TD' standing for Thushan De Silva), followed by a 3-digit number. Sample numbers correspond to the order in which the samples were collected at the field site and have no relationship to any attributes of the patient. For patients where there were historical samples available in the central Caió BioBank, the unique BioBank identifier was also attributed. BioBank codes were not used in this study. Where BioBank samples from earlier time points were used, these samples were relabelled according to the code assigned in 2010 in order to avoid confusion.

## 2.2 Materials and Methods Used in Chapter 3

### 2.2.1 Quantification of patient genomic DNA

Genomic DNA had previously been extracted from $10^6$ PBMCs isolated from patient whole blood samples. Extracted DNA was stored at -80°C from sero-surveys in 2003 and 2010. DNA was thawed on ice and quantified using the Qubit 2.0 Fluorometer with the double stranded DNA (dsDNA) broad-range assay (Life Technologies). Qubit working solution was prepared by diluting Qubit reagent 1:200 in Qubit buffer. 199µL Qubit working solution was added to 1µL DNA and mixed by vortexing for 3 seconds. Samples were incubated at room temperature for 3 minutes and fluorescence was measured immediately after incubation. Assay DNA concentrations were calculated using an inbuilt curve-fitting algorithm (a modified Hill plot) calibrated with fluorescence readings of the standards, 0ng/µL and 100ng/µL in Tris-EDTA (TE) buffer. Sample DNA concentrations were calculated by correcting for the initial 1:200 dilution in Qubit working solution.

### 2.2.2 Amplification of HIV-2 *vpx*

The HIV-2 accessory gene *vpx* was amplified in a nested polymerase chain reaction (PCR) from proviral DNA. Initial primer design was carried out using Primer3 v0.4.0[295]. Briefly, all available HIV-2 group A sequences from Gambia and Guinea-Bissau were downloaded from Los Alamos HIV Sequence Database[24]. A whole genome alignment was built using the inbuilt read assembler on Geneious v6.1.6 with HIV-2 CAM2 (1987, Guinea-Bissau, accession number: D00835) as the reference genome for assembly (**Table 2.3**)[296].

Table 2.3: West African HIV-2 *vpx* sequences.

| Name | Country | Accession Number |
|---|---|---|
| CAM2CG | Guinea-Bissau | D00835 |
| MDS | Guinea-Bissau | Z48731 |
| FG | Guinea-Bissau | J03654 |
| MCN13 | Gambia | AY509259 |
| ISY_SBL | Gambia | J04498 |
| MCR35 | Gambia | AY509260 |
| D194 | Gambia | X52223 |

Reference sequences from The Gambia and Guinea-Bissau that were used in initial *vpx* primer design.

Primers were designed to amplify a genomic segment containing the complete *vpx* coding region in a nested PCR. Following initial primer design, primers were re-designed and optimized using sequence data from initial PCR runs and additional sequence data generated from shotgun RNA sequencing (RNA-Seq) HIV-2 whole genome builds (**Table 2.4**). Primer binding co-ordinates are shown relative to the HIV-2 reference strain UC2 (**Figure 2.2**).

Table 2.4: Primers used in vpx amplification.

| Name | Sequence | Position | Orientation |
|---|---|---|---|
| VPX LOR | TAG CAG TGA CTC TTG CCT GG | 7845-7826 | Reverse |
| VPX LOF | GTG GAC TTG CAG CAG GGT AA | 5022-5041 | Forward |
| VPX LIF | AGA AGA GCC ATC AGR GGR GA | 5236-5255 | Forward |
| VPX LIR | TGC TGT TGC TGC ACT AAT CCC | 7756-7736 | Reverse |
| VPX SIR | YTC TGY TGG WGC TTC AGY CA | 5692-5672 | Reverse |
| VPX SOR | TCA AGG GTG TCT CCT TRY CT | 5803-5784 | Reverse |

Positions are given relative to HIV-2 group A strain UC2 (AN:U38293)

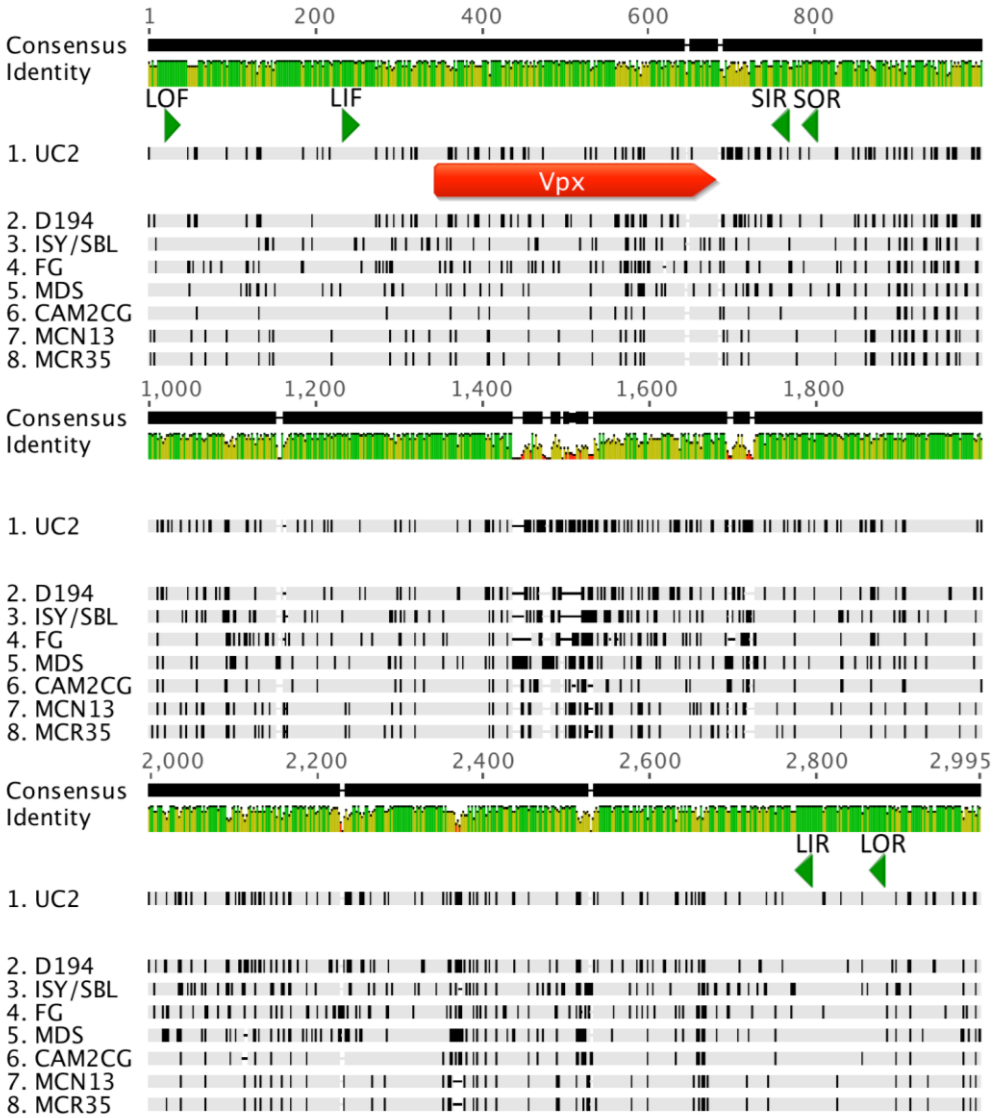



Figure 2.2: Primers used in *vpx* amplification.
Primers are shown in green and the vpx ORF in red. Binding locations are shown relative to the HIV-2 reference strain UC2 (AN:U38293)

HIV-2 *vpx* amplification was carried out using an iterative approach, aiming for long amplicons initially and then re-attempting amplification with primers that would yield shorter fragments if the initial PCR failed. All PCR reactions were carried out using the Advantage 2 polymerase system (TakaraBio). Outer reactions used 250ng genomic DNA and the following components: 1x PCR buffer, 0.2mM each dNTP, 0.6µg Advantage 2 polymerase mix and 0.2µM each of forward and reverse primers. PCR was carried out under the following conditions: initial denaturation at 95°C for 1 minute followed by 40 cycles of 95°C for 15 seconds, touchdown annealing of 58°C to 55° for 30 seconds and an extension at 68°C for 1 minute per kb. Inner round reactions utilised the same PCR components combined

with 1µl of the outer reaction product as a template. Cycling conditions were the same as used in the outer PCRs with extension time adjusted for the shorter inner PCR products. Minimum extension time was 45 seconds, irrespective of whether amplicons were shorter than 750bp. Initial PCR reactions were carried out using VPX LOF & VPX LOR for the outer reaction and VPX LIF & VPX SOR for the inner reactions. If this PCR failed to amplify *vpx*, a second attempt was made using VPX LOF & VPX SOR for the outer reaction and VPX LIF & VPX SIR for the inner reaction.

### 2.2.3 Gel extraction of PCR products

PCR products were purified using QIAQuick Gel Extraction Kit (Qiagen) according to the manufacturer's instructions. PCR products were run on a 2% agarose/ Tris-Acetate-EDTA (TAE) buffer gel at 90V for 1 hour. Appropriately sized products were excised from the gel under ultra-violet (UV) trans-illumination using a clean scalpel. Gel fragments were weighed and 3x volumes of buffer QG were added (100µL QG for 100mg gel). Samples were incubated at 50°C until the gel had completely dissolved, 1x volume of isopropanol was added and samples were applied to a QIAQuick column. Columns were spun at 13,000rpm for 1 minute to facilitate binding of the DNA to the membrane and the membrane was washed with 500µL buffer QG followed by 750µL buffer PE. In order to elute nucleic acids, 30µL DNA-free $H_20$ was applied to the membrane and columns were incubated at room temperature for 5 minutes. Spinning at 13,000rpm for 1 minute facilitated final elution. Purified PCR products were stored at -20°C until needed.

### 2.2.4 Molecular cloning of HIV-2 *vpx* PCR products into *E. Coli* for sequencing

Purified PCR products were ligated into the pCR4-TOPO TA Vector (Life Technologies) and transformed into competent cells according to the manufacturer's protocol (**Figure 2.3**).

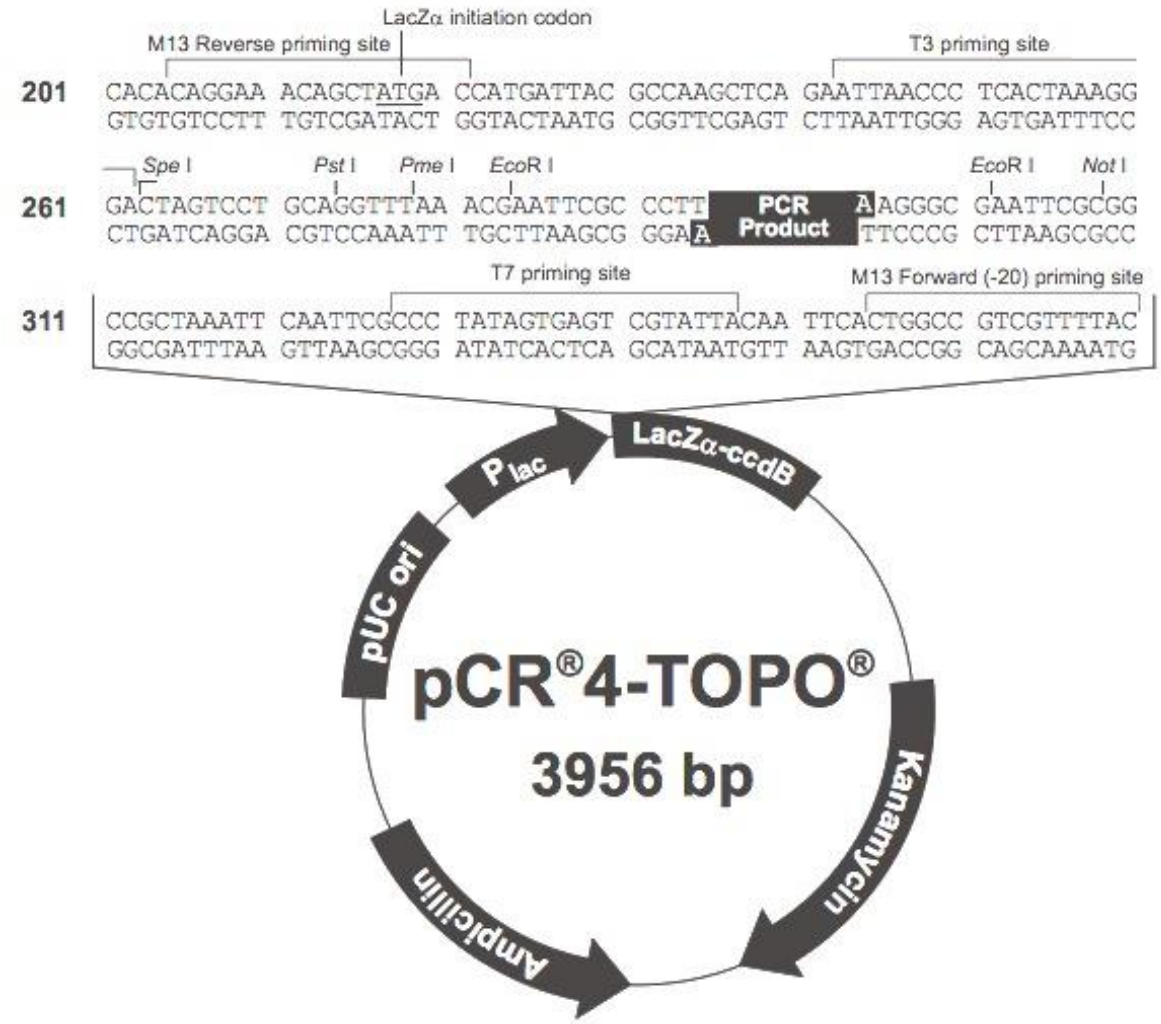



Figure 2.3: Schematic diagram of the pCR4-TOPO TA vector (Life Technologies). Insertion site between the M13F and M13R binding sites is expanded.

4µL PCR product (normalised to a concentration of 5ng/µl in $H_20$) was incubated with 1µl pCR4-TOPO vector and 1µl salt solution (1.2M NaCl and 0.06M $MgCl_2$) at room temperature for 20 minutes. Following ligation, 4µl ligation mixture was added to $1x10^9$ chemically competent DH5$\alpha$-T1 *E. coli* cells. Cells were incubated on ice for 30 minutes. Transfection was conducted via heat shock; cells were heated in a water bath set at 42°C for 30 seconds and then placed on ice for 2 minutes. Following heat shock, 250µl Super Optimal broth with Catabolite repression (SOC) medium (2% tryptone, 0.5% yeast extract, 10mM NaCl, 2.5mM KCl, 10mM $MgCl_2$, 10mM $MgSO_4$ and 20mM glucose, Invitrogen) was added to the cells in an aseptic environment to obtain maximal transfection efficiency. Cells were incubated in a shaking incubator (37°C, 5% $CO_2$) for 1 hour prior to plating. 100µl bacteria were plated onto Lysogeny Broth (LB) Agar plates (2% LB Broth powder, 1.5% Agar)

supplemented with 1.25mg Kanamycin per plate. Plates were incubated for 16 hours at 37°C to allow colonies to grow.

### 2.2.5 Colony PCR of transformed *E. coli* cells

Following overnight incubation, 24 colonies per plate were picked for colony PCR amplification prior to sequencing. Using a fresh sterile tip per colony, a trace of each colony was combined with the following PCR reagents (Platinum Taq, Invitrogen): 1x PCR buffer, 1.5mM $MgCl_2$, 0.2mM dNTP, 0.08µl Platinum Taq polymerase and sequencing primers M13F (3'-TGT AAA ACG ACG GCC ACT-5') and M13R (3'-AGG AAA CAG CTA TGA CCA T-5') at 0.2µM each. PCR was carried out under the following conditions: initial denaturation at 94°C for 2 minutes followed by 30 cycles of 94°C for 30 seconds, 55°C for 30 seconds and 72°C for 1 minute. PCR products were resolved on a 1% agarose (TAE) gel and size was quantified using a 100bp ladder (New England BioLabs). Expected amplicon size following ligation into pCR4-TOPO TA is original amplicon size plus 133bp as the M13R and M13F primer binding sites are in the multiple cloning region of the vector (**Figure 2.3**).

### 2.2.6 Clean up and sequencing of colony PCR products

Colonies containing inserts were identified by the presence of an appropriately sized colony PCR product. Successful reactions were cleaned using a mixture of Shrimp Alkaline Phosphatase (SAP) and Exonuclease I (ExoI) (New England BioLabs). 0.025µl ExoI (0.5 U), 0.25µl SAP (0.25U) and 9.725µl $H_2O$ were combined with 22µL PCR product and the clean-up reaction was conducted at 37°C for 30 minutes followed by an inactivation step at 85°C for 15 minutes. 20 reactions were selected per sample, PCR amplicons were diluted 1:5 with $H_2O$ and sequencing was carried out by Source BioScience using a 48 capillary ABI-3730 DNA analyser and Applied Biosystems BigDye Terminator v3.1 chemistry. Sequencing reactions were run using both the M13 forward and reverse sequencing primers to allow for error correction and ambiguous peak calling.

### 2.2.7 Read cleaning and assembly of full length *vpx* sequences

As *vpx* is 339bp in length, complete gene sequences can be recovered from a single Sanger read (approx. 800bp). Forward and reverse sequencing chromatograms were visualized using FinchTV v1.4 (Geospiza)[297] and aligned using the MUSCLE alignment algorithm[298] implemented in Geneious v6.1.6. As the reads are from clonal populations, MUSCLE was run using 8 iterations, 1 tree build and a default window size of 5 letters. Aligned forward and reverse reads were compared and any conflicts in base calling were resolved by using the highest quality read. Sequences were trimmed to include only the *vpx* coding region, defined as positions 5898-6237 in HIV-2 group A strain UC2 (AN:U38293). Full-length *vpx* sequences from patients were aligned to the reference strain UC2 using MUSCLE implemented in Geneious to ensure they were in the correct orientation. Only full-length *vpx* sequences were included in the analysis. Open reading frames were identified in Geneious and any sequences containing premature stop codons were removed from the final analysis.

### 2.2.8 Nucleotide substitution model testing

For each phylogenetic or evolutionary analysis, the appropriate nucleotide substitution model was identified using JModelTest2[299]. Four models of nucleotide substitution were assessed; Jukes-Cantor (JC)[300], Felsenstein (F81)[301], Hasegawa *et al* (HKY85)[302] and General Time Reversible (GTR)[303]. In order to compare any two models they must be nested so that the null model is equivalent to a restriction of one or more of the parameters in the parameter-rich model. The nesting of the models and the degrees of freedom (df) added when moving from the null to parameter-rich model is depicted in **Figure 2.4**.

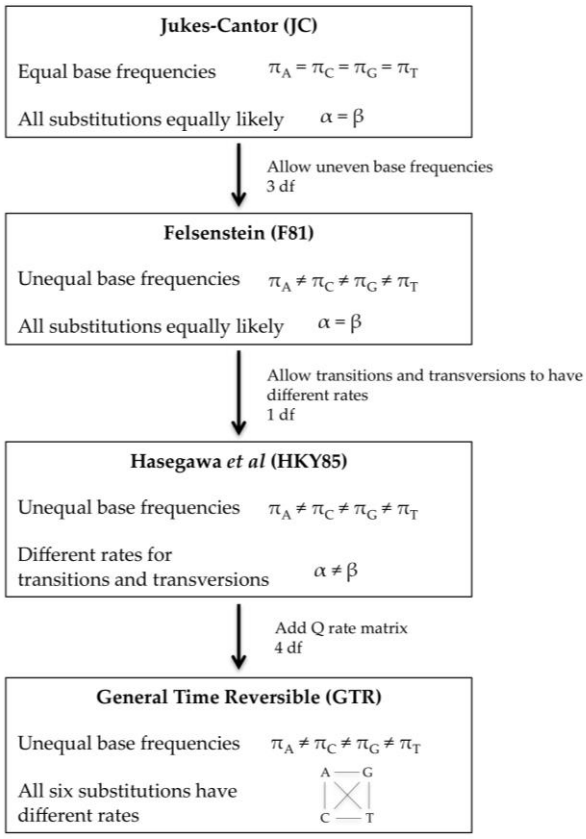



Figure 2.4: Nucleotide substitution models.
Arrows represent null to alternative (parameter-rich) models. Each model is briefly explained and the degrees of freedom between each comparison are shown as df.

The log likelihood was computed for each model and log likelihoods were compared using the likelihood ratio test (LRT) statistic:

$$\text{LRT} = 2(\ell 1 - \ell 0)$$

where $\ell 0$ is the log likelihood of the null model and $\ell 1$ is log likelihood of the alternative model. Under the null hypothesis, the LRT statistic is asymptotically distributed as a $\chi^2$ distribution with df equal to the difference in the number of free parameters between the two models.

The log likelihoods, LRT values, p-values and appropriate model for each comparison are shown in **Table 2.5**. For comparisons of molecular clock rate between patients it is advantageous to perform all analyses under the same substitution and clock models. Therefore, for analysis of longitudinal *vpx*

sequences from individual patients, the HKY85 model of nucleotide substitution was selected. In addition to the nucleotide substitution model, analyses were also run with a proportion of invariant sites (+I) and rate heterogeneity modelled as a gamma distribution (+G).

Table 2.5: Summary of the log likelihood and LRT statistic values.

**Patient Clustering**

| Null Model | Model 1 | $\ell_0$ | $\ell_1$ | LRT | DF | p-value | |
|---|---|---|---|---|---|---|---|
| JC | F81 | -6008.3 | -5960.1 | 96.4 | 3 | 0 | |
| F81 | HKY85 | -5960.1 | -5497.0 | 926.3 | 1 | 0 | |
| HKY85 | GTR | -5497.0 | -5486.6 | 20.8 | 4 | 4.00E-04 | ** |

**Time Point Clustering**

| Null Model | Model 1 | $\ell_0$ | $\ell_1$ | LRT | DF | p-value | |
|---|---|---|---|---|---|---|---|
| JC | F81 | -2822.6 | -2794.3 | 56.6 | 3 | 0 | |
| F81 | HKY85 | -2794.3 | -2579.4 | 429.8 | 1 | 0 | |
| HKY85 | GTR | -2579.4 | -2565.9 | 27.1 | 4 | 2.00E-05 | ** |

**Viral Subtyping**

| Null Model | Model 1 | $\ell_0$ | $\ell_1$ | LRT | DF | p-value | |
|---|---|---|---|---|---|---|---|
| JC | F81 | -2705.2 | -2589.7 | 231.0 | 3 | 0 | |
| F81 | HKY85 | -2589.7 | -2578.5 | 22.4 | 1 | 3.00E-06 | |
| HKY85 | GTR | -2578.5 | -2571.6 | 13.8 | 4 | 0.008 | ** |

**TD035**

| Null Model | Model 1 | $\ell_0$ | $\ell_1$ | LRT | DF | p-value | |
|---|---|---|---|---|---|---|---|
| JC | F81 | -548.2 | -540.9 | 14.7 | 3 | 0.002 | |
| F81 | HKY85 | -540.9 | -537.7 | 6.4 | 1 | 0.011 | ** |
| HKY85 | GTR | -537.7 | -534.5 | 6.3 | 4 | 0.18 | |

**TD041**

| Null Model | Model 1 | $\ell_0$ | $\ell_1$ | LRT | DF | p-value | |
|---|---|---|---|---|---|---|---|
| JC | F81 | -912.7 | -904.0 | 17.4 | 3 | 0.0006 | |
| F81 | HKY85 | -904.0 | -840.4 | 127.2 | 1 | 0 | ** |
| HKY85 | GTR | -840.4 | -838.9 | 3.0 | 4 | 0.55 | |

**TD046**

| Null Model | Model 1 | $\ell_0$ | $\ell_1$ | LRT | DF | p-value | |
|---|---|---|---|---|---|---|---|
| JC | F81 | -672.2 | -662.9 | 18.6 | 3 | 0.00034 | |
| F81 | HKY85 | -662.9 | -634.0 | 58.0 | 1 | 0 | ** |
| HKY85 | GTR | -634.0 | -632.2 | 3.5 | 4 | 0.47 | |

**TD047**

| Null Model | Model 1 | $\ell_0$ | $\ell_1$ | LRT | DF | p-value | |
|---|---|---|---|---|---|---|---|
| JC | F81 | -705.0 | -697.8 | 14.3 | 3 | 0.003 | |
| F81 | HKY85 | -697.8 | -684.4 | 26.9 | 1 | 0.000001 | ** |
| HKY85 | GTR | -684.4 | -683.2 | 2.4 | 4 | 0.66 | |

**TD050**

| Null Model | Model 1 | $\ell_0$ | $\ell_1$ | LRT | DF | p-value | |
|---|---|---|---|---|---|---|---|
| JC | F81 | -570.7 | -565.1 | 11.1 | 3 | 0.01 | |
| F81 | HKY85 | -565.1 | -549.2 | 31.9 | 1 | 0.000001 | ** |
| HKY85 | GTR | -549.2 | -549.2 | 0.0 | 4 | N/A | |

**TD055**

| Null Model | Model 1 | $\ell_0$ | $\ell_1$ | LRT | DF | p-value | |
|---|---|---|---|---|---|---|---|
| JC | F81 | -650.4 | -642.4 | 16.0 | 3 | 0.001 | |
| F81 | HKY85 | -642.4 | -633.7 | 17.4 | 1 | 0.000031 | |
| HKY85 | GTR | -633.7 | -626.1 | 15.3 | 4 | 0.004 | ** |

**TD058**

| Null Model | Model 1 | $\ell_0$ | $\ell_1$ | LRT | DF | p-value | |
|---|---|---|---|---|---|---|---|
| JC | F81 | -652.5 | -644.2 | 16.7 | 3 | 0.000817 | |
| F81 | HKY85 | -644.2 | -634.5 | 19.4 | 1 | 0.000011 | ** |
| HKY85 | GTR | -634.5 | -630.3 | 8.3 | 4 | 0.08 | |

**TD061**

| Null Model | Model 1 | $\ell_0$ | $\ell_1$ | LRT | DF | p-value | |
|---|---|---|---|---|---|---|---|
| JC | F81 | -654.0 | -643.4 | 21.1 | 3 | 0.0001 | |
| F81 | HKY85 | -643.4 | -623.8 | 39.3 | 1 | 0 | ** |
| HKY85 | GTR | -623.8 | -620.8 | 5.8 | 4 | 0.21 | |

**TD093**

| Null Model | Model 1 | $\ell_0$ | $\ell_1$ | LRT | DF | p-value | |
|---|---|---|---|---|---|---|---|
| JC | F81 | -585.7 | -578.5 | 14.5 | 3 | 0.002 | |
| F81 | HKY85 | -578.5 | -567.8 | 21.3 | 1 | 4.00E-06 | ** |
| HKY85 | GTR | -567.8 | -566.5 | 2.8 | 4 | 0.6 | |

P-values were obtained from a $\chi^2$ distribution and the chosen model is shown by **.

### 2.2.9 Preparation of sequencing libraries using NexteraXT

PCR products from a subset of patients were selected for additional sequencing on the Illumina MiSeq platform. Samples included in the subset were all from 2010 and all had been successfully amplified in the first nested PCR, resulting in longer amplicons. The reason for this criterion is that there is a drop off in coverage at the distal ends of amplicons caused by the library preparation protocol and so longer fragments will provide more even coverage over the whole *vpx* coding region. Library preparation was carried out using the NexteraXT protocol (Illumina) according to the manufacturer's instructions. NexteraXT was chosen as it is optimised for preparation of amplicons between 500-1500bp in length[304]. NexteraXT library preparation consists of five stages: tagmentation of DNA, PCR amplification, PCR clean up, library normalisation and library pooling. Briefly, PCR products were quantified using the Qubit dsDNA assay as described above and normalised to a concentration of 0.2ng/µl. 5µl DNA was added to 10µl tagment DNA buffer (TD) and 5µl amplicon tagment mix (ATM). Samples were mixed, briefly centrifuged and incubated at 55°C for 5 minutes. Following tagmentation, neutralisation was performed by adding 5µl neutralise tagment buffer (NT) and incubating at room temperature for 5 minutes. Tagmented DNA was amplified in a limited-cycle PCR, which also adds Illumina index primers 1 (i7) and 2 (i5), required for cluster formation and sample identification (**Figure 2.5**).

| | | | INDEX PRIMERS i7 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | N701 | N702 | N703 | N704 | N705 | N706 | | | | | | |
| | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| INDEX | S501 | A | 1 | 2 | 3 | 4 | 5 | 6 | | | | | | |
| PRIMERS | S502 | B | 7 | 8 | 9 | 10 | 11 | 12 | | | | | | |
| i5 | S503 | C | 13 | 14 | 15 | 16 | 17 | 18 | | | | | | |
| | S504 | D | 19 | 20 | 21 | 22 | 23 | 24 | | | | | | |
| | | E | | | | | | | | | | | | |
| | | F | | | | | | | | | | | | |
| | | G | | | | | | | | | | | | |
| | | H | | | | | | | | | | | | |

Figure 2.5: Illumina index scheme.
Layout of samples 1 → 24 in a 96 well plate. Positions of Illumina index primers i7 and i5 are shown.

PCR amplification was conducted by adding 15µl Nextera PCR master mix (NPM), 5µl index primer 1 and 5µl index primer 2 to tagmented DNA. Cycling was carried out under the following conditions: initial extension at 72°C for 3 minutes, initial denaturation at 95°C for 30 seconds, 12 cycles of 95°C for 10 seconds, 55°C for 30 seconds and 72°C for 30 seconds followed by a final elongation at 72°C for 5 minutes.

PCR clean up was conducted using AMPure XP beads. Following transfer of PCR products to a fresh plate, 30µl of AMPure XP beads were added and incubated at room temperature for 5 minutes. Bead washing was conducted on a magnetic stand; supernatant was removed and discarded and two 30-second washes with 80% ethanol were carried out. Clean libraries were eluted in 52.5µl resuspension buffer (RSB) and all beads were removed using a magnetic stand.

For library normalisation, 20µl of each library was added to 7µl library normalisation beads 1 (LNB1) suspended in 38µl library normalisation additives 1 (LNA1). Samples were shaken at 1800rpm for 30 minutes at room temperature. Clean up was performed on a magnetic stand. Supernatant was removed and discarded and two washes were performed with library normalisation wash 1 (LNW1). Briefly, 45µl of LNW1 were added to the beads, samples were shaken at 1800rpm for 5 minutes and supernatant was removed on a magnetic stand. Samples were eluted in 30µl 0.1 N NaOH and shaken at 1800rpm for 5 minutes before beads were removed on a magnetic stand.

Pooled amplicon library (PAL) was generated by adding 5µl each library to the PAL and multiplexed libraries were sent to the Wellcome Trust Centre for Human Genetics (WTCHG) genomics core for sequencing on a single lane of the Illumina MiSeq platform.

### 2.2.10 Quality control and read cleaning

Read quality was visualised from .fastq files in FastQC (**Figure 2.6**)[305]. Read quality was assessed using Phred quality scores generated by the MiSeq. A Phred quality score (also known as $Q_{Sanger}$) is an integer mapping of the probability that

the position is incorrectly called (p)[306]. MiSeq uses the standard Sanger variant to assess base calling confidence, shown below:

$$Q_{Sanger} = -10 \log_{10} p$$

Quality scores along the length of the read can be summarized in FastQC. A Q score of 20 represents an estimated error rate of 0.01 and a Q score of 30 represents an error rate of 0.001. However, it is worth noting that Illumina quality scores have a tendency to underestimate the error rate for high quality scores and overestimate error scores for low quality bases[307].

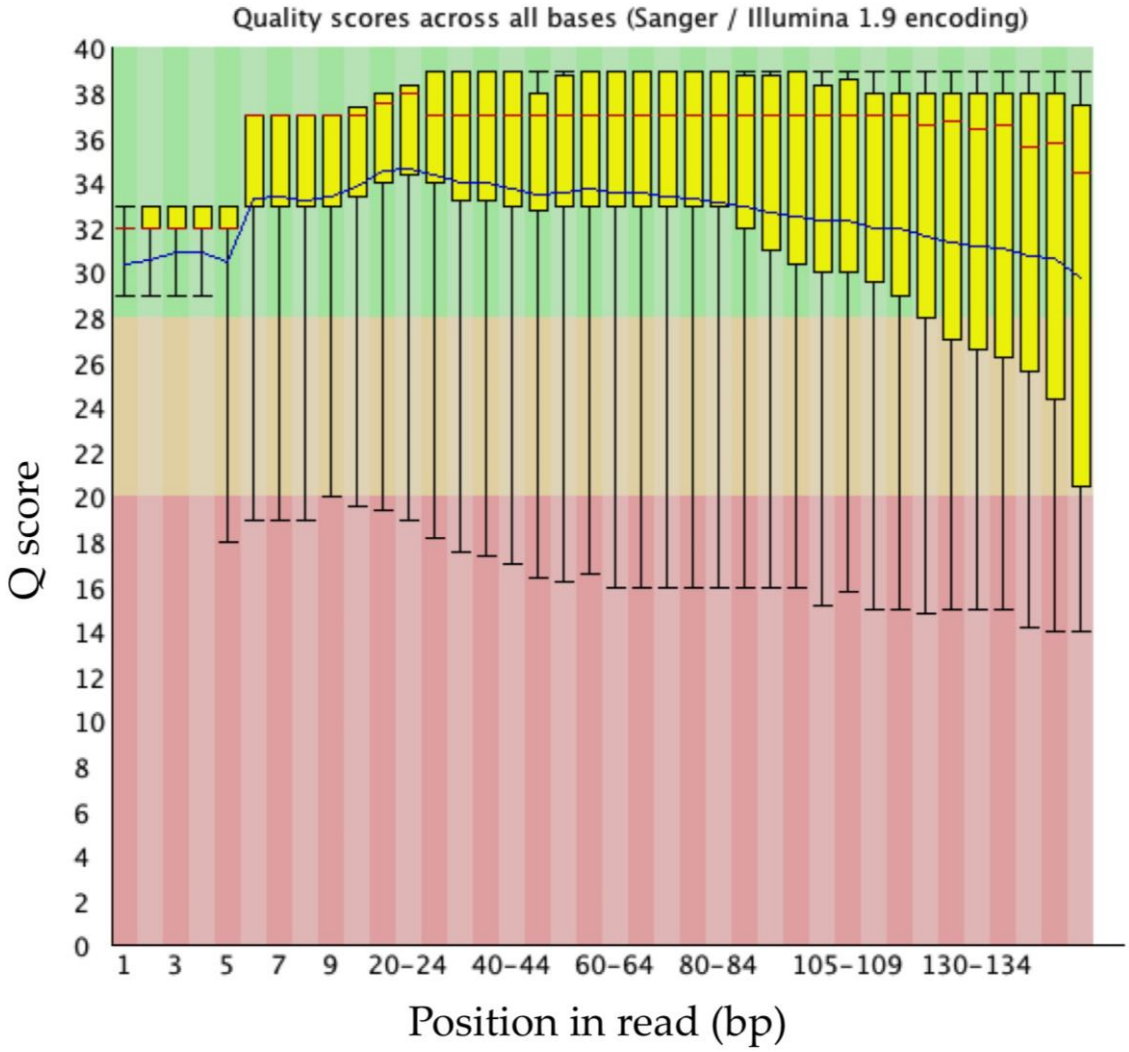



Figure 2.6: Q-score plot.
Quality score is shown for each position in the read. Q score is on the y-axis. Median Q scores are shown in red, yellow boxes represent IQR and whiskers show 95% interval.

Low quality reads were removed using the Sickle package[308]. A conservative Q score of 30 was chosen as the cut off for removal and a minimum read length of 40bp following quality trimming was stipulated (**Figure 2.7**).

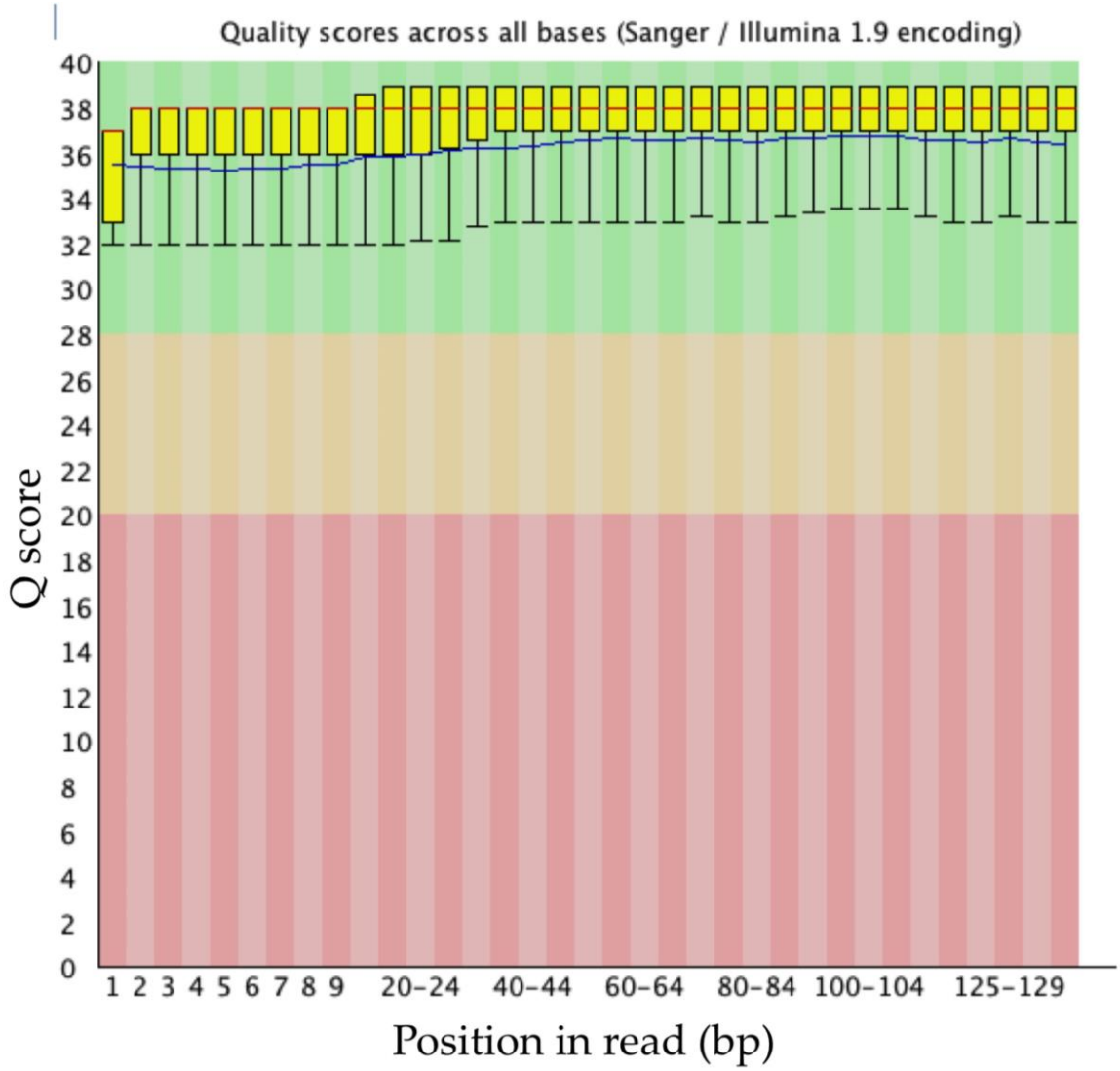



Figure 2.7: Q score plot following read trimming.
Reads were trimmed based on Q score (Q>30).

### 2.2.11 Read assembly and variant calling

Trimmed .fastq files were assembled into .bam files using Bowtie2[309]. For each sample, the patient-specific consensus sequence generated from clonal sequences was used as the reference genome. Reference genomes were indexed using the Bowtie2-build function and assembled using Bowtie2. Sam files containing assembled reads were converted into sorted bam files, unmapped reads were filtered out, PCR duplicates generated during library preparation were removed

using Picard and a bai index file was created using the package Samtools[310], [311]. Bam files were visualised in Integrative Genomics Viewer (IGV) with average depth of coverage was 7,000x per position[312]. The frequency of each variant position identified in the analysis of clonally derived sequences was assessed.

## 2.3 Materials and Methods Used in Chapters 4 & 5

### 2.3.1 Propagation of viral reference strains

Viral isolates HIV-2 ROD and HIV-2 CBL-20 were propagated *in vitro* prior to RNA-Seq in the lymphocyte cell line H9, a single cell clone derived from a HUT 78 cell line. H9 cells were propagated in RPMI 1640 medium supplemented with 10% foetal bovine serum, 2mM L-glutamine and 50U/ml penicillin/streptomycin (commonly known as R10 medium) at 37°C and 5% $CO_2$. H9 cells were assayed for mycoplasma infection prior to infection and all cultures tested negative[313].

Infection with viral isolates used $5x10^6$ cells per experiment. Cells were counted, pelleted and then gently re-suspended in 200µL cell-free virus, concentrated to $9 \times 10^3$ TCID50/mL. Initial incubation of cells in highly concentrated virus stock was conducted at 37°C for 1 hour. Following initial incubation, 5mL R10 was added to the cells (Day 0). Cells were fed with 5mL R10 every third day or as needed and incubated at 37°C and 5% $CO_2$. Supernatant was harvested at days 4,7,11,14 and 18 post infection (HIV-2 ROD) and days 5,7,9,12 and 14 (HIV-2 CBL-20) depending on how the cell lines were growing during the incubation period. Care was taken not to let the cell cultures over-grow, whilst simultaneously minimizing the total culture volume, in order to maximize virus concentration. To harvest supernatant, cells were gently pelleted at 250xg for 10 minutes, supernatant was removed and frozen at -80°C and the cell pellet was re-suspended in 5mL fresh R10 and incubated at 37°C and 5% $CO_2$.

### 2.3.2 Quantification using reverse transcriptase (RT) activity assay

Supernatants were assayed for HIV-2 concentration with a reverse transcriptase colorimetric assay (Roche, UK) according to manufacturer's instructions. Briefly, virus was precipitated from 1mL supernatant overnight at 4°C in Polyethylene glycol (PEG) solution (30% PEG 6000 in 1.2M NaCl). Precipitated virus was pelleted at 800xg for 45 minutes and supernatant was removed and discarded. The viral pellet was re-suspended in 80µl lysis buffer and incubated at room temperature for 30 minutes. A serial dilution series of reverse transcriptase of

known concentration was prepared for assay calibration. Calibration curve standards were prepared at concentrations of 0,2,1,0.5,0.25,0.125 and 0.0625 ng/well. 40µl reaction buffer 3 was added to each sample and samples were incubated at 37°C for 1 hour. Following lysis, 40µl of sample was transferred to a 96-well reaction plate and samples were measured in duplicate. The plate was incubated at 37°C for 1 hour. The reaction plate was washed 5 times with 250µl washing buffer per well per wash. 200µl anti-DIG-POD solution was added to each well and the plate was incubated in the dark at 37°C for 1 hour, followed by another 5 washes with 250µl washing buffer. 200µl ABTS solution was added to each well and the plate was incubated at 37°C for 30 minutes. Absorbance was measured at 405nm in an ELISA plate reader and a standard curve was plotted in Excel. The standard curve was used to convert sample absorbance readings into reverse transcriptase concentrations in ng/µL (**Table 2.6**).

Table 2.6: Mean reverse transcriptase concentrations.

| Sample | Day | RT Concentration (ng/µL) |
|---|---|---|
| ROD | 4 | 0 |
| ROD | 7 | 0 |
| ROD | 11 | 0.17 |
| ROD | 14 | 1.41 |
| ROD | 18 | 1.31 |
| CBL-20 | 5 | 0 |
| CBL-20 | 7 | 0.19 |
| CBL-20 | 9 | 1.64 |
| CBL-20 | 12 | >2 |
| CBL-20 | 14 | 1.55 |

RT activity was quantified for cultured HIV-2 reference isolates at each collection point and used to ascertain the level of viral replication.

### 2.3.3 Patient plasma samples

Dr Thushan De Silva collected plasma samples from the Caió cohort in 2010. Samples were stored at -80°C and transported to Oxford. The Gambia Government/MRC joint ethics committee and the Oxford Tropical Research Ethics Committee (OXTREC) granted ethical approval for the use of patient samples in this study. Clinical characteristics of the patients included in this study are explained in detail in Chapter 2.1.

### 2.3.4 RNA extraction

RNA was extracted from 1mL culture supernatant using a modified version of the QIAamp UltraSens Virus Kit (Qiagen, UK). 1ml frozen supernatant was thawed on ice and added to 800µl buffer AC with 2µl linear acrylamide (5mg/mL, Ambion, USA). Linear acrylamide was used as the nucleic acid co-precipitant in place of carrier RNA. Carrier RNA is a poly-A RNA that is commonly used as a nucleic acid co-precipitant. Linear acrylamide was used instead as carrier RNA would contaminate the RNA-Seq library, reducing the number of informative reads. Following addition of buffer AC, samples were mixed on a vortex for 10 seconds and then incubated at room temperature for 10 minutes. Samples were spun at 1200xg for 3 minutes and supernatant was removed and discarded. Pelleted cell debris and nucleic acids were re-suspended on a vortex in 300µl buffer AR heated to 60$^{o}$C and 20µl proteinase K. Proteinase K treatment was conducted for 10 minutes at 40$^{o}$C in a shaking incubator. 300µl of binding buffer AB was added to each sample and samples were passed through a QIAamp MinElute column to facilitate binding of nucleic acids to the membrane. QIAamp MinElute columns were used in place of the QIAamp Mini Spin Columns provided with the kit, as they allow for final elution into a volume of 10µl, resulting in a more concentrated RNA sample. Columns were washed with 500µl buffer AW1 followed by 500µl buffer AW2 and dried by spinning at full speed for 2 minutes. Final elution of nucleic acids was conducted by adding 12µl $H_2O$ to the membrane, incubating at room temperature for 3 minutes and spinning at 6000xg for 1 minute. Following elution, good RNA practice was followed, samples were kept on ice or stored at -80$^{o}$C if not needed and all downstream experiments were conducted as quickly as possible after elution.

RNA was extracted from patient plasma samples using the same protocol as the viral culture supernatants. The only modification was that a smaller starting volume of patient plasma was used. Initially, 500µl patient plasma was diluted in 500µl of phosphate buffered saline (PBS) to create a starting volume of 1mL.

### 2.3.5 DNase treatment

Eluted nucleic acid samples contain both RNA and DNA. Therefore it is important to perform a DNase treatment to remove DNA, which would contaminate the RNA-Seq libraries. DNase treatment was conducted using the TURBO DNA-free kit (Ambion, USA) according to the manufacturer's protocol. Briefly, 0.1 volumes of both TURBO DNase buffer and TURBO DNase (1µl) were added to the RNA and DNase treatment was conducted at 37$^{o}$ for 25 minutes. DNase was inactivated by addition of 0.1 volumes of DNase inactivation reagent (1µl), gentle mixing by inversion and incubation at room temperature for 5 minutes. Inactivation reagent was removed by spinning at 10,000xg for 90 seconds and transfer of supernatant to a clean, RNase free tube. RNA samples were used immediately or stored at -80$^{o}$C until needed.

### 2.3.6 RNA quantification

RNA concentration was estimated using the Qubit 2.0 Fluorometer with the RNA assay. RNA quantification follows the same method as DNA quantification, described earlier in this chapter, but the standards and fluorescence reagents differ. For samples that fell below the level of detection (250pg/µl), concentration was estimated from the raw fluorescence readings. Normally, a curve-fitting algorithm is used to fit a curve to the standard and sample readings in order to calculate concentration (**Figure 2.8**).

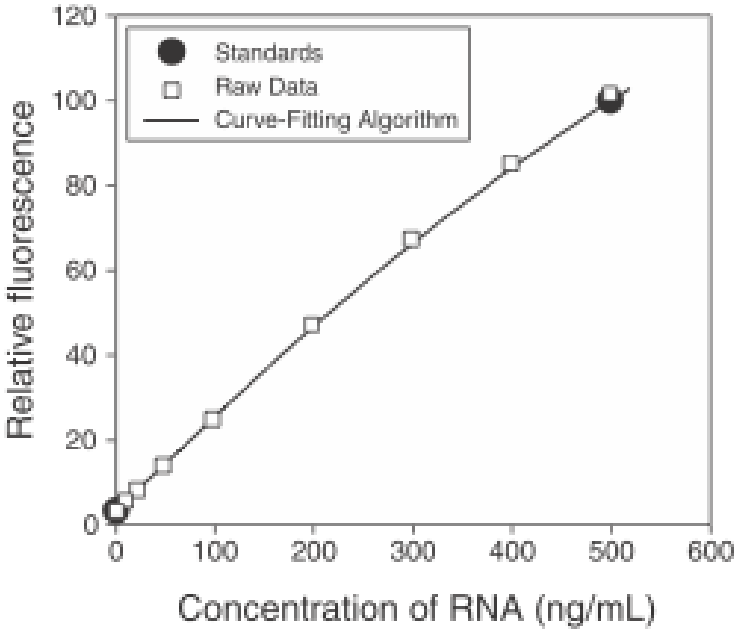



Figure 2.8: Qubit RNA assay standard curve.
A modified Hill plot is used to fit the standard curve, allowing an estimation of RNA concentration.

As there are only two standard measurements, the correlation was modelled as a straight line and the raw fluorescence readings from the samples below detection were used to estimate RNA concentration.

### 2.3.7 NEBNext RNA-Seq library preparation

DNase-treated RNA samples were prepared for sequencing using the NEBNext Ultra RNA Library Prep Kit for Illumina (New England BioLabs). Post-quantification, all samples fell below the maximum recommended input of 1µg of RNA, so library preparation was performed without any dilution of samples. Library preparation was performed according to the manufacturer's protocol, without mRNA isolation. Briefly, 5µl of RNA was added to 4µl NEBNext first strand synthesis reaction buffer and 1µl random primers. RNA denaturation and primer annealing was performed at 94°C for 15 minutes followed by incubation on ice. 0.5µl of Murine RNase inhibitor, 1µl M-MuLV RNase H- reverse transcriptase and 8.5µl $H_2O$ was added to each sample followed by first strand synthesis of 25°C for 10 minutes, 42°C for 50 minutes and 70°C for 15 minutes. Second strand cDNA synthesis was performed by adding 8µl second strand synthesis reaction buffer,

4µl second strand synthesis enzyme mix and 48µl $H_2O$, followed by incubation at 16°C for 1 hour.

Double-stranded cDNA was purified using 1.8x volumes (144µl) AMPure XP beads. Following addition of beads, samples were incubated at room temperature for 5 minutes, placed on a magnetic stand and supernatant was removed and discarded. Beads were washed twice on the magnetic stand with 80% EtOH and DNA was eluted in 55µl $H_2O$. Beads were removed and samples transferred to a clean tube. End repair/dA-tail addition was performed by addition of 6.5µl NEBNext end repair reaction buffer and 3µl NEBNext end prep enzyme mix followed by incubation at 20°C for 30 minutes and 65°C for 30 minutes. Adaptor ligation was performed by the addition of 15µl Blunt/TA ligase master mix, 1µl NEBNext adaptor and 2.5µl $H_2O$. Ligation was carried out at 20°C for 15 minutes. Following adaptor ligation, fragmented libraries were size selected using AMPure XP beads. 0.6 volumes (60µl) beads were added to each sample, incubated at room temperature for 5 minutes and placed on a magnetic rack. Supernatant was transferred to a fresh plate and beads bound to larger DNA fragments were discarded. 0.25 volumes (25µl) beads were added to the supernatant, incubated at room temperature for 5 minutes, supernatant was discarded and beads were washed twice with 200µl 80% EtOH. Beads were air-dried for 10 minutes to ensure all EtOH was removed and DNA was eluted in 20µl $H_2O$.

USER excision and PCR library enrichment was carried out in one step. To the size selected DNA, 3µl uracil-specific excision reagent (USER) enzyme, 25µl NEBNext high-fidelity PCR master mix, 1µl universal PCR primer and 1µl index primer were added. A unique index primer was used for each sample, allowing resolution of reads by patient following multiplexing. USER digestion was carried out at 37°C for 15 minutes followed by a limited cycle PCR of 98°C for 10 seconds and then 12 cycles of 98°C for 10 seconds, 65°C for 30 seconds and 72°C for 30 seconds, followed by a final extension of 72°C for 5 minutes. Prepared libraries were stored at -20°C until sequencing.

### 2.3.8 Sequencing library quality control

Library quality was assessed using an Aligent DNA high sensitivity chip read on a Bioanalyser. Libraries were diluted 1:2 with $H_2O$ before quality control and chips were prepared according to the manufacturer's protocol. Briefly, 9µl gel-matrix containing high sensitivity DNA dye was injected under pressure onto a high sensitivity DNA chip. 5µl of marker was added to each well and 1µl DNA ladder was added to the ladder well. 1µl of each sample was loaded onto the chip, the chip was vortexed at 2400 rpm for 1 minute and then immediately read on the Bioanalyser.

Library quality was assessed for four factors; fragment size, library concentration, presence of adaptor-dimer contamination and presence of PCR artefacts (**Figures 2.9 – 2.11**).

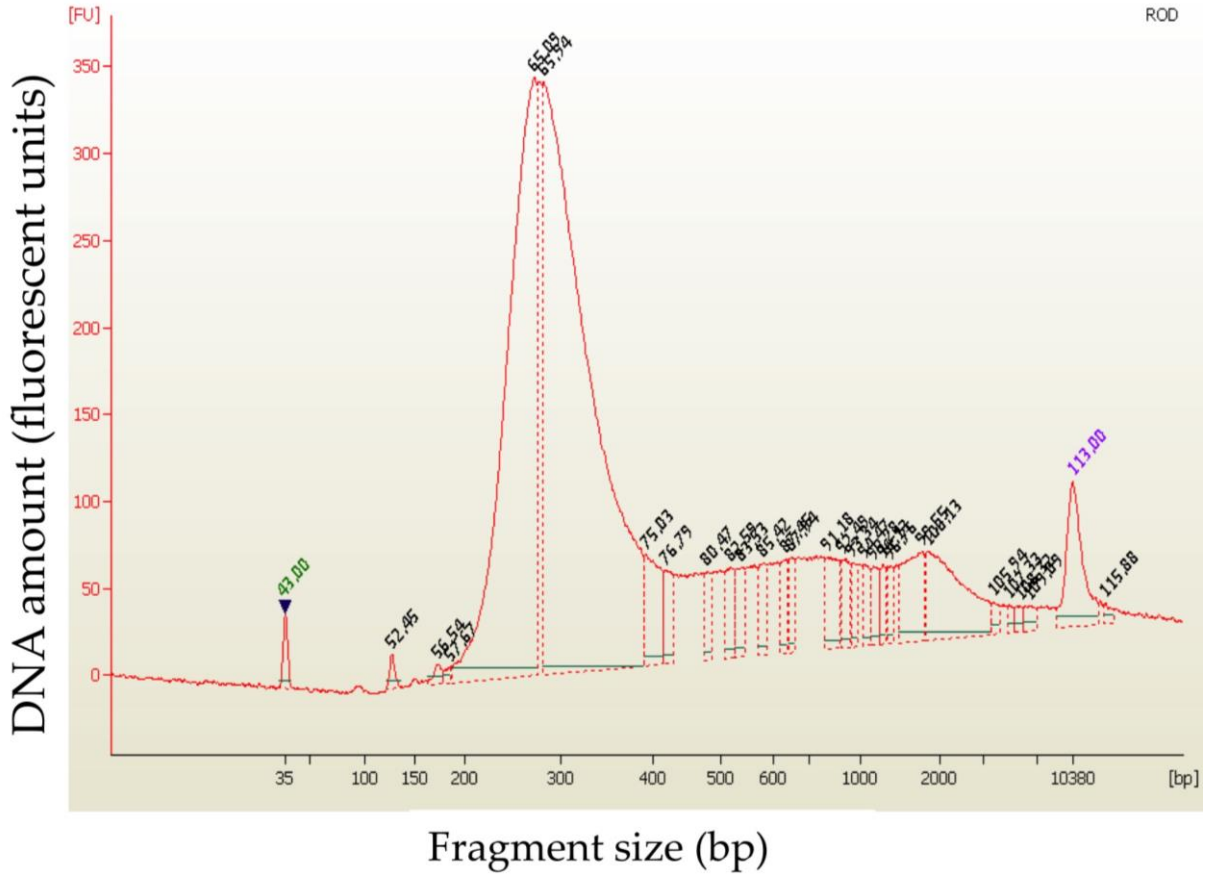



Figure 2.9: Library visualisation.
Trace depicts a good quality library with fragments concentrated between 200-400bp with little adaptor-dimer contamination

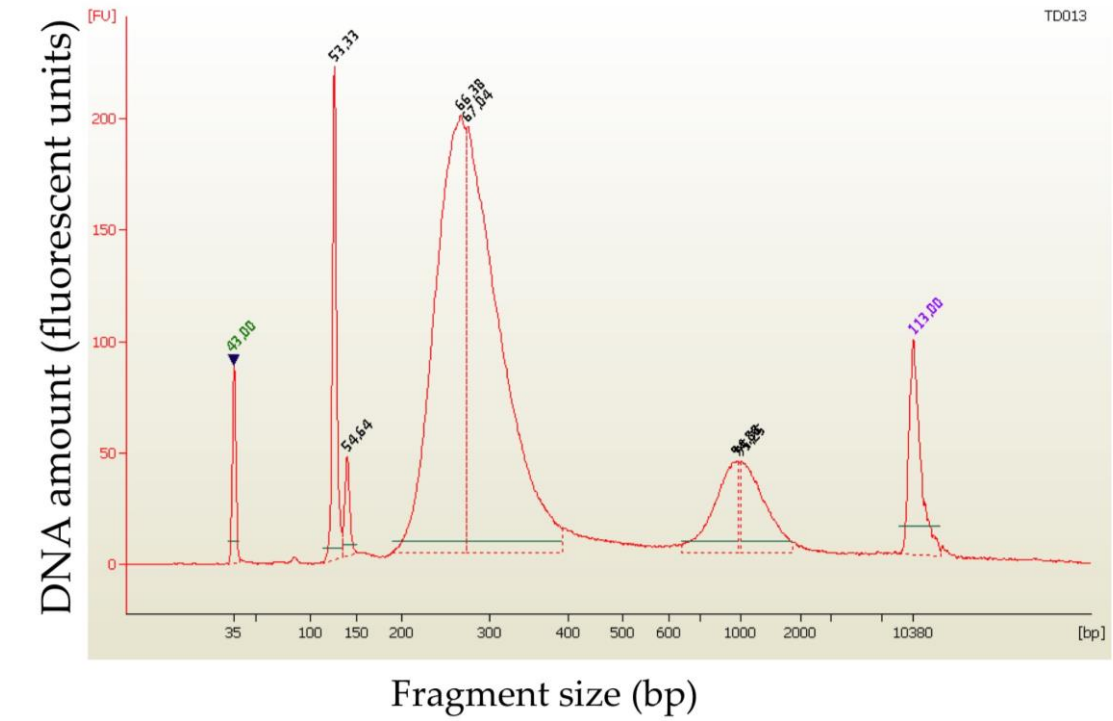



Figure 2.10: Library visualisation showing PCR over-amplification.
Peak at 53.33 seconds (127bp) is adaptor-dimer contamination. Secondary peak at 600-2000bp is due to a PCR artefact from over amplification. In this instance, the large secondary peak was removed by size-selection with AMPure beads but the adaptor-dimer was not due to the high concentration of the library

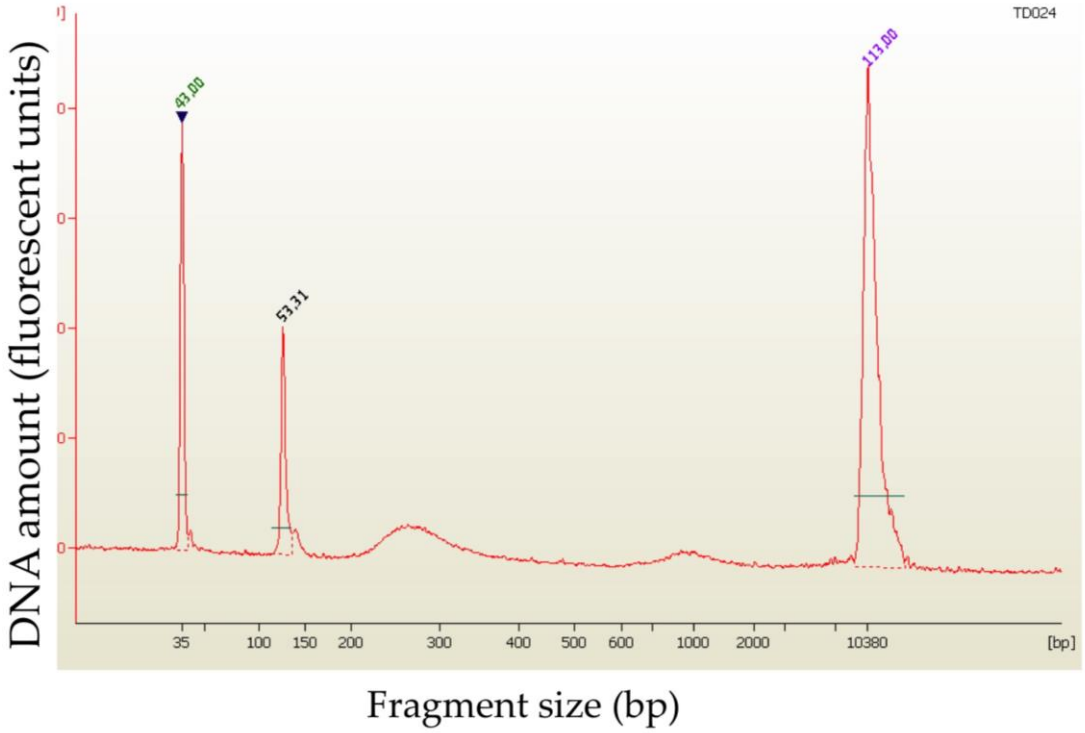



Figure 2.11: Library visualisation of a low concentration library.
Trace shows a library with a high proportion of adaptor-dimer contamination at 127bp. Adaptor-dimer was removed using size-selection prior to sequencing

Fragment size was further verified post read assembly using Geneious. Library fragment size was estimated using the mapping co-ordinates of pairs of read mates and the distribution of fragment sizes was plotted to ensure that the library was correctly identified by the Bioanalyser plots (**Figure 2.12**).

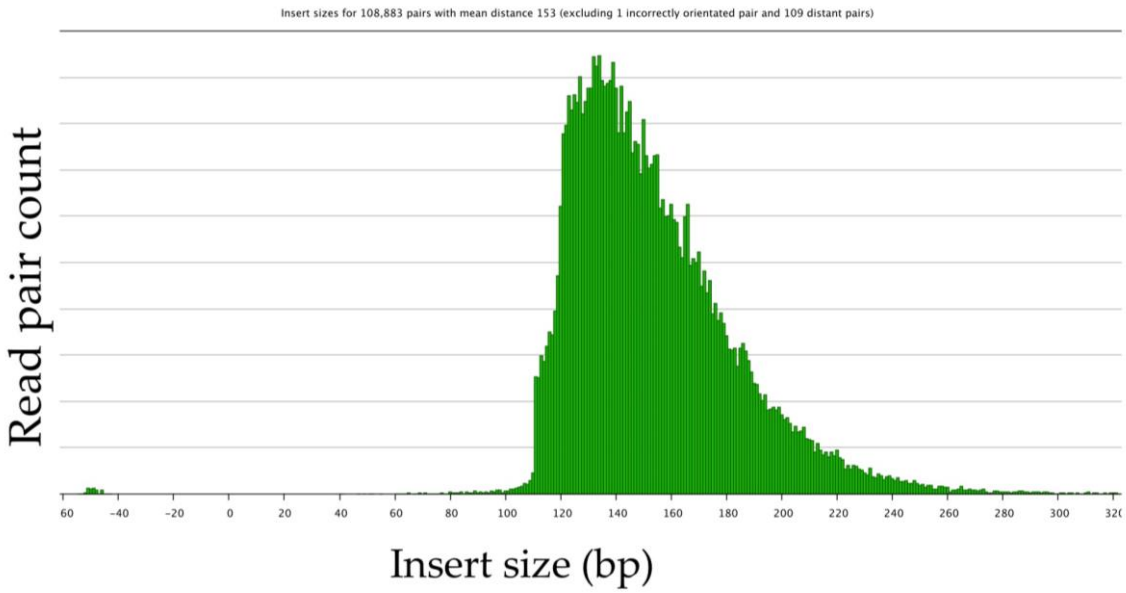



Figure 2.12: Post-assembly verification of fragment size.
Unsequenced region lengths are shown in green and are distributed tightly between 100-200bp, showing most library fragments were between 300-400bp in length

### 2.3.9 Library sequencing

The WTCHG genomics core performed library sequencing. Prepared libraries were multiplexed and sequenced on Illumina platforms. The HIV-2 reference isolates were multiplexed 2/lane and sequenced on the MiSeq platform, generating 150bp paired-end reads. Patient samples were multiplexed 6/lane and sequenced on the HiSeq 2000 platform, generating 100bp paired-end reads. Raw data were obtained as .fastq files. Quality control was conducted as described above and whole genome assemblies were generated using VICUNA and a panel of assembly algorithms, as outlined in Chapter 5.

# Chapter 3: Diversity, Evolution and Selection Pressure in the HIV-2 Accessory Gene *Vpx*

## 3.1 Introduction

*Vpx* is an SIVsmm/HIV-2 and SIVrcm/SIVmnd2 specific accessory gene[156]. The main function of *vpx* is antagonism of the host restriction factor SAMHD1, allowing infection of myeloid cells such as dendritic cells (DCs), which may be non-permissive to HIV-1 infection[178]. Several SIVs lacking *vpx* are instead able to use the neighbouring gene *vpr* to antagonise SAMHD1[202]. The absence of *vpx* or an apparent alternative SAMHD1 antagonist in HIV-1 shows that efficient degradation of SAMHD1 is not necessary for successful establishment of infection. The consequences of SAMHD1 degradation in HIV-2 infection remain unclear but it has been hypothesised that post-infection, sensing of HIV-2 by DCs leads to potent activation of innate immune responses[49]. The differences in pathogenicity between HIV-1 and HIV-2 may suggest activation of DCs is detrimental to the virus, leading to more efficient control of viral replication.

Elucidation of the structure of SIVsmm *vpx* highlighted sites of interaction between *vpx* and SAMHD1, leading to an explanation for *in vitro* mutagenesis studies that have shown site-specific mutations that impact the ability of *vpx* to antagonise SAMHD1[185]. However, there are few data on variation of the gene *in vivo*. A study by Yu *et al* showed no correlation between the ability of *vpx* alleles derived from a small number of patient samples to degrade SAMHD1 *in vitro* and clinical outcome following HIV-2 infection[223]. This study also showed that following SIVsmm *vpx* mediated infection of DCs with HIV-1 and HIV-2 there was a reduction in innate immune activation following HIV-2 infection when compared to HIV-1 infection. The ability of HIV-1 to infect DCs *in vivo* in the absence of SAMHD1 degradation remains debatable and so the consequences of apparent reduced immune activation following HIV-2 infection on disease progression are still unclear.

Yu *et al* also identified a naturally occurring substitution, K68M, in two alleles derived from the same viraemic patient that significantly reduced SAMHD1 degradation *in vitro*. Position 68 is located in the highly conserved nuclear localisation signal (NLS) and *in vitro* mutagenesis of this position back to the WT residue restored SAMHD1 degradation for one of the alleles. Reversion of an additional mutation, E15G, in the second allele restored SAMHD1 degradation, suggesting that the effects of the mutation at position 68 are somewhat context dependent. The identification of an allele derived from a viraemic patient that abrogated the ability of *vpx* to degrade SAMHD1 supports the hypothesis that infection of DCs is detrimental for HIV-2. The patients used in this study were members of the well-defined and longitudinally studied Rotterdam Cohort. A major consideration when interpreting these results is that *vpx* sequences were generated from clonal viral isolates obtained by co-culture of patient PBMCs with healthy donor PBMCs. In addition to the selection pressure introduced during clonal isolation, passage of the virus through donor PBMCs may also cause a shift in the isolated population away from the true viral quasi-species. Additionally, this study used a total of 20 sequences, generated from 11 patients, meaning that there was insufficient resolution to look at intra-patient diversity.

Whilst knowledge about the structure and mechanism of *vpx in vitro* has grown rapidly over the past few years, the genetics of *vpx in vivo* remain largely unstudied. A lack of *vpx* sequences derived from primary patient samples means that the variation, evolutionary rate and selection pressures moulding *vpx* evolution are unknown. Further study of primary *vpx* sequences is needed in order to assess whether naturally occurring *vpx* mutations may be implicated in disease progression and to answer the question of whether *vpx* is an important factor in the differences between HIV-1 and HIV-2 pathogenesis. Whether or not *vpx* is evolving at the same rate as other HIV-2 genes gives an estimation of the selection pressures acting on the gene, allowing a possible elucidation of the importance of *vpx* in overcoming host anti-viral defences.

This study therefore aimed to generate and analyse *vpx* sequences from proviral DNA in a well-characterised cohort of HIV-2 mono-infected progressors and non-progressors in order to address these questions and further investigate how the differences in *vpx* may relate to clinical outcome following HIV-2 infection.

## 3.2 Results

Laboratory methods used to generate proviral *vpx* sequence data, initial data clean up and tests of nucleotide substitution model are described in Chapter 2.2.

### 3.2.1 Cohort characteristics

Patient recruitment, stratification and clinical markers of HIV-2 disease progression are outlined in Chapter 2.1. V*px* amplification and sequencing was successful for 40 samples from 31 patients, with 8-20 clones sequenced per sample. Sequences were generated from proviral DNA, the integrated form of HIV-2. Proviral DNA can represent archived copies of the virus, which may not be functionally active[314]. However, recent work has shown that for HIV-1, RNA from plasma virions and proviral DNA in PBMCs from multiple time points from the same patient formed patient-specific populations, suggesting that proviral DNA is a good proxy for the plasma viral population[315]. In order to ensure that sequences were unlikely to represent replication-deficient viruses, sequences were removed that contained premature stop codons or frame-shift mutations anywhere in the *vpx* coding region and only sequences containing the full length and intact *vpx* ORF were included in the final analysis. Additionally the two time points chosen for longitudinal analysis were 7 years apart (2010 and 2003), making it less likely that any archived viruses from the earlier time point would persist through to the later time point.

Of the 31 patients for whom amplification was successful, clonally expanded sequences were available from 2003 and 2010 for 9 patients, from 2010 only for 15 patients and from 2003 only for 7 patients. Sequences from a single time point were generated for 10 non-viraemic patients (defined in this study as non-progressors) and 12 patients with viraemia >500 copies/mL (defined as progressors). Sequences were generated from both 2003 and 2010 for 5 non-progressors and 4 progressors. *Vpx* PCR products from a subset of 17 patients with 20 clonal sequences available were re-sequenced on the Illumina MiSeq platform, generating 2x150bp paired end

reads to allow comparison of traditional and Next-Generation Sequencing (NGS) methods.

### 3.2.2 A comparison of diversity estimates and variant calling between Sanger and Sequencing by synthesis methods

One of the major considerations when quantifying variation, diversity and evolution in viral populations using clonally expanded PCR products is the number of colonies that must be sequenced in order to gain an accurate representation of the true underlying population. There are several factors that [316]must be taken into account when deciding on a number of colonies to pick. These include power to resolve low frequency variants accurately, cost and time involved in increasing the number of colonies sequenced and an expectation of the underlying variation in the population as a whole and therefore an estimation of a lower bound frequency that is thought to be biologically informative[228].

Next generation sequencing (NGS) technologies allow the depth of sequencing and overall sequence data available for an amplified genomic region to be massively increased without an associated increase in time or expense. However, NGS data is not without its drawbacks in the context of viral evolutionary studies. Many of the most commonly used tools for analysis are not able to cope with the large amount of sequence data generated by NGS platforms due to the exponential relationship between sequence data and computing power needed. Additionally, data output in the form of short reads generated from randomly fragmented PCR products during library preparation means that the power to infer viral haplotypes can be lacking. In contrast, clonal expansion allows repeat sequencing of longer (~800bp) Sanger reads from the same DNA molecule.

In order to assess how accurately the clonal *vpx* sequences used in this analysis capture the true variation in the underlying population, a subset of samples was selected for sequencing on the Illumina MiSeq platform. Reads were mapped to the patient-specific *vpx* consensus sequence and assembled .bam files were generated for 17 samples with an average depth of coverage of 7,000x (Chapter 2.2).

Nucleotide pairwise diversity ($\pi$) was calculated for each sample using the Nei and Li method[317], where n is the number of sequences in the sample, $\chi i$ and $\chi j$ are the respective frequencies of the *i*th and *j*th sequences and $\pi ij$ is the number of nucleotide differences per site between the *i*th and *j*th sequences.

$$\pi = \sum_{ij} x_i x_j \pi_{ij} = 2 * \sum_{i=1}^{n} \sum_{j=1}^{i-1} x_i x_j \pi_{ij}$$

As diversity estimates are derived from the same underlying sample population (the *vpx* PCR product pool), estimates were compared with a 2-group paired-Wilcox Signed Rank Test (p=0.0017), which showed a significant difference between the pairwise diversity measures from Sanger and MiSeq platforms (**Table 3.1**). Correlation between the estimates of $\pi$ for each sample was also assessed (**Figure 3.1**). There was evidence of statistically significant correlation between the estimates of $\pi$ from each method ($r_s = 0.7009$, p=0.0023).

Table 3.1: Per patient estimates of nucleotide pairwise diversity.

| Sample | $\pi$ (MiSeq) | $\pi$ (Clonal) |
|---|---|---|
| TD004 | 0.00027368 | 0.0008539 |
| TD005 | 0.00033681 | 0.00169228 |
| TD013 | 0.00023086 | 9.32E-05 |
| TD020 | 0.00019192 | 0.00102469 |
| TD034 | 0.00019746 | 0.00074523 |
| TD037 | 0.00015574 | 0.0007297 |
| TD041 | 0.00048524 | 0.00380799 |
| TD046 | 0.00019361 | 0.00121099 |
| TD049 | 0.00012428 | 0.0001242 |
| TD050 | 0.00010743 | 6.21E-05 |
| TD052 | 0.00011843 | 7.76E-05 |
| TD053 | 0.00026365 | 0.00026393 |
| TD055 | 0.00030871 | 0.00094706 |
| TD058 | 0.00021209 | 0.00107126 |
| TD061 | 0.00016178 | 0.00088496 |
| TD062 | 0.00027099 | 0.0020649 |
| TD093 | 0.00014057 | 0.00018631 |

Raw estimates of $\pi$ (the average number of nucleotide differences per site for any two sequences randomly chosen from the sample population) are shown for both sequencing methods for each patient. Estimates were compared using a Wilcoxon Signed Rank Test (p=0.00167, p=0.013 (TD041 outlier excluded))

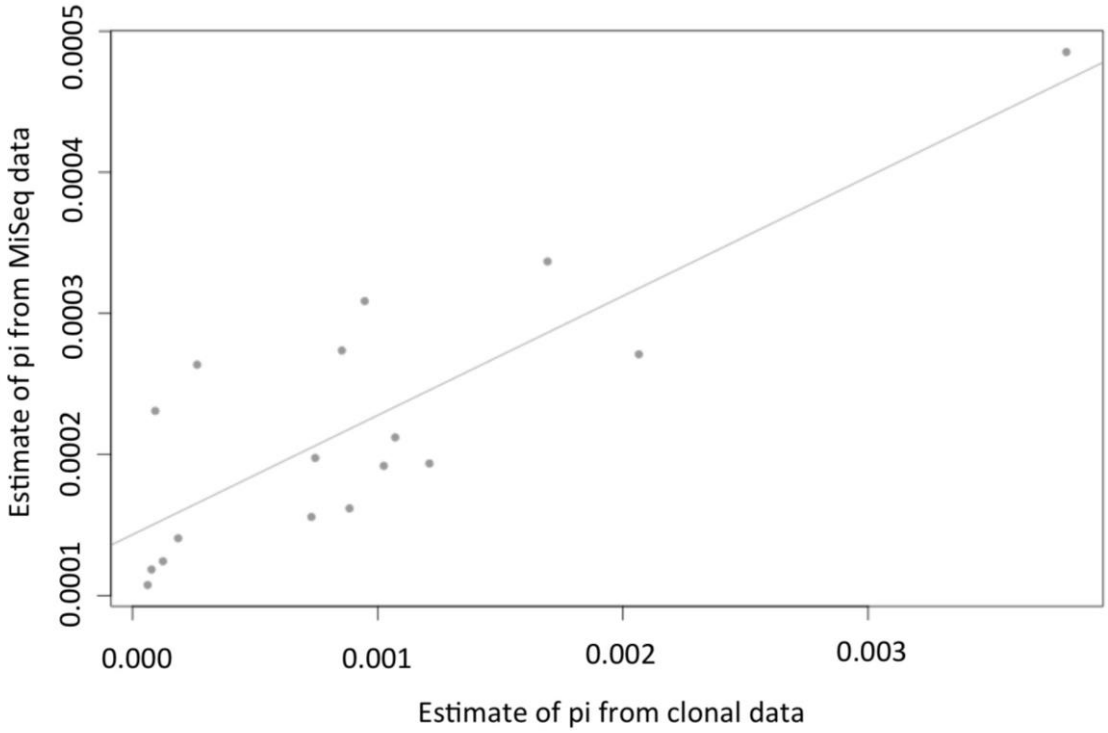



Figure 3.1: Comparison of nucleotide pairwise diversity estimates. Points per patient and linear regression are shown

Additionally, the estimates of frequency for each variant from both methods were compared (**Figure 3.2**). Computing the $R^2$ values for all data combined and per patient samples assessed the fit of data points to the expected linear regression under the null hypothesis. Overall there was a high $R^2$ value ($R^2$ = 0.889, slope=0.915; 95% CI 0.883-0.948), suggesting that the majority of the variation in estimates can be explained by fitting a linear model as the relationship (**Table 3.2**). Although the majority of patients showed a strong relationship between estimates from both methods (**Figure 3.3**), for 3 of the 17 patients the $R^2$ values were less than 0.65, suggesting that there is a large proportion of variation between the two estimates that is not expected under a linear regression model (**Table 3.2**). One of the factors that will influence the power of traditional cloning to assess variant frequency accurately is the size of the underlying viral population. As the viral load increases, the fraction of the total population that is captured in a small number of clones will decrease. The data presented in this chapter were derived from proviral DNA, and a direct comparison between proviral and plasma viral loads is not straightforward. Following HIV-1 infection, the plasma and proviral

populations are known to behave differently, although recent work has shown a steady decay in the proviral load following early treatment and long term viral control [318,319]. Additionally recent work on the latent HIV-1 reservoir has shown that the subset and number of cells harbouring provirus have an impact on the course of disease progression in HIV-1. Analysis of the proviral reservoir in different subsets of CD4+ T cells showed that CD4+ T memory stem cells show a high per-cell HIV-1 proviral load, suggesting that HIV-1 is able to exploit the longevity of these cells to establish persistent infection[320]. Analysis of the correlation between proviral load and virological control (as measured using immunological markers of disease progression and the time to plasma viral rebound following ART interruption), showed that the magnitude of the HIV-1 proviral population is predictive of disease outcome, with smaller proviral populations predicting a longer time to viral rebound following treatment interruption[321]. This observation is similar to that seen following HIV-2 infection, where control of viral replication as measured by the plasma viral load is highly predictive of the disease outcome[321]. In contrast, little is known about the dynamics of the HIV-2 reservoir. Early work suggested no correlation between plasma and proviral loads during HIV-2 infection, instead suggesting that the high proviral loads seen in HIV-2 resulted from replication of latently infected cells rather than from active HIV-2 replication[322]. However, more recent work has shown a correlation between plasma and proviral loads during HIV-2 infection[323]. In the absence of proviral load data, this study used plasma viral load as a proxy for the total HIV-2 population size, in line with the most current data on HIV-2 plasma and proviral loads. In order to determine the effect of viral population size on the accuracy of variant calls, the correlation between $R^2$ values and plasma viral load was assessed (**Figure 3.4**).

Table 3.2: Variant frequency comparisons between MiSeq and PCR cloning experiments.

| Patient | R^2 | Slope | 95% CI | p Value |
|---|---|---|---|---|
| TD004 | 0.927 | 0.992 | 0.858-1.126 | 3.15E-12 |
| TD005 | 0.91 | 0.953 | 0.842-1.064 | 3.49E-16 |
| TD013 | 0.982 | 1.017 | 0.9575-1.075 | < 2E-16 |
| TD020 | 0.997 | 1.055 | 1.017-1.094 | 1.59E-11 |
| TD034 | 0.923 | 0.983 | 0.859-1.107 | 1.22E-12 |
| TD037 | 0.552 | 0.769 | 0.360-1.178 | 0.00358 |
| TD041 | 0.961 | 0.915 | 0.840-0.989 | < 2E-16 |
| TD046 | 0.959 | 1.032 | 0.941-1.122 | 4.06E-16 |
| TD049 | 0.997 | 0.936 | 0.909-0.963 | < 2E-16 |
| TD050 | 0.997 | 0.955 | 0.927-0.983 | < 2E-16 |
| TD052 | 0.997 | 0.923 | 0.898-0.948 | < 2E-16 |
| TD053 | 0.915 | 1.034 | 1.040-1.305 | <2e-16 |
| TD055 | 0.647 | 0.59 | 0.429-0.751 | 8.45E-08 |
| TD058 | 0.997 | 0.938 | 0.917-0.960 | < 2E-16 |
| TD061 | 0.392 | 0.66 | 0.300-1.021 | 0.00182 |
| TD062 | 0.928 | 0.852 | 0.775-0.929 | < 2e-16 |
| TD093 | 0.99 | 0.959 | 0.914-1.004 | <2e-16 |
| Overall | 0.889 | 0.915 | 0.883-0.948 | < 2e-16 |

The spread of variant frequency estimates was measured using $R^2$.

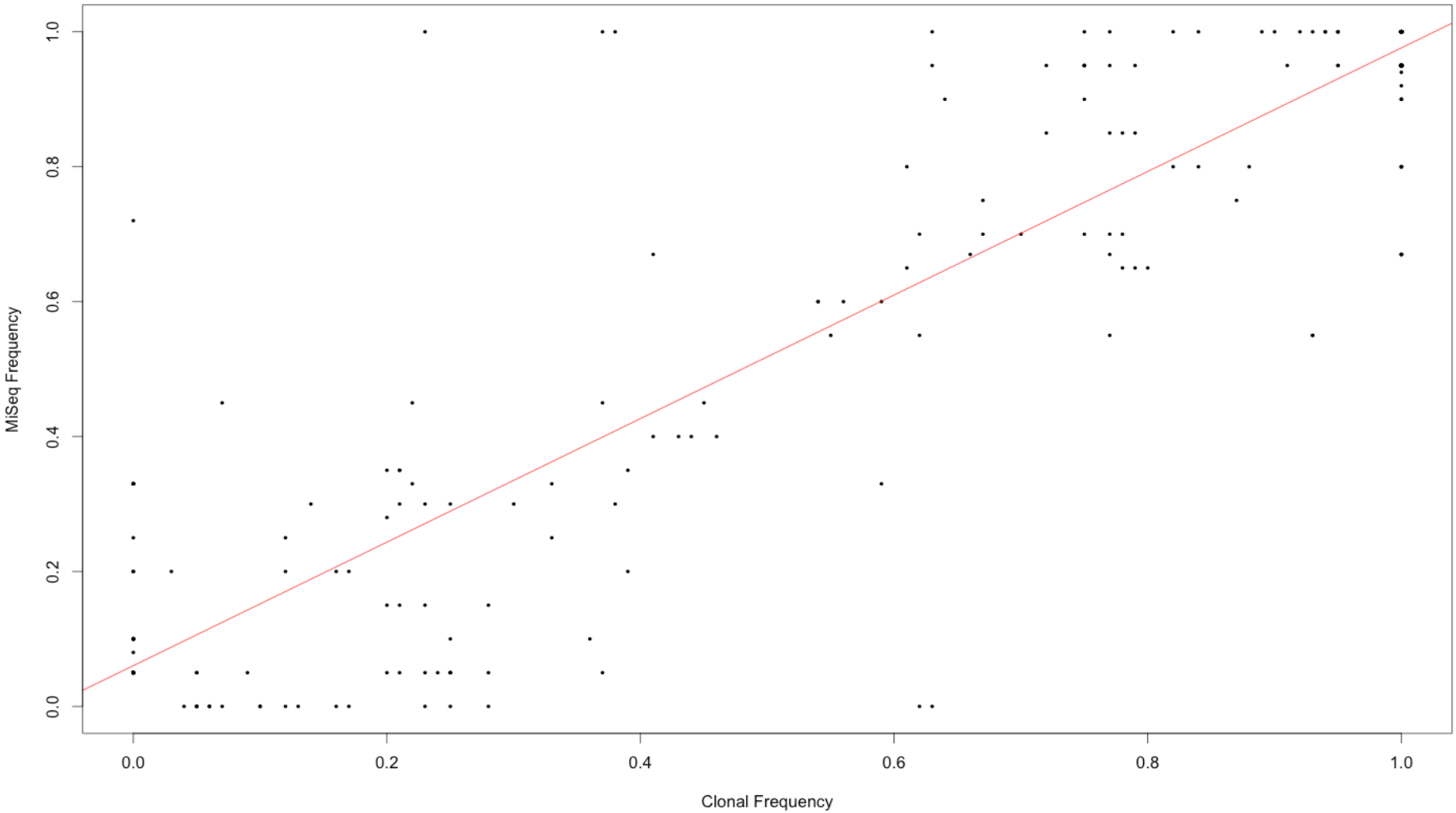



Figure 3.2: Comparison of sequence variant quantification by Illumia MiSeq deep sequencing and PCR cloning/Sanger sequencing. Frequency estimates from the two methods are plotted for each variant in 17 patients. Linear regression line is shown in red ($R^2$ =0.889)

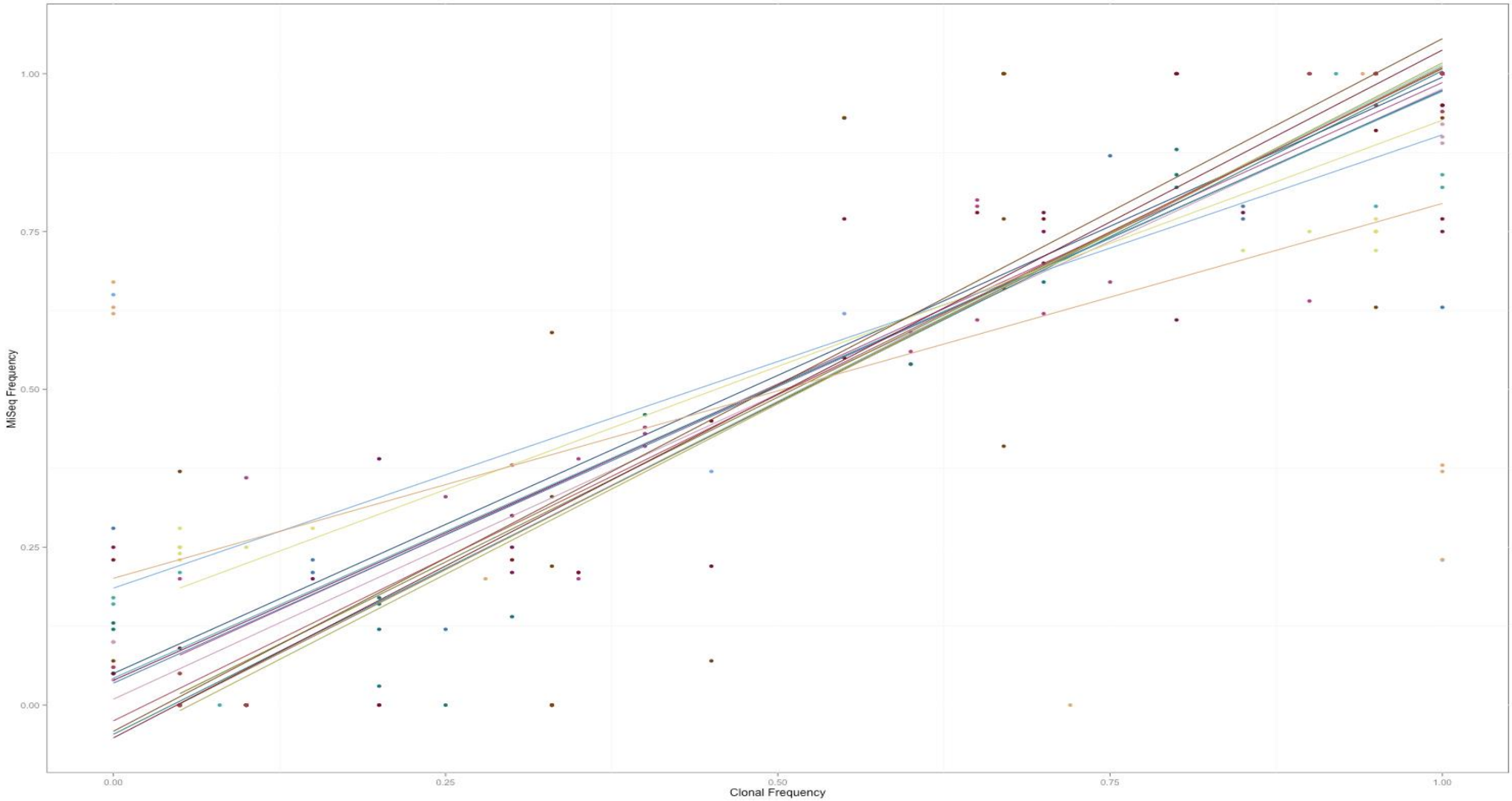



Figure 3.3: Comparison of sequence variant quantification by Illumia MiSeq deep sequencing and PCR cloning/Sanger sequencing. Frequency estimates and linear regression lines are overlaid and coloured by patient.

There was a significant negative correlation between $R^2$ and log plasma viral load ($r_s$ = -0.76, p= 0.0003), suggesting that the discrepancy in frequency estimates between the two methods may be a factor of the underlying virus population size.

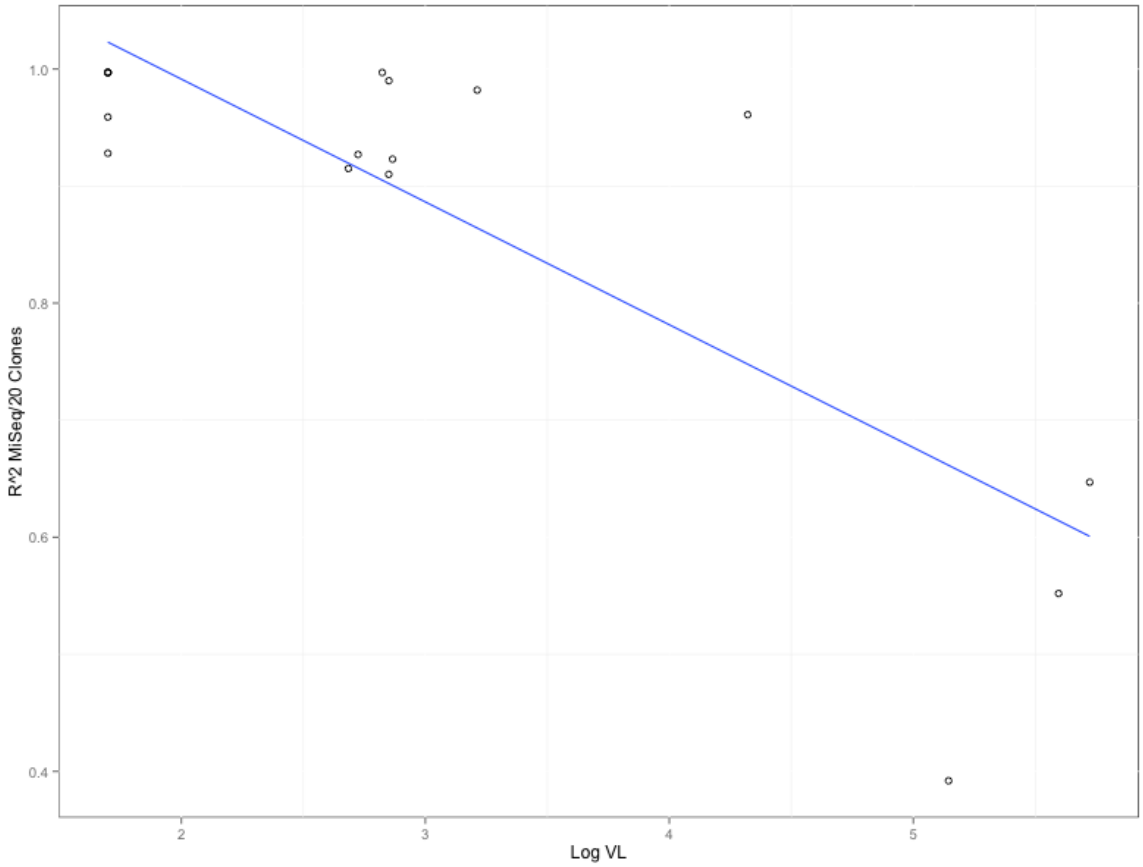



Figure 3.4: Relationship between $R^2$ and log plasma viral load.
$R^2$ fit between MiSeq and PCR cloning experiments was plotted against the log of the plasma viral load. ($r_s$ = -0.76, p= 0.0003)

Overall, these analyses highlighted several significant differences between the estimates of diversity and variant frequencies when calculated using MiSeq or clonal data. There was a significant difference between the magnitude of estimates of nucleotide pairwise diversity ($\pi$). However, there was a significant correlation between the estimates, showing that between patient comparisons using the same method are still valid. Comparing the fit to a linear regression model assessed variant frequencies estimated from both data sets. The slope of the line (0.95) suggests that variants are called at a slightly lower frequency using MiSeq than clonal data. This is likely to be a function of the sample size of the two data sets. All samples had 20 clonal sequences, leading to a lower bound of f=0.05 for minor

variants when using clonal data and all variants frequencies were assigned to bins of 0.05. In the MiSeq data the depth of coverage (analogous to sample size) was 7,000x, allowing more accurate representation of variant frequencies. The step-wise nature of the variant frequencies called by data from 20 clones also explains the variation in $\pi$, as assigning frequencies to bins is likely to inflate the values for rare variants, which are more numerous than variants at higher frequencies, leading to an overall increase in the estimate of diversity. We also showed that viral load is an important factor to consider when deciding how many clones to sequence. There was a negative correlation between log viral load and the proportion of variation explained by a linear model. Therefore, large viral populations may lead to reduced power to determine variant frequencies accurately when using small numbers of clonally derived sequences. However, the overall $R^2$ of 0.889 between the two methods showed that sequencing 20 clones gives good resolution to estimate variant frequencies in the underlying population and therefore, clonal sequence data could be used with confidence in further analyses in this chapter. However, this study suggests a reduction in the accuracy of frequency estimates with increasing (log) viral loads, which must therefore be considered when selecting a sequencing method to look at viral evolution. This is even more important in the context of HIV-1 evolution, where log plasma loads are normally one or two orders of magnitude greater than those seen in HIV-2 infection[110].

### 3.2.3 Exclusion of PCR contamination and sample mix up

In order to exclude the possibility of sample mix up during the retrieval of historical samples or contamination during PCR amplification, a series of phylogenetic analyses were carried out. To exclude contamination during PCR amplification, a maximum likelihood (ML) phylogenetic tree was generated in FastTree under a GTR+I+G model of nucleotide substitution, based on an alignment of all clonal sequences from all patients[324]. For patients with sequences from 2003 and 2010, sequences from both time points were included. Statistical support for each node in the tree was estimated using the Shimodaira-Hasegawa (SH) test, providing 1000 resamples of each split (**Figure 3.5**)[325]. Monophyletic clusters were observed for all patients except for three. Sample TD046 showed separate monophyletic clusters for samples from 2010 and 2003, which is indicative of intra-patient evolution. There is low statistical support for the individual TD046 clusters (75% and 31%) when compared to the support for the monophyletic clustering of samples from TD042 (100%). Samples TD035 and TD005 were both split over two subclusters (labelled 'A' and 'B') with sequences from other patients separating the two. This is not likely to be indicative of PCR contamination as the sequences branching from within the radiation of TD005 and TD035 form patient-specific monophyletic clusters with high statistical support (93%-99%). These multi-patient clusters are most likely to represent transmission clusters, as inter-cohort transmission has previously been described in Caió[290].

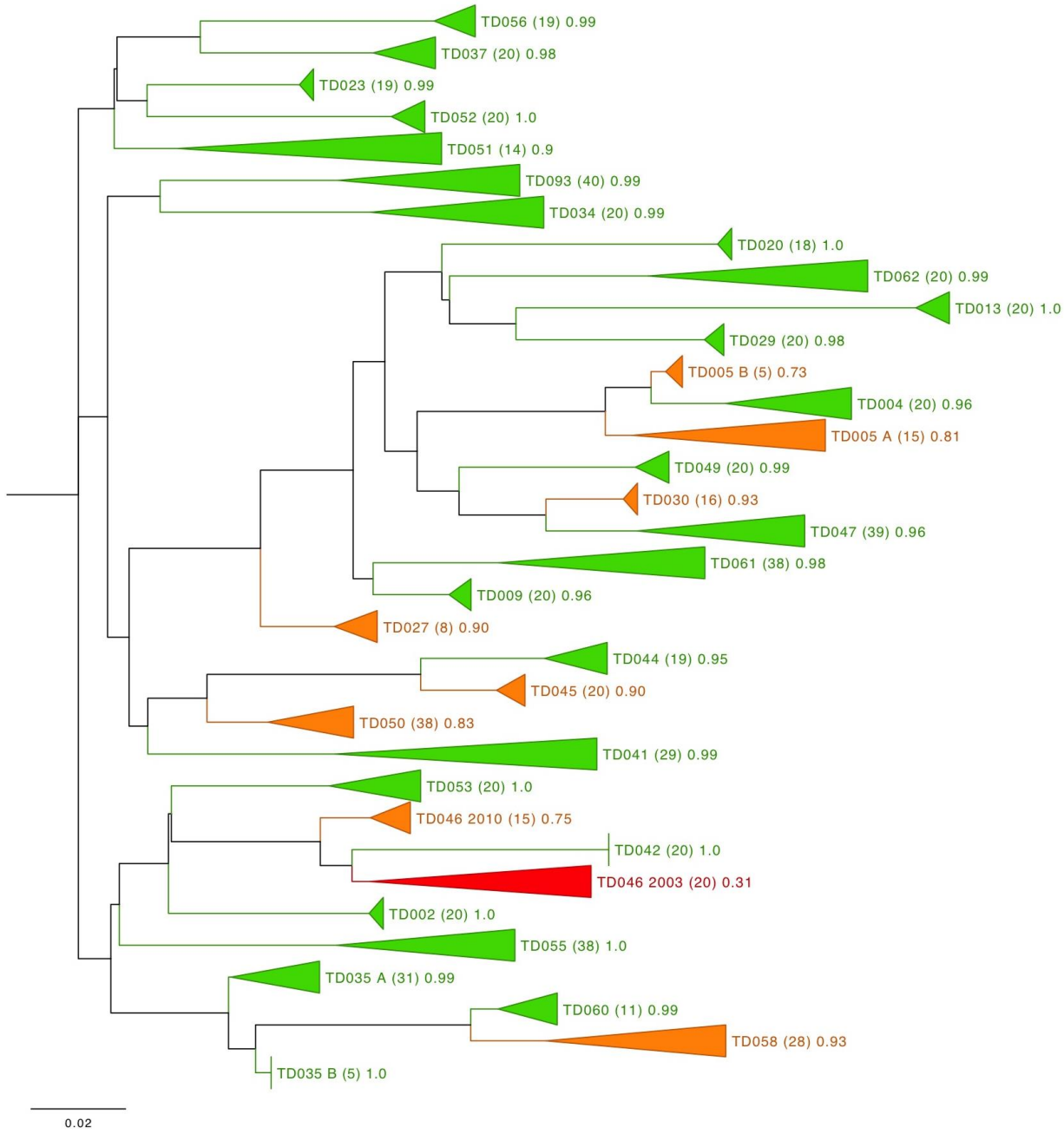



Figure 3.5: Midpoint rooted ML phylogeny of *vpx* sequences.
Tips are annotated with patient ID, the number of sequences comprising the collapsed cluster and support for the cluster estimated under the SH method. Branches with support values >0.95 are shown in green, >0.75 in orange and ≤0.75 in red.

In order to verify that 2003 samples retrieved from the Caió BioBank have been correctly linked to those collected in 2010, a phylogenetic analysis was carried out on sequences from nine patients with samples from both 2010 and 2003 (**Figure 3.6**). An ML phylogeny was created in FastTree under a GTR+I+G model of nucleotide substitution using an alignment of all longitudinal sequences. Statistical

support for the topology was estimated using the SH test. Monophyletic clusters were observed for all patients with robust statistical support (>85%) for all patients except one. The statistical support for the cluster containing sequences from patient TD035 was low (26%), however, the support for the node separating these sequences from the nearest neighbour (TD058) was high (89%), providing strong statistical support that these two groups form distinct populations. Therefore, this analysis shows that all samples from 2003 have been correctly identified and matched to samples from 2010.

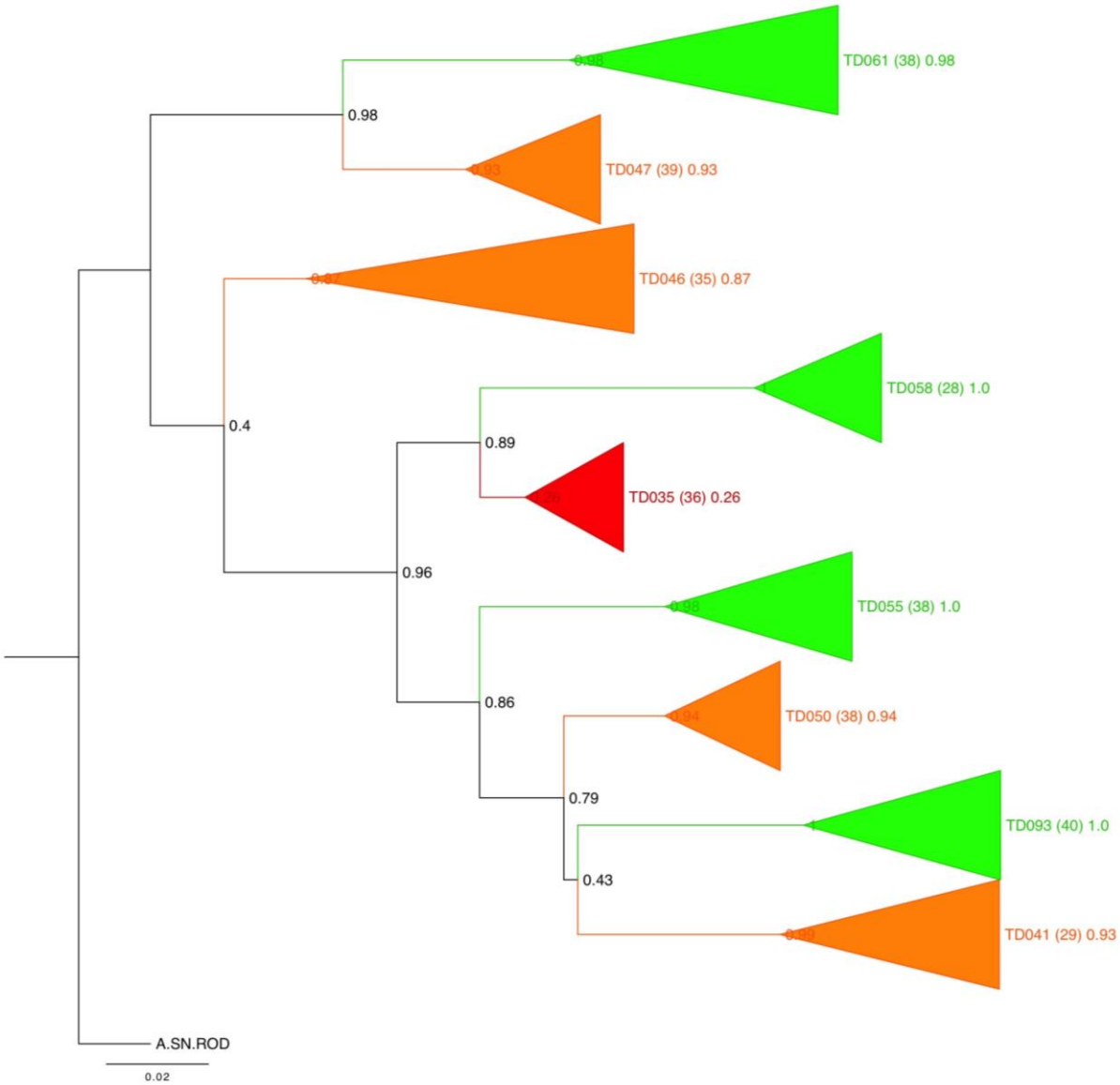



Figure 3.6: Midpoint rooted ML phylogeny of longitudinal *vpx* sequences. Tips are annotated with patient ID, the number of sequences comprising the collapsed cluster and support for the cluster estimated under the SH method. Branches with support values >0.95 are shown in green, >0.75 in orange and ≤0.75 in red.

### 3.2.4 HIV-2 group analysis

Previous analyses of *nef*, *env* and *gag* sequences from the patients involved in this study have shown that all are infected with HIV-2 group A[326]. This is unsurprising as HIV-2 A is the major form of the virus in Guinea-Bissau, with the majority of HIV-2 B infections seen further East, in Côte d'Ivoire and Ghana[327]. A Bayesian phylogenetic analysis was carried out to verify that all *vpx* sequences generated for this study are from HIV-2 group A and that no patients represent previously misidentified alternative HIV-2 types (**Figure 3.7**).

A panel of reference sequences was downloaded from the LANL HIV database, representing 20 group A *vpx* sequences, 4 group B *vpx* sequences and 4 *vpx* sequences from the CRF01_AB recombinant form[24]. SIVmac was included as an outgroup, representing an independent zoonotic transmission of SIVsmm (**Table 3.3**).

Patient clonal sequences were aligned using MUSCLE implemented in Geneious 6.2 and a single consensus sequence was generated per patient for use in the subtype analysis[328]. Where there were multiple time points for a single patient, a separate consensus was created for each time point. An alignment of patient and reference panel sequences was created using ClustalW[329]. Phylogenetic relationships were inferred by Bayesian Markov Chain Monte Carlo (MCMC) inference under the GTR+I+G model of nucleotide substitution, with a prior distribution under a constant population size (Kingsman's Coalescent) and constant molecular clock (clock rate prior was normally distributed with a mean of 0 and a standard deviation of 0.1). MCMC estimates were generated using BEAST v1.8.0; the MCMC search was performed using 100,000,000 iterations with trees sampled every 10,000 generations[330]. Convergence of posterior probabilities was assessed by calculating the effective samples size (ESS) using Tracer v1.6 and convergence was assumed once the ESS for all parameter estimates was >200. TreeAnnotator v1.8.0 was used to select a maximum clade credibility tree (MCCT) from the sampled posterior distribution, following a burn-in of 1000 trees

(equivalent to 10% of the posterior distribution)[331]. The MCCT was visualised and edited using FigTree v1.4.1[332].

All *vpx* sequences generated in this study clustered with reference HIV-2 group A sequences with high support (posterior probability of the HIV-2 A node >0.9). CRF01_AB sequences clustered with HIV-2 group B sequences with high support, suggesting the recombinant has a 'B-like' *vpx*. Therefore it is unlikely that any of the sequences generated in this study were derived from HIV-2 group B or CRF01_AB viruses.

The MCCT was also used to assess the relationship between HIV-2 group A *vpx* sequences. The HIV-2 group A sequences formed two distinct clusters. The major HIV-2 A cluster contained all sequences generated in this study, all of the sequences from Guinea-Bissau and all but one of the sequences from the neighbouring countries of The Gambia and Senegal. The second cluster contained sequences from Côte d'Ivoire and Ghana. Non-African samples were split over both clusters, probably representing the different geographic locations of these infections. HIV-2 transmission is relatively rare outside Africa and the majority of reported cases represent infections that were not contracted in the country of diagnosis. The geographical split of the countries in the two clusters recapitulates the known history of HIV-2 group A, where both Guinea-Bissau and Côte d'Ivoire acted as hubs early in the epidemic, leading to distinct viral populations in each geographic region. *Vpx* sequences generated in this study were dispersed amongst the major HIV-2 A cluster, showing that infections in Caió are not the result of a narrow source outbreak and therefore are representative of the radiation of HIV-2 A in Guinea-Bissau and neighbouring countries. Additionally, there was no obvious clustering of sequences generated from progressors or non-progressors, implying that the differences in disease progression are not simply due to intra-cohort transmission of a defective HIV-2 A virus.

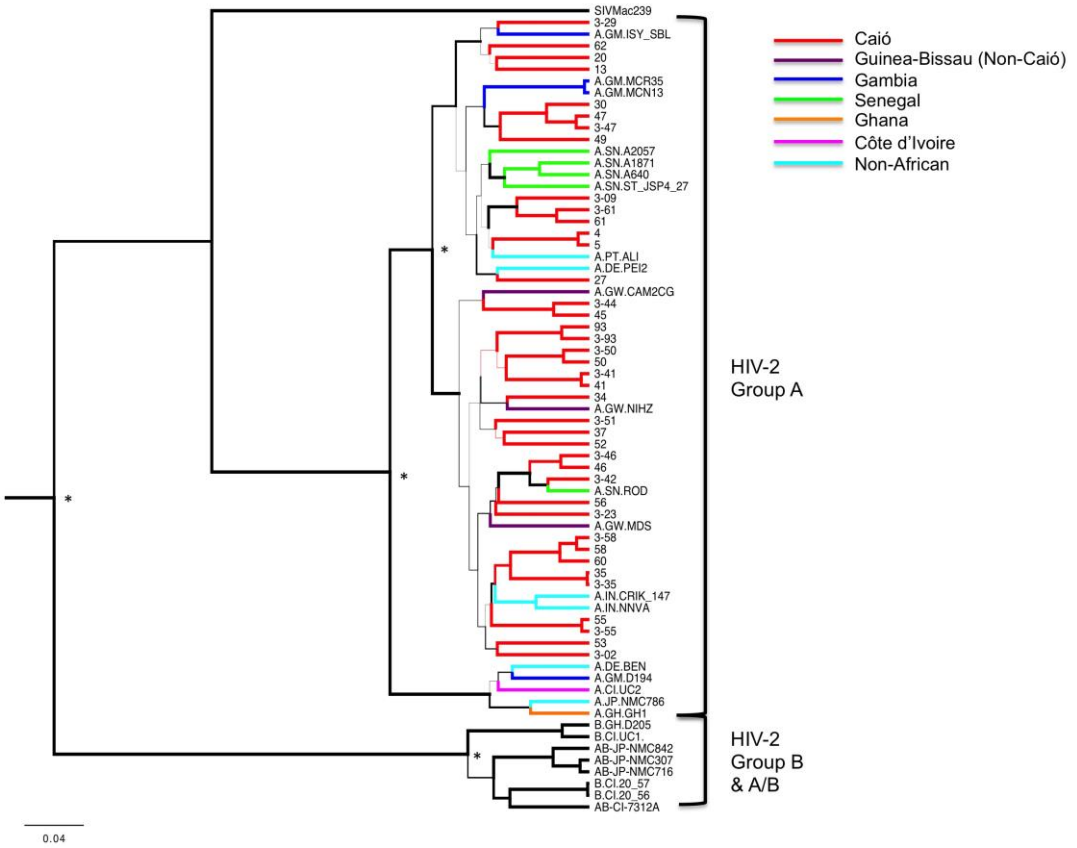



Figure 3.7: Midpoint rooted MCCT of reference and patient *vpx* sequences. Branches are weighted by posterior probability, internal nodes with a posterior probability of >0.9 are denoted by * and group A sequences are coloured by country of origin.

Table 3.3: Reference sequences used for viral subtyping.

| Name | Group | Country of Origin | Accession Number |
|---|---|---|---|
| ST_JSP4_27 | A | Senegal | M31113 |
| A640 | A | Senegal | U81837 |
| A1871 | A | Senegal | U81839 |
| PEI2 | A | Germany | U22047 |
| CAM2CG | A | Guinea-Bissau | D00835 |
| CRIK_147 | A | India | DQ307022 |
| NNVA | A | India | EU980602 |
| ROD | A | Senegal | BD413542 |
| MDS | A | Guinea-Bissau | Z48731 |
| NIHZ | A | Guinea-Bissau | J03654 |
| MCN13 | A | Gambia | AY509259 |
| MCR35 | A | Gambia | AY509260 |
| ISY_SBL | A | Gambia | J04498 |
| A2057 | A | Senegal | U81843 |
| ALI | A | Portugal | AF082339 |
| NMC786 | A | Japan | AB731742 |
| GH1 | A | Ghana | E02138 |
| BEN | A | Germany | M30502 |
| D194 | A | Gambia | A09995 |
| UC2 | A | Côte d'Ivoire | U38293 |
| 20_56 | B | Côte d'Ivoire | AB485670 |
| 20_57 | B | Côte d'Ivoire | AB485671 |
| UC1 | B | Côte d'Ivoire | L07625 |
| D205 | B | Ghana | X61240 |
| 7312A | AB | Côte d'Ivoire | L36874 |
| NMC716 | AB | Japan | AB499694 |
| NMC307 | AB | Japan | AB731738 |
| NMC842 | AB | Japan | AB499695 |
| SIVMac239 | SIVMac | U.S.A. | M33262 |

Group, country of origin and accession numbers are shown.

### 3.2.5 Genetic variation in *vpx*

In order to classify variation in existing *vpx* sequences, a subset of reference sequences was accessed via the LANL HIV database that represented all publically available HIV-2 group A *vpx* sequences (n=20). Variant frequencies were summarised in a folded-site frequency spectrum (**Figure 3.8A**). The spectrum was folded in the absence of a robust HIV-2 group A *vpx* ancestral sequence. Variant frequencies in the study cohort were estimated using all clonal data rather than consensus patient sequences. This allows inter-patient, as well as population level variation to be quantified. In order to prevent biasing from multiple sampling of the same patient, only one time point per patient was included in the analysis and for patients with multiple time points, 2010 was chosen arbitrarily (**Figure 3.8B**).

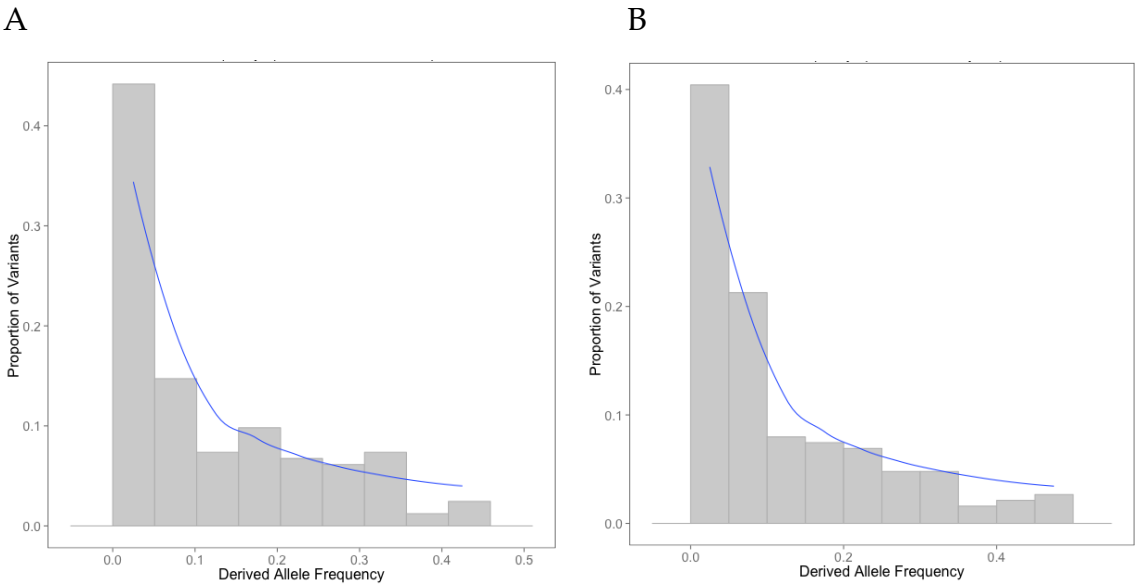



Figure 3.8: Folded site frequency spectra for *vpx*.
Allele frequency distributions are shown for LANL sequences (A) and study sequences (B). Expected spectrum shape under a neutral model is shown in blue

*Vpx* sequences from both the LANL database and the study population show an excess of rare variants when compared to the expectation under a neutral model. An excess of rare variants is commonly attributed to either purifying selection or a recent population expansion[333,334]. In the case of HIV-2, both factors could contribute to the observed excess of rare mutations. Previous studies have demonstrated that HIV-2 is under purifying selection in the majority of patients[231]. Additionally, Bayesian skyline analyses of the population history of HIV-2 in West

Africa suggest that HIV-2 underwent an exponential population growth between 1955-1970, although the rate of new infections and therefore the effective population size ($N_e$) has been falling since 1970[113]. The population dynamics of acute HIV-2 infection remain largely unstudied and it is not known whether HIV-2 infection is the result of one or multiple transmitted founder viruses. However, it is reasonable to assume that there is a significant growth in population size following initial infection, as is seen in HIV-1 and therefore the site frequency spectrum from patient data could reflect this population expansion in early infection. The similarity between the spectra generated for database and study samples suggests that the selective pressures shaping the variant distribution in the study population may be similar to those acting on HIV-2 group A at a population level and therefore conclusions drawn from this study may be applicable to HIV-2 group A *vpx* more generally.

### 3.2.6 Identification of non-synonymous variation in *vpx*

In order to classify non-synonymous variation in existing *vpx* sequences, nucleotide sequences accessed from LANL database were translated into amino acid sequences using the human genetic code and aligned using Blosum62[335]. An HIV-2 group A *vpx* consensus protein sequence was generated, with each position defined by the most common residue in the alignment (**Figure 3.9**).

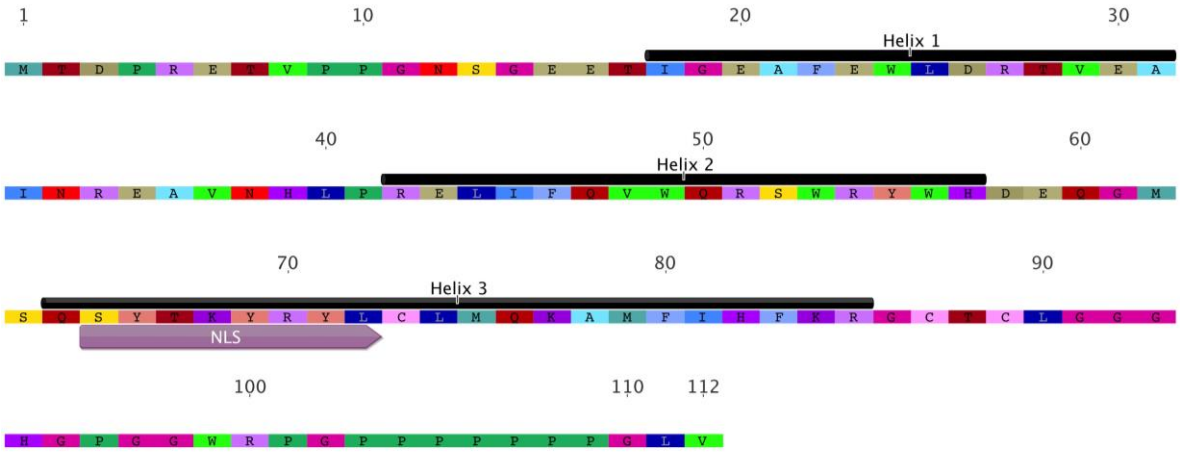



Figure 3.9: *Vpx* amino acid consensus sequence.
Consensus was generated from 20 HIV-2 group A sequences accessed from the LANL HIV database. Locations of the three $\alpha$-helical domains and nuclear localisation signal are shown

Non-synonymous variants were called if they differed from the consensus. A total of 75 variants were seen in the database sequences, found at 45 different positions in *vpx*. The total length of *vpx* is 112 amino acids and therefore approximately 40% of sites in *vpx* showed variability and 60% were invariant. Sequences deposited in LANL database are taken from multiple sampling locations and time points and so may be considered to be a good representation of the total variation in HIV-2 A *vpx*. However, there are considerably fewer *vpx* sequences than are available for other genes of HIV-2 A (*gag* ≈300, *env*≈400) and the majority of those available have been sampled from viruses isolated and subsequently cultured *in vitro*, rather than derived directly from primary patient samples. Non-synonymous variant frequencies in the study cohort were estimated using all clonal data rather than consensus patient sequences. Each site in the amino acid sequence was analysed for variants and the frequency was calculated per patient for each variant. In total, 49 non-synonymous mutations were identified at a frequency of >0.01 at 33 sites.

In order to compare observed variation in *vpx* with previously published sequences, a consensus sequence was generated for each patient using total clonal data and consensus sequences were aligned to HIV-2 reference sequences and clonal data from Yu *et al* (**Figure 3.10**). Variation in the study cohort was similar to previous studies, with helix 3 showing the most variation and helix 2 the least. There were no mutations observed in the poly-proline motif, highlighting the high conservation in this region, possibly linked to a strong functional constraint. Several novel mutations were observed in the NLS, including an L to V mutation of the anchor residue in patient TD030. The location and nature of these variants will be discussed in further detail later in this chapter.

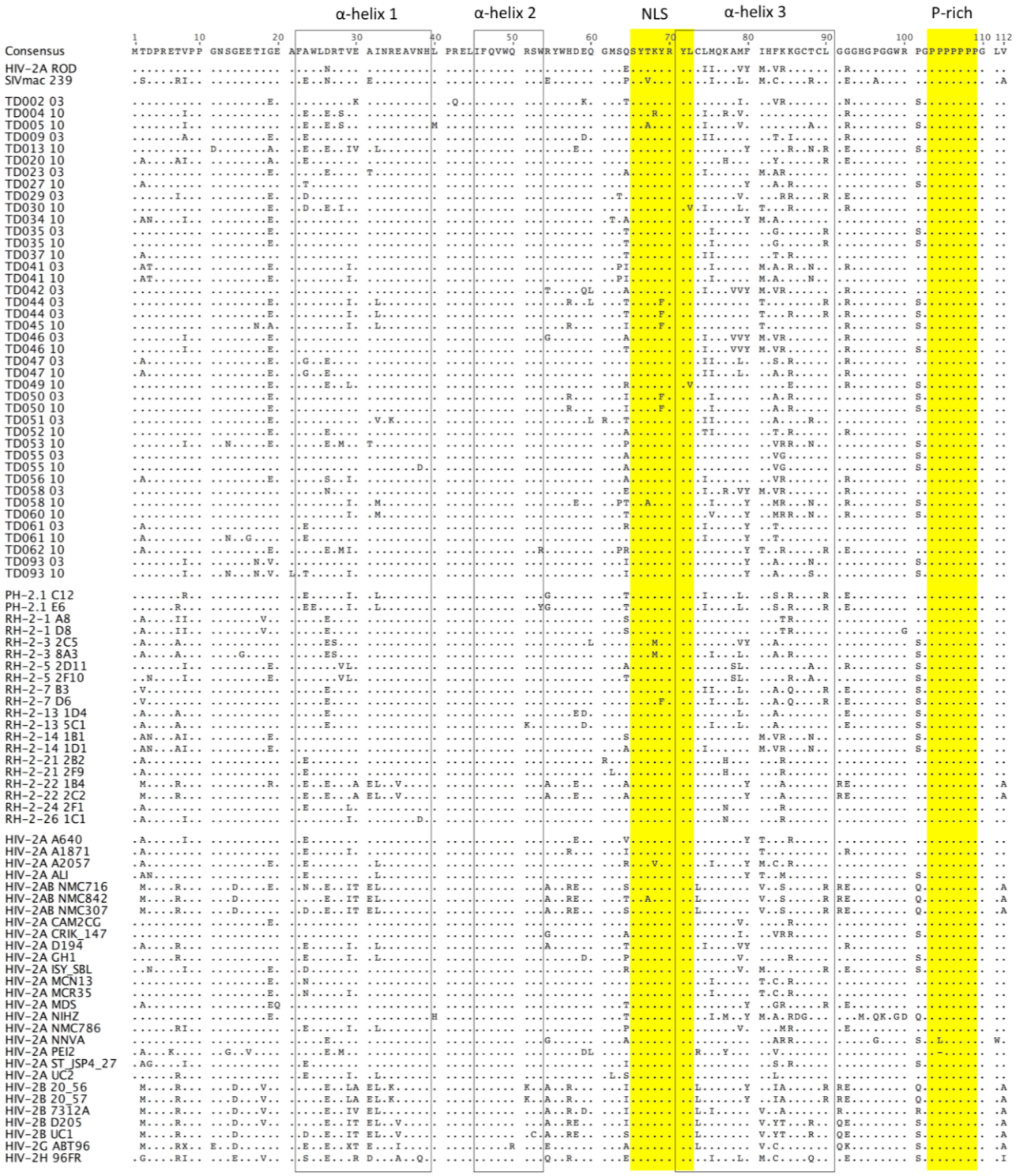



Figure 3.10: Alignment of HIV-2 vpx sequences.
The HIV-2 ROD sequence is shown above for comparison. The three predicted alpha-helical domains are highlighted alongside the nuclear localisation signal and poly-proline motif. Study samples are annotated with patient ID and year of sampling. Dots indicate amino acid identity.

### 3.2.7 Genomic location of non-synonymous variants:

In addition to estimating the nature and the frequency of the variants in HIV-2 A *vpx*, it is also informative to look at which genomic region they are in. Recent work has shown that there are three helical domains and a nuclear localisation signal (NLS) in *vpx* (**Figure 3.9**). These are helix 1 (position 18-37), helix 2 (position 42-57), helix 3 (position 64-85) and NLS ('SYTKYRYL' at position 65-72)[185]. The number and proportion of variant sites in each helical domain and the NLS were calculated, in addition to the proportion of variant sites in non-helical domain sites and over the whole gene for both database and study sequences (**Figure 3.12**). In addition to the major structural and functional domains of *vpx*, variation was also assessed over the known sites of interaction with DCAF1 and SAMHD1. 35 sites have been shown to be involved in these interactions (**Figure 3.11**, summarised in Chapter 1.7).

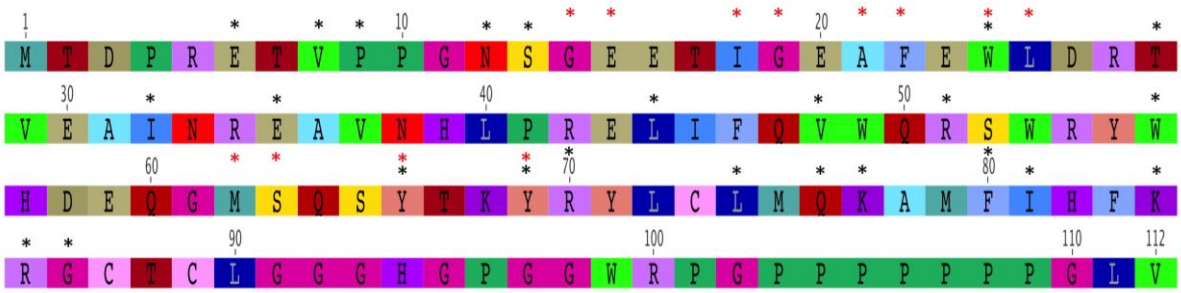



Figure 3.11: Sites of interaction in *vpx*.
Residues involved in interaction with SAMHD1 (shown in red) or DCAF1 (shown in black) in *vpx*

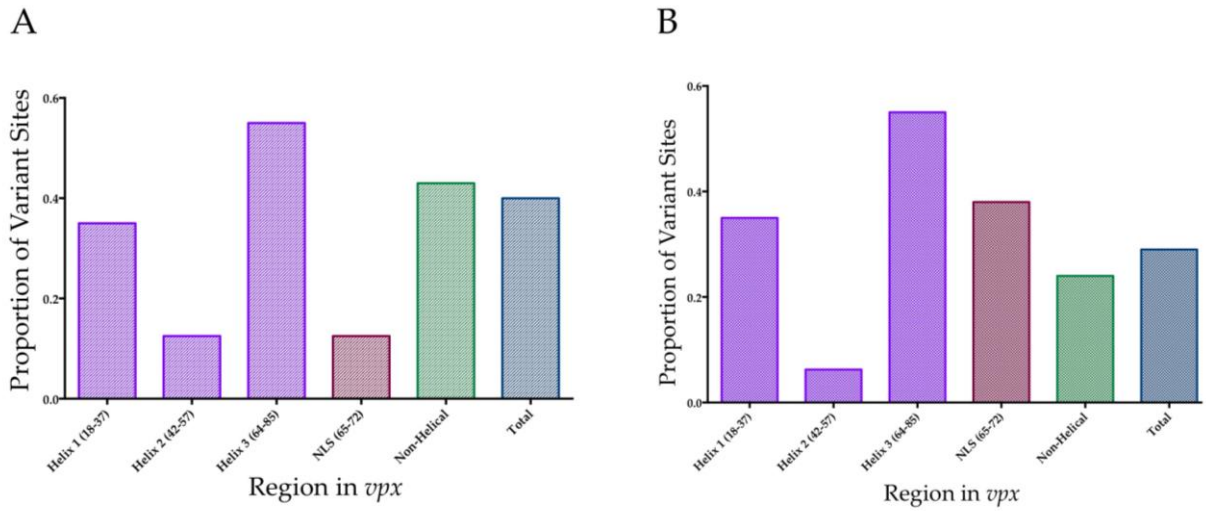



Figure 3.12: Proportion of sites showing non-synonymous variation in *vpx*. Proportions are shown for LANL (A) and study population (B) and are partitioned according to genomic region in *vpx*

Analysis of the LANL database sequences showed the proportion of variant sites over the whole gene is 0.4 (45/112) and is similar to that seen in non-helical sites 0.43 (23/54), helix 1 0.35 (7/20) and helix 3 0.55 (12/22). Helix 2 and the NLS have notably lower proportions of variant sites; both helix 2 and the NLS have 0.13 variant sites (helix 2 =2/16 and NLS = 1/8), indicative of sequence conservation that may result from functional constraint. For the study population, the proportion of variant sites over the whole gene is 0.29 (33/112), which is lower than the proportion seen in the reference sequences. This is probably due to the distribution of samples in both data sets. Whilst the reference panel includes sequences from multiple geographical locations sampled over many decades, the study population represents only two time points and all samples were collected in the same location. However, HIV-2 infections in Caió do not represent a narrow source outbreak and therefore variation seen in this study population is informative for overall HIV-2 group A genetic heterogeneity. Additionally, reference sequences may harbour mutations that have been introduced through artificial selection pressures during *in vitro* isolation and culture, leading to an over-estimation of the true variation. The proportion of variant sites in non-helical regions is 0.24 (13/54), which is similar to the overall proportion of variant sites, an observation that is also made in the reference sequences. The distribution of variant sites over the

three $\alpha$-helical domains is also similar to that seen in the database. Helix 1 and helix 3 show a higher proportion of variant sites, helix 1 has 0.35 (7/20) and helix 3 has 0.54 (12/22). Helix 2 has 0.063 variant sites (1/16), which is lower than seen in the database but in line with the previous observation that helix 2 appears to be more conserved than helix 1 or helix 3. The NLS showed much greater variation in Caió with 0.38 (3/8) sites showing variation. This suggests that in natural populations, the NLS is less conserved than would be predicted from previous sequence data.

When sites involved in interaction with SAMHD1 or DCAF1 were considered, the database sequences showed conservation at 22/35 sites, implying variation might be tolerated at some of the interaction sites. For the majority of sequences in LANL, there are few data on viral phenotype or clinical symptoms of HIV-2 infection and it would be informative to assess the variation at these sites in primary patient samples for which progression data are available. Additionally, the two mutations identified in Yu *et al* that affected SAMHD1 degradation, K68M and E15G, were investigated. K68M was not seen in the database, however, an alternative substitution, K68V, was identified at position 68 in one sequence. E15G was also not seen in the database sequences and no alternative substitutions at position 15 were identified. Yu *et al* also identified a second mutation in the NLS. Y69F was identified in one patient and the ability of *vpx* alleles harbouring Y69F to degrade SAMHD1 was not significantly affected by the presence of this mutation. However, Yu *et al* postulated that its location in the NLS might make it functionally important. Y69F was not seen in the reference panel and no alternative mutations were seen at position 69 or at any other position in the NLS.

Non-synonymous variation at sites of known interaction with SAMHD1 or DCAF1 was also assessed in the study population. Conservation was seen at 22/35 sites, which is the same as the reference panel. However, the interaction sites showing variation were not the same in both populations. In Caió, variation was seen at sites 62, 63 and 69, which were conserved in the reference population and are involved in SAMHD1 interaction. Site 69 is located in the NLS and was shown

to be variable in the Yu *et al* study. The mutation seen at this site in the study population, Y69F, was the same substitution that Yu *et al* observed and was seen at a frequency of around 0.1 in both studies. *In vitro* studies of *vpx* alleles with Y69F showed no effect on SAMHD1 degradation and so the consequences of this substitution remain unclear. Sites 6, 77 and 86 are involved in DCAF1 interaction and showed variation in the reference sequences but were conserved in the study population. The previously identified mutation K68M that is associated with reduced SAMHD1 degradation was not seen in the study population and no alternative mutations at site 68 were seen. This raises questions over whether this is a naturally occurring mutation or an artefact of *in vitro* viral propagation.

### 3.2.8 Comparison of non-synonymous variants in *vpx* between the study population and LANL HIV database

Coding variant frequencies were compared between the study population and reference panel in order to determine how well the reference panel predicts variation in *vpx* sequences derived from primary patient samples. Variants identified in the reference panel were stratified into two groups, major variants were those seen at a frequency of 0.15 or higher (observed in 3 or more sequences) and minor variants were at a frequency of 0.1 or 0.05 (observed in 1 or 2 sequences only). Initially the frequency of each major variant in both the study population (comprising all clonal sequences with a single time point per patient) and the reference panel was calculated. Median frequencies were compared with a non-parametric test and 6 variants were seen at significantly different frequencies in the two populations (**Figure 3.13**). The most highly significant was an isoleucine to threonine substitution at position 81. This was at a frequency of 0.5 in the database, whereas it was only seen at a frequency of 0.16 in the study population. A second mutation, F83C was seen in the reference panel at 0.2 but not at all in the study population. Additionally three variants were identified at a frequency >0.5 in the study population, resulting in a different 'wild-type' (WT) amino acid at this position when the two groups are compared.

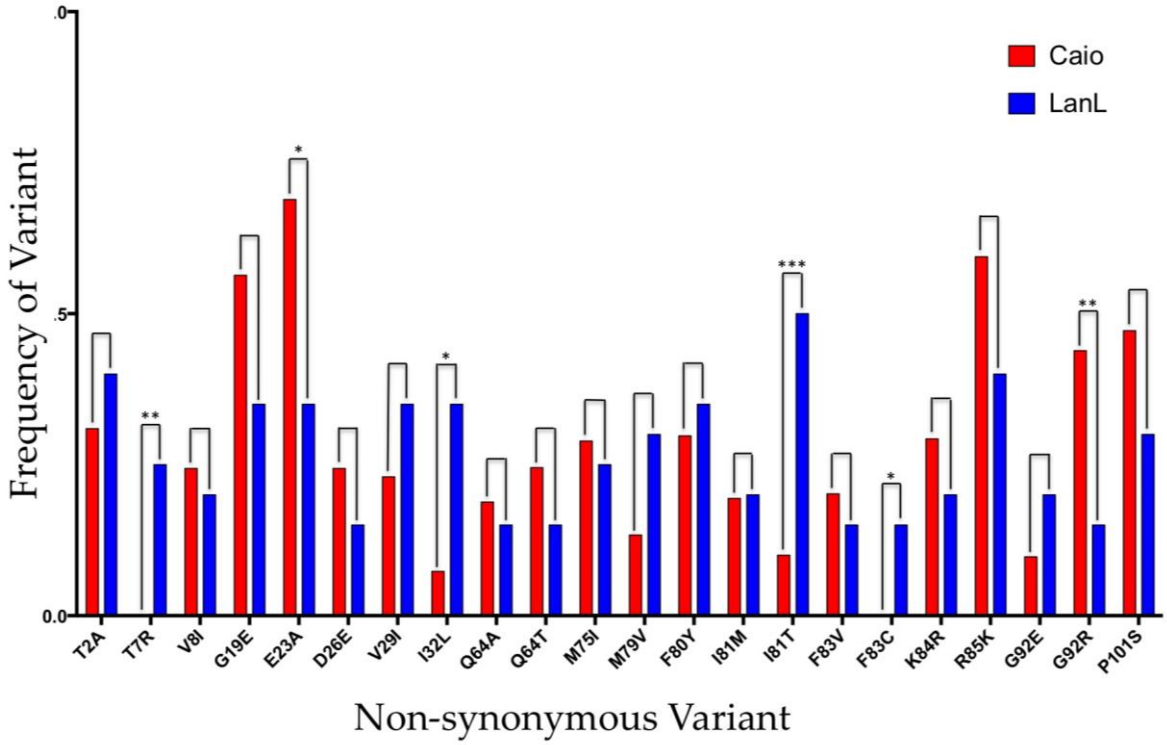



Figure 3.13: Frequencies of major variants (defined by a frequency >0.15 in LANL) in LANL HIV database and study population. Mann Whitney U test used to assess significant differences. *p≤0.05 **p≤0.01 ***p≤0.001

Minor frequency variants were selected for comparison if they were present in the study population at f >0.05. Five mutations were identified that fulfilled these criteria (**Figure 3.14**). Only one of these reached statistical significance. A substitution of leucine to isoleucine at position 74 was seen in 37% of sequences in the study population compared to 10% in the reference panel.

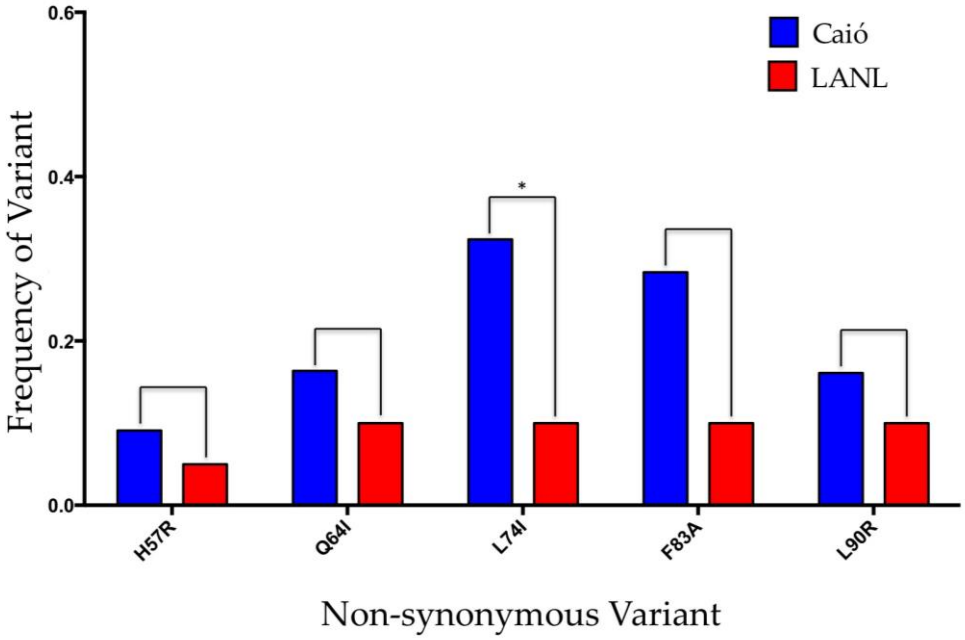



Figure 3.14: Frequencies of minor variants
(defined by a frequency <0.15 in LANL) in LANL HIV database and study population. Mann Whitney U test used to assess significant differences. *p≤0.05

Five variants that are not present in the reference sequences were identified in the study population (**Figure 3.15**). Two of these (T67A and Y69F), were located in the NLS, which was previously thought to be one of the most conserved regions of *vpx*. The highest frequency of a novel mutation was 0.22, showing that these previously unidentified mutations can reach non-trivial frequencies in natural populations. This suggests that the current knowledge of variation in *vpx* is somewhat lacking and highlights the importance of studying sequences from primary patient samples rather than cultured isolates.

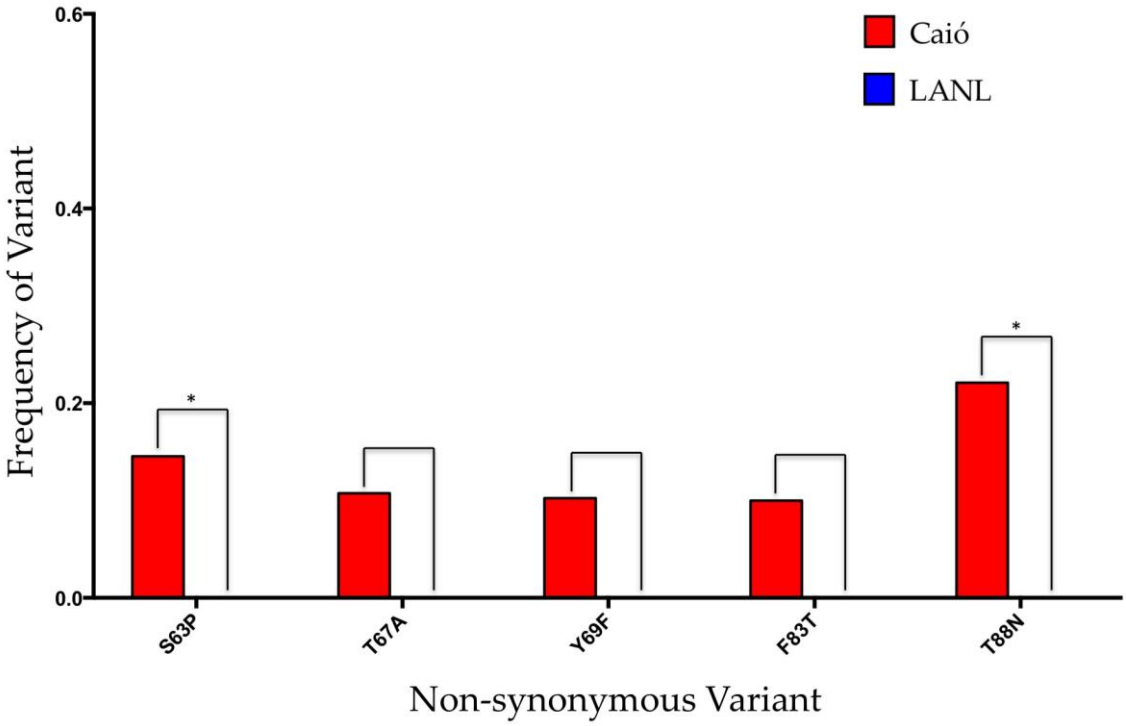



Figure 3.15: Frequencies of variants seen exclusively in the study population. Mann Whitney U test used to assess significant differences. *p≤0.05

### 3.2.9 Comparison of *vpx* variant frequencies in progressors and non-progressors

The role of *vpx* in influencing clinical outcome following HIV-2 infection remains unknown. In order to address this question, allele frequencies were compared between progressors and non-progressors (**Figure 3.16**). Mean frequencies of each allele in the two groups were compared using an unpaired t-test with a Bonferroni correction for multiple comparisons. Although no statistically significant differences were seen, several alleles showed interesting distributions. The R → K substitution at position 85 was seen at a frequency of 0.47 in viraemic patients but at 0.63 in non-viraemic patients, showing that the majority 'WT' allele at this position is different between the two groups. The largest difference was seen at position 26 where the D → E substitution was at a frequency of 0.13 in non-viraemic patients and at a frequency of 0.43 in viraemic patients. Additionally, several mutations were present in only one of the progression group. Variants D3T, T28S, A31T, Q64R and K84G were present only in patients with detectable viraemia and were never seen in non-progressors. The largest difference

was seen in A31T, which was at a frequency of 0.13 in the progressors. Two mutations were seen in non-viraemic patients only. These were Y69F and M79L. Y69F has previously been identified as a possible mutation of interest due to its location in the NLS and in this study it seems to be associated with non-viraemic patients, although the reasons behind this association are not clear, as it has no effect on SAMHD1 degradation *in vitro*.

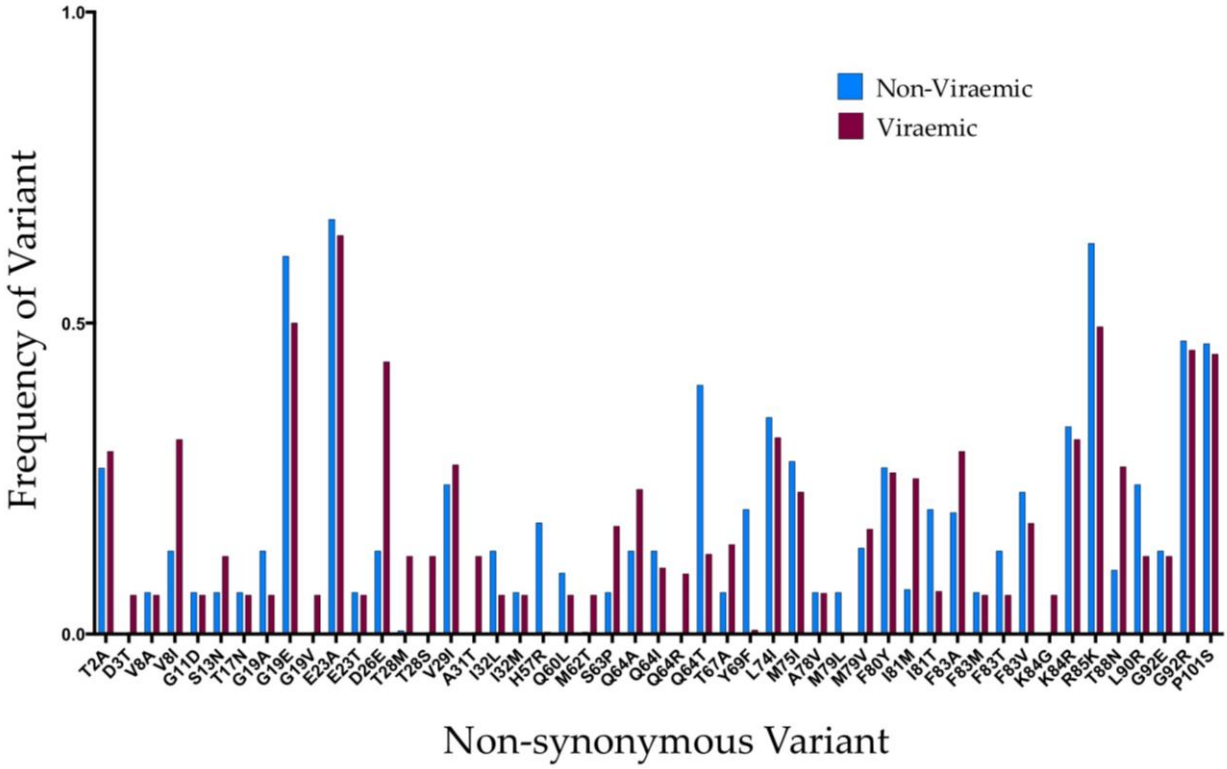



Figure 3.16: Comparison of variant frequencies between viraemic and non-viraemic individuals.
Frequencies were estimated for all variants seen at >0.05 in the study population and plotted according to location in the gene

### 3.2.10 Intra-Patient genetic diversity

In order to assess whether diversity in *vpx* was correlated with HIV-2 disease progression, the mean diversity over the tree in *vpx* sequences from a single time point was compared to two clinical markers of progression, absolute CD4 count (cells/μL) and CD4% (as a percentage of total lymphocyte count). Diversity was calculated using a custom perl script estimating the average pairwise distance between all sequences in the specified cluster from a maximum likelihood

phylogeny generated in Garli with 100 bootstrap replicates. Mean diversity was compared with absolute CD4 count and CD4% (**Figure 3.17**). No significant correlation was seen in either comparison (CD4%; $r_s$=-0.11 p=0.57, absolute CD4; $r_s$=-0.26 p= 0.10). Therefore it seems that diversity in *vpx* is not correlated with HIV-2 disease progression.

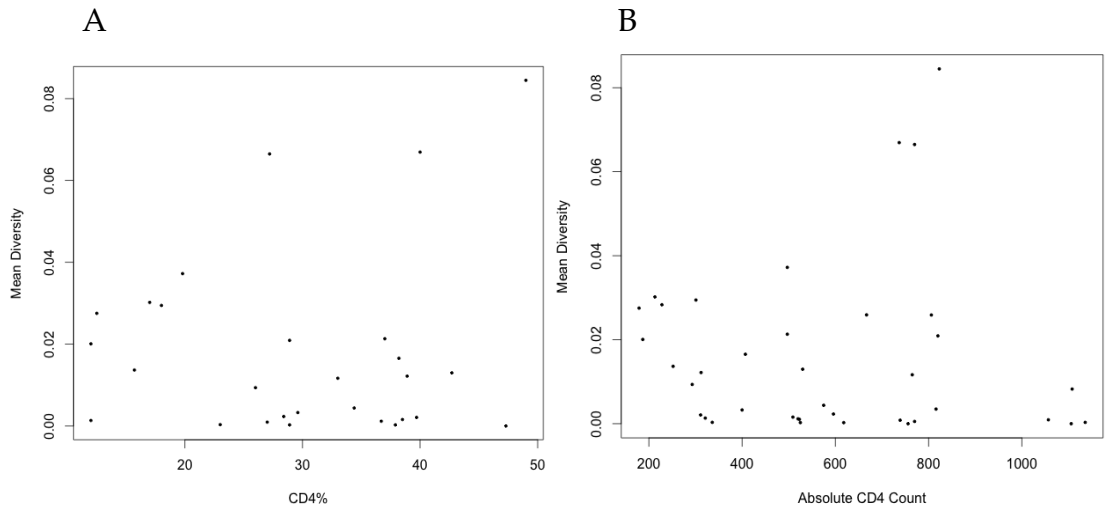



Figure 3.17: Diversity as a function of CD4% and absolute CD4 count. Scatterplots showing mean diversity against CD4% (A) and absolute CD4 counts per μL (B). No significant associations were seen. CD4%; $r_s$=-0.11 p=0.57, absolute CD4; $r_s$=-0.26 p= 0.10

### 3.2.11 Evolutionary rate analysis

Nucleotide substitution rates were estimated for 9 patients with longitudinal samples available. Rates were estimated using Bayesian MCMC inference under the HKY+I+G model of nucleotide substitution, a strict molecular clock and a constant population size. Both strict and relaxed molecular clock models were tested independently and compared by estimating a Bayes Factor (BF)[336]. The strict clock provided a significantly better fit to the data, which was indicated by a BF >20. Alignments of in-frame time-stamped sequences were used, which also enabled relative evolutionary rate estimations to be made separately for codon positions 1+2 and codon position 3. MCMC estimates were generated using BEAST v1.8.0 and the MCMC was performed using 100,000,000 iterations with posterior probability sampled every 10,000 generations.

The estimated median nucleotide substitution rate for *vpx* was $3.77x10^{-3}$ substitutions/site/year, which was inline with previous estimates of HIV-2 clock rate (**Table 3.4**)[337]. Median rates between viraemic and non-viraemic patients were compared and no statistically significant difference was observed (**Figure 3.18A**). Whilst the viraemic patients had a relatively constant clock rate ($2.58x10^{-3}$ – $3.77x10^{-3}$), non-viraemic patients showed much more variation in evolutionary rate and appeared to be stratified into two groups. Two patients had an extremely slow clock rate whilst the other 3 had a rate that was higher than the viraemic patients. This could be due to differing viral replication dynamics between patients.

The relative evolutionary rate estimates for codons 1+2 and codon 3 were used to assess the ratio of the rate attributable to substitutions at positions 1+2 and position 3 (**Figure 3.18B**). Partitioning of the rate estimate gives an approximation of the selection pressures acting on the region. There is no significant difference between the median rate ratio between the two progression groups but the median is higher for the viraemic patients and patient TD055 has a rate ratio of >1, indicative of positive selection. The other patients had a rate ratio of <1, which suggests *vpx* may be under purifying selection in these patients.

Table 3.4: Substitution rates in *vpx*.

Substitution Rate at $10^{-3}$ subs/site/year

| Sample | Group | Rate | 95% HPD |
|---|---|---|---|
| TD035 | Non-Viraemic | 1.08 | (0.10-2.20) |
| TD041 | Viraemic | 3.77 | (1.36-6.60) |
| TD046 | Non-Viraemic | 6.49 | (2.73-11.1) |
| TD047 | Viraemic | 3.17 | (1.14-5.61) |
| TD050 | Non-Viraemic | 6.39 | (1.19-12.9) |
| TD055 | Viraemic | 2.58 | ( 0.582-4.74) |
| TD058 | Non-Viraemic | 1.99 | (0.455-3.99) |
| TD061 | Non-Viraemic | 6.87 | (2.89-11.8) |
| TD093 | Viraemic | 4.38 | (1.65-7.55) |

Mean rate and 95% highest posterior densities (HPD) are shown per patient.

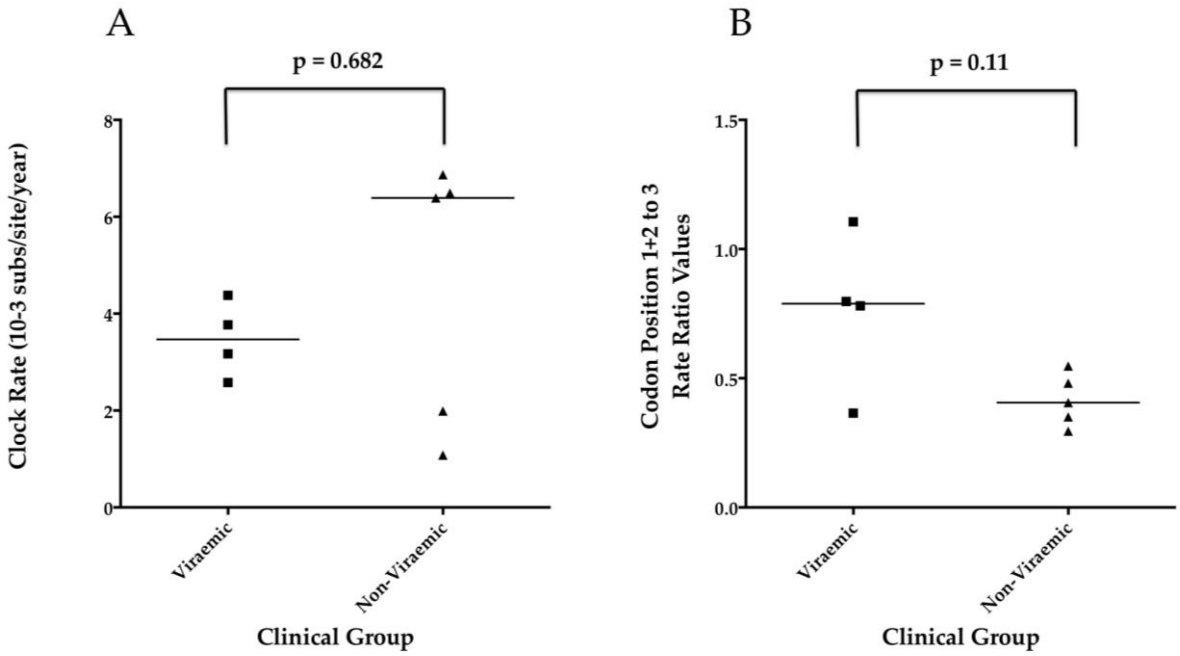
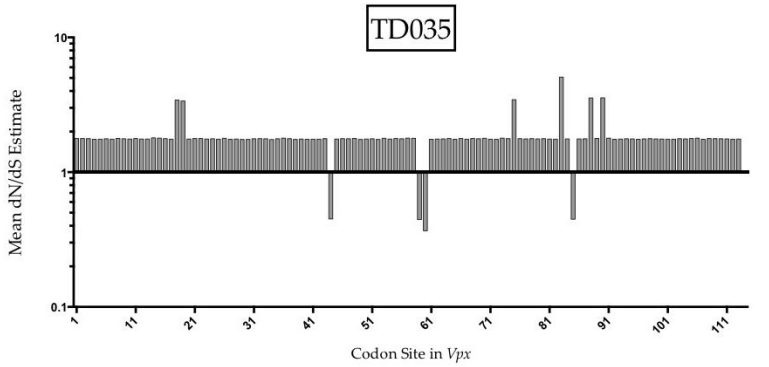


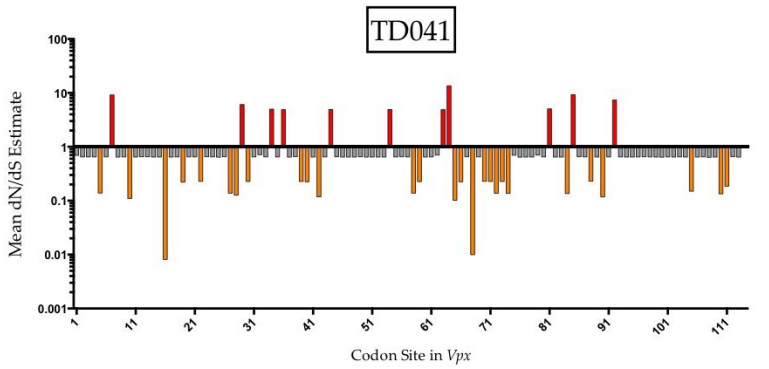

Figure 3.18: Evolutionary rate analysis of *vpx*.
Mean clock rate is shown for viraemic and non-viraemic patients and the group median is shown as a horizontal line (A). Codon rate ratio values of the 1st + 2nd to 3rd codon positions for viraemic and non-viraemic patients (B). Horizontal line shows the group median. Differences between the median values for the two groups was assessed using the Mann-Whitney U test and p values for the comparisons are shown

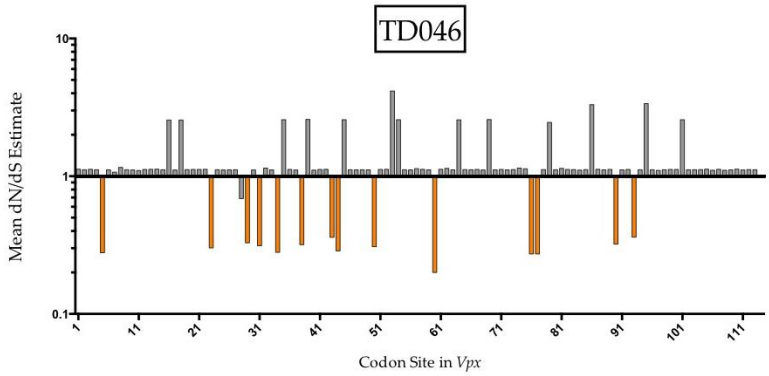
In order to further characterise the selection pressures acting on *vpx*, the ratio of non-synonymous to synonymous mutations (dN/dS) was estimated for each codon position. A dN/dS ratio >1 is caused by an excess of non-synonymous mutations, indicative of positive selection. A dN/dS ratio <1 shows the converse, indicating negative or purifying selection. Estimates of dN/dS were made using counting renaissance implemented in BEAST v1.8.0[338]. Counting renaissance is a hybrid method that combines crude counting estimates with the posterior distribution of ancestral state reconstructions and phylogenies taken from a Bayesian framework to give a quantification of the uncertainty of site-specific dN/dS ratio estimates. Estimates of dN/dS ratios were made under Bayesian MCMC inference using 100,000,000 iterations with the posterior probability sampled every 10,000 generations. The significance of the signal of selection was assessed using the 95% highest posterior density (HPD) for the estimate. A 95% HPD that did not encompass 1 was taken as significant evidence of either positive (>1) or negative

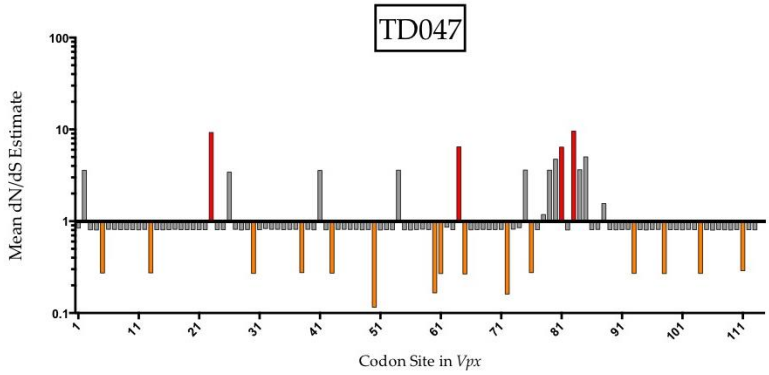
(<1) selection. Estimates of dN/dS were made for each of the 9 patients with longitudinal sequences available and the significance of each estimate was assessed (**Figure 3.19**).

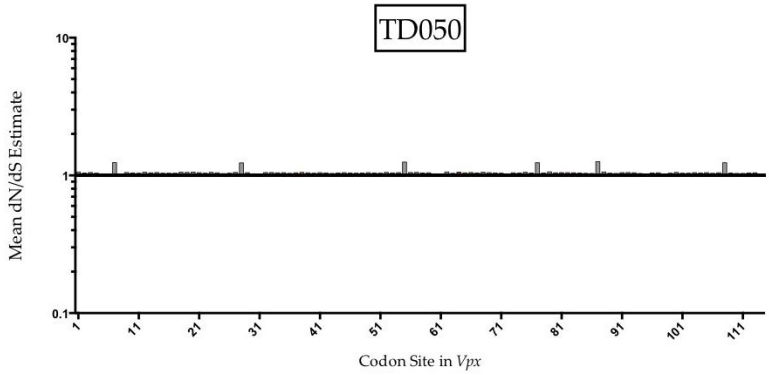
As is seen in other HIV-2 genes, there is evidence of negative selection acting on more sites in *vpx* than positive selection. Three patients (TD035, TD050 and TD093) showed no significant signals of either positive or negative selection acting at any codon in *vpx*. A further three patients (TD046, TD058 and TD061) only showed evidence of negative selection and three patients (TD041, TD047 and TD055) showed evidence of both positive and negative selection. Patient TD055 also showed a codon position 1+2:3 rate ratio of >1, indicative of positive selection. When the patients with no evidence of directed selection were excluded, no sites showed global negative or positive selection. In total, there was evidence of directed selection in at least one patient for 58 sites. Of these sites, 45 showed negative selection only, 7 showed positive selection and 6 showed evidence of both positive and negative selection. The absence of consistent signals of directed selection from all patients for any site means that no globally conserved sites can be identified. Sites 65-72 represent the nuclear localisation signal (NLS) and a mutation in the NLS (K68M) has previously been demonstrated to have an effect on SAMHD1 degradation *in vitro*. Positive selection was observed at site 65 in the NLS and 5/8 sites showed evidence of negative selection pressure in at least 1 patient. Of the two sites in the NLS that showed evidence of global neutral selection (sites 67 and 69), site 69 has been demonstrated to harbour a tyrosine to phenylalanine in approximately 10% of patients. The lack of evidence of negative selection pressure acting on this site is inline with previous observations that showed the Y69F substitution has no effect on SAMHD1 degradation *in vitro*. Site 68, which has previously been shown to contain a K68M mutation, abrogating the ability of *vpx* to antagonise SAMHD1 *in vitro* showed evidence of purifying selection in one patient, indicative of a functional constrain on this residue.


TD035
Mean dN/dS Estimate
10
1
0.1
1
11
21
31
41
51
61
71
81
91
101
111
Codon Site in Vpx


TD041
Mean dN/dS Estimate
100
10
1
0.1
0.01
0.001
1
11
21
31
41
51
61
71
81
91
101
111
Codon Site in Vpx


TD046
Mean dN/dS Estimate
10
1
0.1
1
11
21
31
41
51
61
71
81
91
101
111
Codon Site in Vpx


TD047
Mean dN/dS Estimate
100
10
1
0.1
1
11
21
31
41
51
61
71
81
91
101
111
Codon Site in Vpx


TD050
Mean dN/dS Estimate
10
1
0.1
1
11
21
31
41
51
61
71
81
91
101
111
Codon Site in Vpx

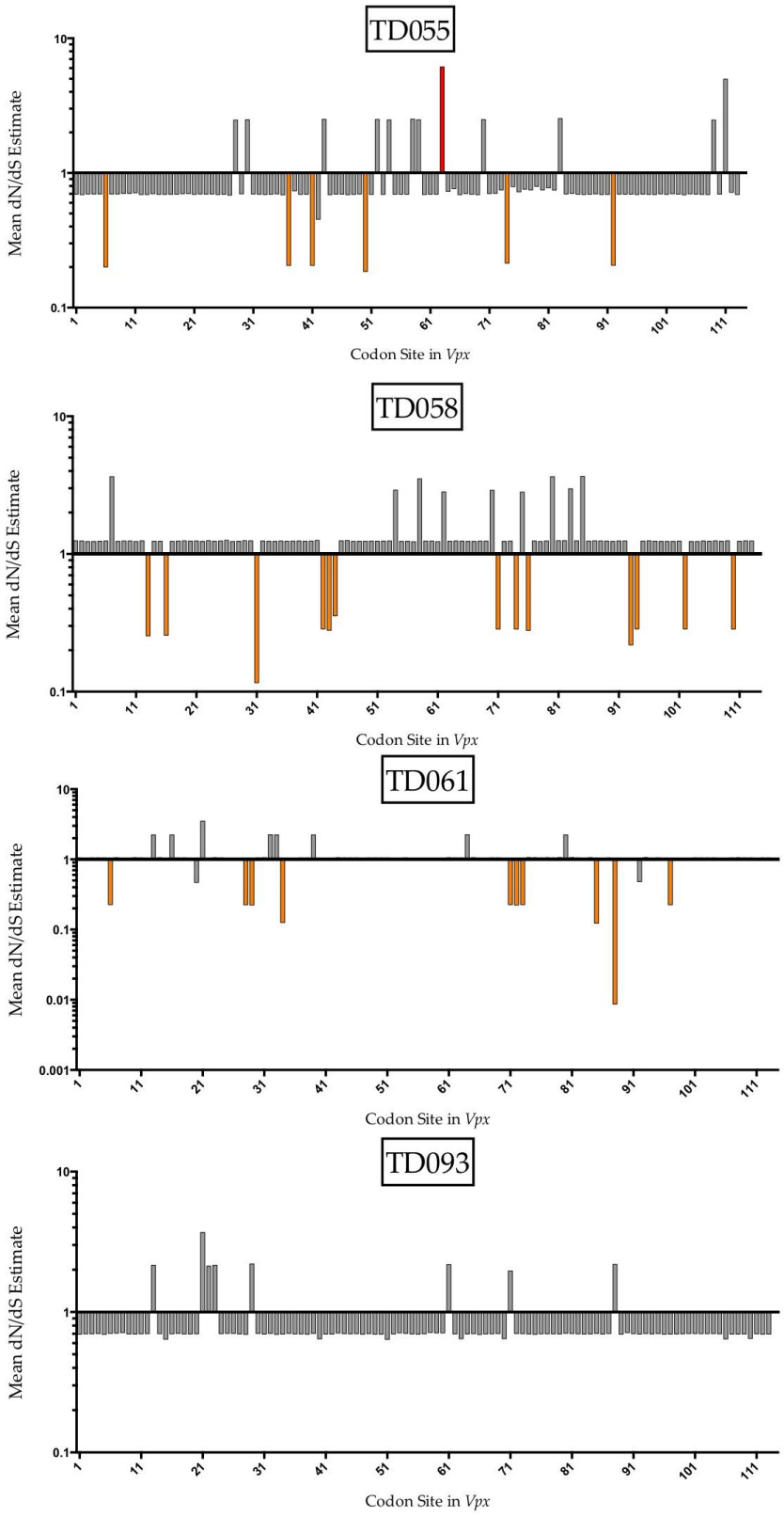



Figure 3.19: dN/dS estimates for each codon position in *vpx.*
Significance was assessed using the 95% HPD of the estimate and sites that showed significant positive selection are shown in red and significant purifying selection in orange

Therefore, whilst the majority of sites in the NLS show evidence of negative selection acting in at least 1 patient, sites 67 and 69 appear to be evolving under neutral selection, suggesting the role of these residues in correct nuclear localisation and efficient SAMHD1 antagonism may be less conserved than for the other residues in the NLS.

Sequences were partitioned according to patient progression status and the proportion of samples in each group showing evidence of either positive or negative selection was calculated for each codon position (**Figure 3.20**).

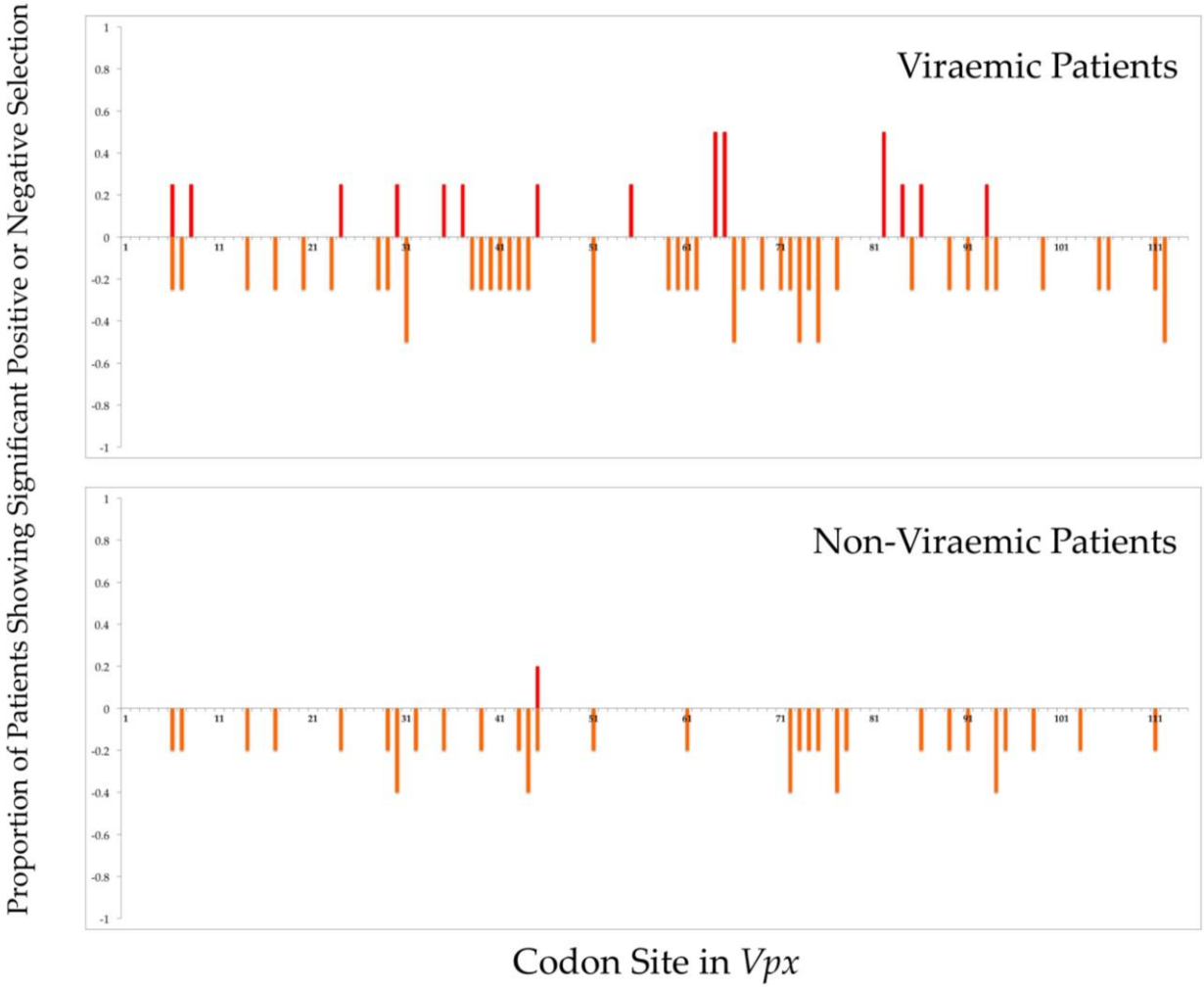



Figure 3.20: Selective pressures acting on *vpx*.
The proportion of patients showing significant evidence of positive (red) or purifying (orange) selection are shown for each codon position

There is evidence of considerably more selection acting on sequences taken from viraemic patients when compared to the non-viraemic patients. 29 sites showed evidence of purifying selection in at least one patient for the non-viraemic patients

compared to 40 positions in the viraemic patients. Equally, only one position showed evidence of positive selection in the non-viraemic patients compared with 15 sites in the viraemic patients. Positive selection pressure was seen in at least one patient from each group at position 45, which is located in the second $\alpha$-helical domain of *vpx*. The positions with evidence of positive selection in more than one patient were positions 64, 65 and 82. Position 65 is located at the start of the NLS, which has been hypothesised to have a critical role in efficient antagonism of SAMDH1.

Signals of positive selection in HIV are normally attributed to selection pressure from the host immune system. The potent block that restriction factors place on retroviral replication can lead to strong selective pressure for the virus to evolve antagonists to these proteins. The presence of signals of selection in *vpx* may point to pressures exerted by SAMHD1 on the HIV-2 genome. Additionally, the evidence for more positive selection pressure in viraemic patients when compared to non-viraemic patients suggests a dynamic interplay between SAMHD1 and *vpx* that is more apparent in patients where there is reduced control of viral replication.

### 3.2.12: Structural modelling of novel *vpx* mutations

The crystal structure of SIVsmm vpx in complex with DCAF1 and SAMHD1 has recently been solved[185], allowing the mapping of variants in the context of this structure. In order to further investigate the putative function of the novel variants identified in this study, the HIV-2 vpx structure was modelled with Chimera using the HIV-2 consensus sequence and the published SIVsmm structure[339]. Along with the high level of amino acid homology between HIV-2 and SIVsmm vpx (84%), HIV-2 vpx was also predicted to comprise three alpha helical domains, sitting within the binding pocket of DCAF1 (**Figure 3.21**). As outlined earlier in this chapter, 5 novel non-synonymous mutations were identified in the study cohort at a frequency of more than 15% (**Table 3.5**).

Table 3.5**:** Novel non-synonymous mutations in the study population

| Position | HIV-2 Consensus | Study Population Variant |
|---|---|---|
| 63 | S | P |
| 67 | T | A |
| 69 | Y | F |
| 83 | F | T |
| 88 | T | N |

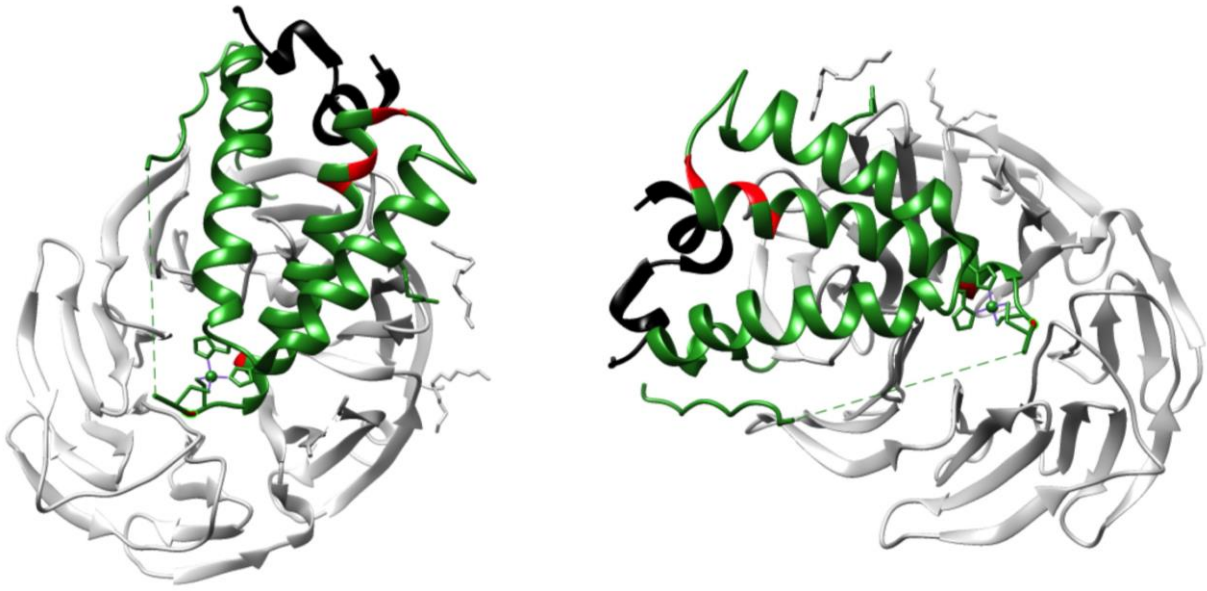


Figure 3.21: Structural model of HIV-2 vpx.
HIV-2 is shown in complex with DCAF1 and SAMHD1. DCAF1 is depicted in grey, SAMHD1 in black and the predicted structure of HIV-2 vpx in green. Mapped non-synonymous variants are shown in red.

Mapping of the novel non-synonymous variants identified in this study onto the predicted structure of HIV-2 vpx showed that 3 of the 5 mutations were closely clustered at one end of the second alpha helical domain. The 2 additional variants mapped to the other end of the second alpha helix and to the bridging region between the second and third helical domains. The close spatial clustering of these variants around the second alpha helix is suggestive of an important role of this domain in the pathogenesis of HIV-2 infection. Under the neutral model of evolution, novel variation is predicted due to the high mutation rate of HIV-2. Novel mutations then are either eliminated through purifying selection, often due to the detrimental effect they confer on the virus harbouring them, remain at low frequencies in the population with the frequency varying according to stochastic factors such as genetic drift or are positively selected for and increase in frequency

in the population, due to a selective advantage conferred on the virus harbouring them. The presence of these novel mutations in *vpx* sequences generated from multiple patients suggests that they fall into the latter group, indicative of a putative fitness gain for HIV-2 strains containing them. The predicted location of 4 of these variants in the second alpha helix highlights this as an important structural domain of vpx, making it a good candidate for future assays focussing on the effects of variation in vpx on HIV-2 disease progression.

## 3.3 Discussion

SAMHD1 antagonism by *vpx* is one of the key differences between the life cycles of HIV-1 and HIV-2, allowing HIV-2 to infect a broader range of target cells *in vivo*[178]. Whilst the absence of *vpx* or an alternative SAMHD1 antagonist in HIV-1 may be the result of selection pressures resulting from the inability of SIVrcm *vif* to antagonise chimp or human APOBEC3G, necessitating the deletion of *vpx* and overprinting of *vif* to facilitate a successful zoonosis, the lack of SAMHD1 antagonism clearly does not pose a barrier to the establishment of successful HIV-1 infection[208]. Therefore, the implications of SAMHD1 antagonism for the natural history of HIV-2 infection remain unclear. Structural studies have elucidated the method of *vpx*/SAMHD1/DCAF1 interaction and a small study of *vpx* alleles from isolated viruses showed that the mutation K68M can abrogate SAMHD1 antagonism *in vitro*[223]. The number of *vpx* sequences derived from HIV-2 group A is much lower than for other genes and no previous studies have generated *vpx* sequences from primary patient samples. The data presented in this study greatly increase the sequence knowledge of *vpx* and were generated in order to assess the distribution of naturally occurring alleles and signals of selection pressure in *vpx*.

The advent of next generation sequencing technologies has revolutionised the study of human genetics. Projects such as The 1000 Genomes Project have allowed a comprehensive catalogue of human variation to be created, something that is prohibited by the cost of Sanger sequencing of large genomes[282]. In contrast, the size of the HIV genome and the availability of tools capable of dealing with large NGS datasets have meant that clonal expansion and Sanger sequencing of PCR products is still commonly employed when looking at HIV evolution.

An alternative approach employed to avoid the uncertainties and biases involved in colony selection and amplification is single genome amplification (SGA). During SGA of the region of interest, the initial template is diluted to ensure that PCR amplification occurs from a single template, followed by direct sequencing of the amplicon. SGA allows unbiased assessment of the genetic

diversity of the target population. The amplification reaction is run in parallel and the probability of amplification occurring from a single template is a function of the ratio of amplification reactions that are successful (under a Poisson distribution, the probability of amplification occurring from a single template is more than 80% when the success rate of amplification falls below 30%)[316]. SGA or terminal dilution is a robust assay, however, it is dependent on equal primer binding affinity between templates and is highly cost and labour intensive. A recent study by Henn *et al* compared the variant calls from 454 pyrosequencing, traditional clonal expansion (768 clones sequenced) and SGA sequencing (87 single genomes) of a 1544bp region spanning *vif* to *tat*. Overall there was a 95.6% concordance between the three methods in the detection of variant/invariant sites, suggesting that 454 pyrosequencing has a similar ability to profile HIV quasispecies to SGA and traditional cloning methods[340]. Therefore, SGA was not conducted as the variant sites identified using SGA and traditional cloning show a high concordance.

This study also aimed to compare the results of different sequencing methods on a subset of *vpx* sequences generated. Comparisons were made between the Illumina MiSeq platform (sequencing by synthesis) and traditional cloning and Sanger sequencing (20 clones sequenced per patient). A reduced number of clones were used in this study when compared to Henn *et al*, and this number was chosen as it represents a typical number used in studies of HIV evolution. In line with the report of Henn *et al*, this study showed high concordance between variant frequencies estimated using MiSeq and traditional cloning ($R^2$=0.889). However, when the estimates of nucleotide pairwise diversity ($\pi$) were compared, there was a significant difference in the magnitude of estimates (paired Wilcox Signed Rank Test p=0.00167). Comparisons of the correlation between per patient $\pi$ estimates for both methods showed a significant positive correlation ($r_s$ = 0.7009, p=0.0023) and therefore inter-patient comparisons of $\pi$ are valid, provided that sequences have been generated using the same method.

The possible causes of the differences in estimates of $\pi$ are likely to be due to the features inherent in either sequencing method. It is worth noting that both methods were performed on the same PCR amplicons and therefore an identical underlying PCR bias that can lead to an erroneous increase in genetic homogeneity will be present in both methods[341]. Sampling of 20 clonally expanded products allows a lower bound of 0.05 for frequency estimations, which may lead to an over estimation of the true frequency of rare variants, leading to a reduction in the estimation of $\pi$. Equally, a cut off of 0.05 means that rare variants present in the population at a $f<0.05$ may be missed in the analysis entirely. High depth (~7000x) MiSeq data gave much lower estimates of $\pi$ than traditional cloning, showing that there maybe a bias in cloning, leading to inaccurate variant calling and diversity estimates. In order to assess this bias, the correlation between log plasma VL and concordance of variant calling (given by $R^2$) was assessed. A significant inverse correlation was observed ($r_s$= -0.76, p=0.0003) suggesting that the accuracy of traditional cloning decreases as a function of the underlying viral population size. Although this study utilised proviral sequences, VL measurements were taken from the free virion population in the plasma, as proviral load data was not available. This means that some caution has to be exercised when interpreting the results. However, there is some evidence that plasma and proviral loads are correlated in HIV-2 infection and therefore we can conclude that the power to accurately call variants from a fixed number of clones decreases with increasing VL[323]. Overall, the comparisons between the different methods showed that there is a significant correlation between both for estimates of diversity and the ability to call variant frequencies. However, the magnitude of the estimates of $\pi$ are significantly different and the power of cloning to estimate variant frequencies accurately decreases with increasing VL, leading to a set of considerations for the design of future studies of HIV evolution.

Analysis of the distribution of variant frequencies in previously existing and newly generated *vpx* sequences showed an excess of rare variants when compared to the expected distribution under a neutral model of evolution. An excess of rare

variants is normally attributed to purifying selection or a recent population expansion. Selective pressures acting on *vpx* were assessed using renaissance counting, which allows a robust estimation of the selective pressure acting on each codon and gives statistical support for the estimate. Sequences used in this study were generated from proviral DNA in PBMCs, which has previously been shown to cluster with sequences derived from plasma RNA in HIV-2, suggesting a continuous and unrestricted flow in the viral population between these two compartments[342]. Previous studies have shown evidence for negative selection pressure acting on HIV-2 *env* and this study also showed more evidence of negative selective pressure than positive pressure[231]. This study used sequences from 9 patients from two time points (2003 and 2010) to assess *vpx* evolution. The clock rate was in line with previous estimates for HIV-2 (mean = 3.14 $x10^{-3}$ subs/site/year) and no significant difference in clock rate was seen between the two patient groups. Partitioning of evolutionary rate estimates between codon positions 1&2 and 3 suggested *vpx* was evolving under negative selection pressure in all patients but one (TD055). Counting renaissance showed greatest evidence of positive selection in patient TD041, which was not supported by the partitioning of rate ratios, suggesting that counting renaissance offers a more robust and detailed view of selection pressure. When dN/dS ratios were partitioned by patient progression status, there was evidence of more sites of positive selection in the viraemic compared to the aviraemic patients. Only one site (site 45) showed evidence of positive selection in aviraemic patients, whereas 15 sites were under positive selection in at least one of the viraemic patients. Equally, more evidence of negative selection was seen in the viraemic patients (40 sites vs. 29 sites in aviraemic patients). Previous studies have suggested that the NLS is crucial for *vpx* function and appears to be highly conserved. This observation stems from the identification of two non-synonymous mutations in the NLS, one of which reduces the efficiency of SAMHD1 antagonism *in vitro*. This study observed 5 sites in the NLS that showed evidence of purifying selection in at least one patient, one site showed evidence of positive selection and two sites showed globally neutral

selective pressure. This observation is in line with the hypothesis that the NLS is conserved in *vpx* and the majority of sites appear to be under functional constraint. Additionally, the proportion of variable sites in the NLS was shown to be lower than other regions of *vpx*, such as helices 1 and 3. Of the two sites showing global neutral selection, site 69 has previously been described as harbouring a mutation (Y69F) in an aviraemic patient that has no effect on SAMHD1 degradation. This mutation was observed in a similar frequency (0.1) in this study and was entirely present in aviraemic patients.

Selection pressure on the HIV genome is normally attributed to the host immune system. As *vpx* is an accessory gene targeting the host restriction factor SAMHD1, we postulate that the increased signals of selection pressure seen in viraemic patients are driven by more active viral replication and higher turnover of viral populations, leading to an advantage in overcoming host restriction factors, allowing more efficient viral replication. It would be informative to assess whether other accessory genes in HIV-2 show similar patterns of selection pressures, elucidating the impact of restriction of viral replication by host genes on viral evolution.

A comparison was made between previously reported *vpx* sequences and those generated in this study. In total, 9 non-synonymous mutations were identified at significantly different frequencies. Variant frequencies were then partitioned according to the progression status of the patients and no significant differences in variant frequencies were observed, suggesting there is not an obvious link between variation in *vpx* and HIV-2 disease progression. Additionally, no evidence was found for the K68M mutation (that reduces SAMHD1 antagonism *in vitro*) in sequences derived from primary patient samples. Therefore the frequency of this variant in natural populations appears to be extremely low, generating uncertainty around the impact it is likely to have on HIV-2 disease progression generally. However, this study did highlight several differences that merit further investigation. Variants D3T, T28S, A31T, Q64R and K84G were only present in the viraemic patients. In this progression group, positions 28 and 31 showed evidence

of purifying selection whereas positions 64 and 84 showed evidence of positive selection, suggesting variants at these positions may have a critical role in SAMHD1 antagonism. Additionally, modelling of the structure of HIV-2 vpx and the novel variants identified in this study highlighted the putative importance of the second alpha helical domain in HIV-2 disease progression.

*In vitro* analysis of the influence of these variants on SAMHD1 antagonism would give a better idea of the role of variation in *vpx* may play in the lack of control of HIV-2 replication in viraemic patients.

# Chapter 4: Using Shotgun RNA Sequencing (RNA-Seq) to Characterise *in vitro* Divergence in HIV-2 ROD

## 4.1 Introduction

Next Generation Sequencing (NGS) has the potential to revolutionise the study of pathogens, revealing at a fine-mapping level the genetic factors that may determine outcomes such as survival following exposure to communicable diseases. A common application of NGS is re-sequencing, where information gained from multiple reads is used to generate a robust consensus sequence of the underlying genome[343]. Assembled genomes are then compared to one another to look for associations between genetic variation and complex, heritable traits. In contrast, NGS of pathogen genomes uses deep sequencing to assess the genetic structure of a viral population with a single experiment[344]. Therefore, the information in each read is considered to be informative, with multiple reads at a single locus potentially originating from different members of the viral population. This fundamental shift in the purpose of generating high read depth (from error correction to the determination of population level variant frequencies) leads to a set of considerations that are unique in the context of pathogen sequencing projects.

One of the major obstacles in sequencing viral populations is the limited amount of genetic material available for library preparation. In the case of HIV, even in the presence of high viral loads, the small 10kb RNA genome has meant that target enrichment is needed prior to library preparation and sequencing. Most commonly this takes the form of reverse transcription and PCR amplification of the whole genome in overlapping amplicons, or targeted amplification of the genomic region of interest. Whilst PCR amplification allows robust capture of the whole HIV-1 genome from samples with low viral load (Gall *et al* demonstrated a pan HIV-1 strain primer set with a sensitivity of 3,000 copies/mL)[283], PCR primer design is reliant on sufficient prior sequence knowledge over the length of the whole genome to allow identification of highly conserved regions between multiple

independent samples, something that is more challenging when studying HIV-2 infection.

Additionally, PCR can introduce biases and errors prior to library preparation. PCR biases are introduced by mutations in primer binding sites that lead to differential primer binding affinities between templates[345]. Re-sampling of amplicons in subsequent rounds (known as 'PCR duplicates') may lead to the erroneous creation of apparent genetic homogeneity[346],[347]. One method that has been used to quantify PCR bias is Primer ID tagging of the cDNA synthesis reaction[348]. During cDNA synthesis of the viral RNA genome, primers are designed to target reverse transcription of the genomic region of interest. Included in the 5' tail of each primer is a unique 8-mer, encoding a general 3bp 'tag' and a unique 5bp ID. Following sequencing, the presence of the ID allows each read to be traced to a single template. Primer ID sequencing has revealed allelic frequency skewing of 2-fold to 100-fold when compared to a run without corrections, showing that the effects of PCR bias may greatly alter the results of viral sequencing projects. However, primer ID is unsuitable for sequencing platforms that contain a random fragmentation step during library preparation, one-step RT-PCR protocols and the generation of tens of thousands of unique primers means cost prohibits routine usage[283].

PCR errors may also be introduced through chimera formation and the addition of mis-matched nucleotides by the polymerase[349]. PCR mediated recombination has been shown to be in the region of 1-3% and is therefore unlikely to be a major problem, depending on the nature of the study[283]. Commercial polymerase error rates are generally low, although one study has recently reported an error rate of 0.2% when amplifying clonal HIV-1 cDNA and it is worth noting that polymerase errors may appear as high quality reads following downstream sequencing[350]. The likelihood of introducing a polymerase mediated error increases in proportion to the number of cycles of PCR performed and therefore in light of these considerations, sample preparation that minimises the use of PCR would be beneficial.

RNA-seq was initially developed to allow sequencing and quantification of the human transcriptome[351]. Whole RNA sequencing is achieved through random hexamer (RH) priming of cDNA synthesis, followed by fragmentation, adaptor ligation of the cDNA, library enrichment through PCR and sequencing[352]. Therefore, RNA-Seq library preparation can be used to generate data without the need for efficient primer design or prior PCR enrichment of target material. However, RNA-seq is not a bias-free approach and biases maybe introduced during preparation of RNA for sequencing or during post-sequencing analysis. Sources of bias in RNA-Seq include non-random RH priming and the generation of PCR duplicates during library preparation, both leading to non-uniform coverage over the target region. Additionally, assembly can be problematic in the absence of an appropriate reference sequence or when assembling a divergent population. In this instance a trade-off must be made between allowing mis-matches and therefore recapitulating the true diversity of the sample, and reducing the number of mis-mapped reads that introduce false heterogeneity into the assembly[353],[354]. Local sequence context has also been shown to influence efficiency of assembly and sequencing and both positive and negative GC biases have been demonstrated[355], [307].

In spite of the biases involved in RNA-Seq it remains an appealing solution for the sequencing of viral populations. This study aimed to utilise RNA-Seq to generate whole genome sequence data for the HIV-2 A isolate strain ROD. HIV-2 ROD was isolated in Senegal in 1985 from a patient with advanced AIDS[5]. The early isolation and ease of *in vitro* propagation has lead to ROD being thought of as the HIV-2 A prototype strain. A literature search reveals 138 publications listing HIV-2 ROD in the materials section, far higher than any other publically available HIV-2 isolate. This study aimed to look at the potential biases involved in using RNA-Seq to generate whole genome sequences for HIV-2 and to address the question of how much variation is seen in the lab adapted isolate HIV-2 ROD compared to the published reference genome sequence.

## 4.2 Results

### 4.2.1 Initial sample preparation and RNA-Seq protocol

The HIV-2 isolate ROD was cultured in the H9 lymphocyte cell line and viral replication was quantified by measuring reverse transcriptase activity (as described in Chapter 2.3.2). Nucleic acid was extracted from 1mL of cell-free culture supernatant and DNA was removed by Turbo DNase treatment. Total RNA was prepared for sequencing using the NEBNext Ultra RNA-Seq library preparation kit for Illumina (**Figure 4.1**).

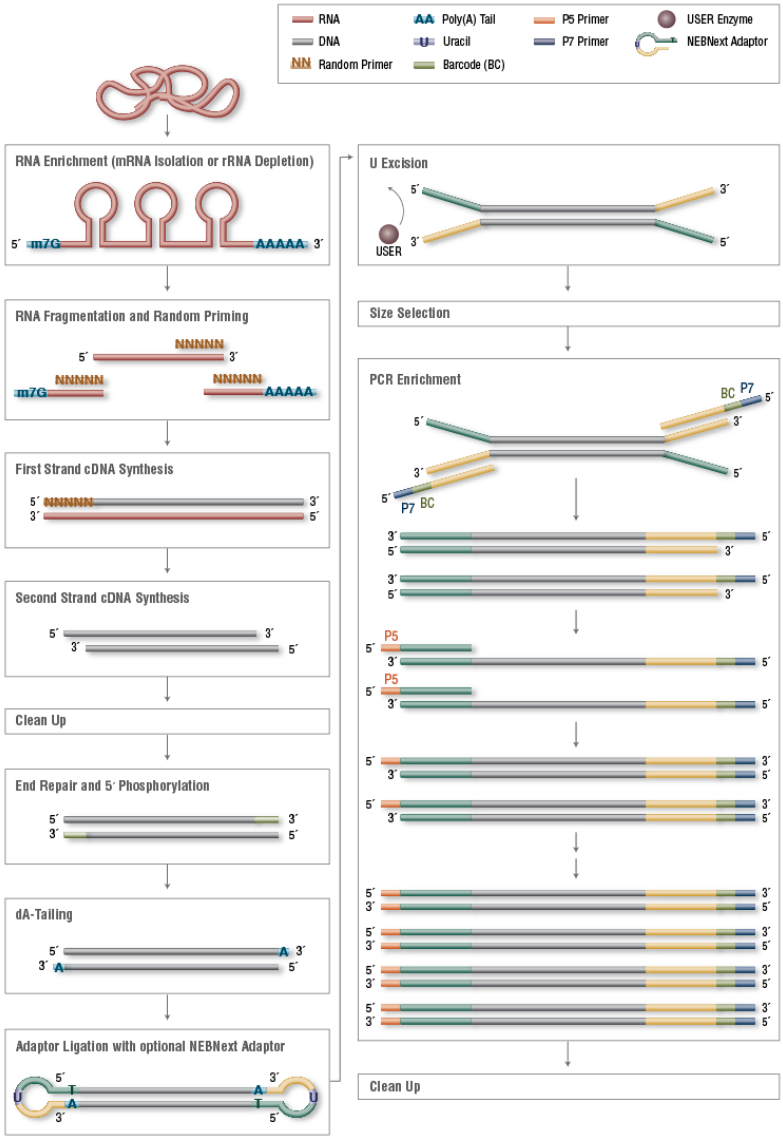



Figure 4.1: RNA-Seq library preparation protocol.
Total RNA was prepared according to the manufacturer's protocol with the exclusion of the mRNA isolation step in order to maintain a sufficient RNA concentration in the sample to allow good quality library preparation.

Library quality was assessed using a Bioanalyser DNA HS chip (Chapter 2.3.8) and libraries were run on the Illumina MiSeq platform, generating 2 x 150bp paired end reads.

### 4.2.2 Quality control and removal of PCR duplicates

Following initial quality control and removal of low quality reads and adaptor contamination (Chapter 2.3.8), reads were assessed for the presence of biased random hexamer (RH) priming. Previous studies have shown that RH biases cause distinctive patterns in the nucleotide composition of the first 13 bases of the 5' end of reads[356]. Under the assumption of truly random fragmentation and cDNA synthesis priming, each read can be thought of as an independent sample from the total population. Therefore the base composition at each position in the read averaged over all reads should be constant, re-capitulating the GC content of the sample sequenced. Base composition per read position was assessed using FastQC for both mates in the read pair (**Figure 4.2**).

In both traces a strong distinctive pattern of base composition was seen in the first 13bp of the 5' end of the reads. This is indicative of random hexamer bias, as has been shown in previous studies. The similarity of the pattern shows this is due the same underlying RH bias and not caused by a reduction in read quality at the terminal ends of the reads. The reverse reads show a slight deviation at the 3' end of the read, but this is not due to RH bias. Previous studies have shown that predicted binding energies of hexamers do not explain the observed frequencies of hexamers seen at the 5' end of reads and therefore the bias is a function of the RNA sample sequenced rather than an inherent bias in the RH pool[356]. Generally it is not necessary to trim the reads to remove the RH bias as trimming will remove the pattern from the first 13bp but it will still be seen when the upstream nucleotides are mapped to the genome. This study shows RH bias inline with other studies, which did not extend past the first 13bp of each read. RH priming of cDNA synthesis is still preferable to oligo(dT) priming of RNA molecules which results in a strongly directional bias towards the 3' end of the RNA[356].

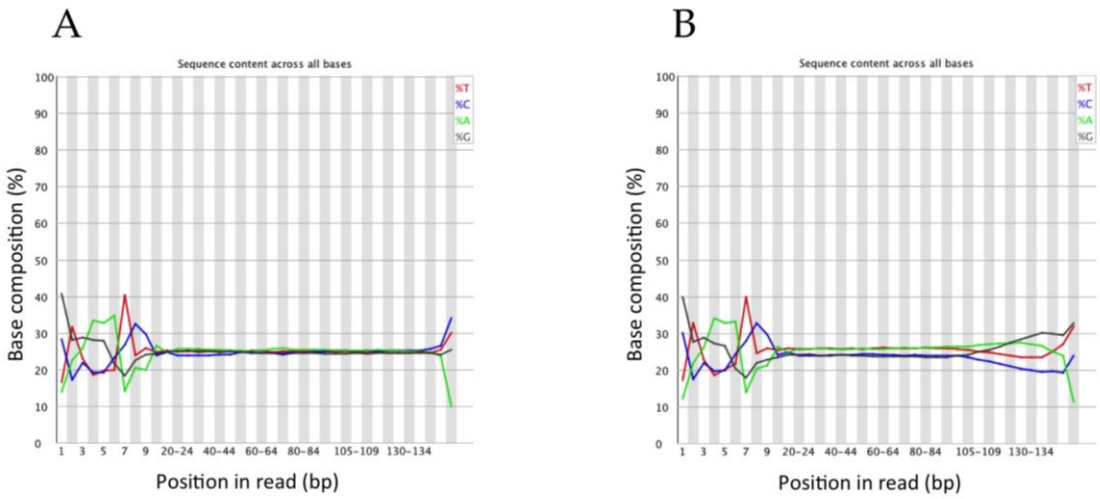



Figure 4.2: Base composition for forward (A) and reverse (B) reads. Random hexamer bias can be seen in the first 13bp of the 5′ end of both traces as a deviation from the expected base composition.

Library preparation involves a PCR enrichment step, using primers that bind to sequence motifs in the adaptor molecule. Enrichment can introduce PCR duplicates where the same fragment becomes over-represented in the final pooled amplicon library[357]. Optical duplicates are generated when signals from one cluster are mistakenly attributed to multiple adjacent clusters by the base-calling algorithm[281]. Both optical and PCR duplicates were removed from assembled reads using Picard[310]. Picard marks reads as potential duplicates if they have identical genome co-ordinates and reads that flagged as duplicates were removed from the alignment file (**Figure 4.3**). A total of 27% of reads were flagged as duplicates and removed from the assembly. Comparison of the coverage plots before and after duplicate removal reveals that whilst some of the larger peaks in the assembly were reduced (e.g. at 8001 and 9400), the overall shape of the coverage plot remains constant. This could be due to two factors: either duplicates are evenly spaced over the genome or the small length of the HIV genome means that unique reads that happen to begin at the same position are falsely identified as duplicates. It is likely that the similar coverage plots imply overly stringent duplicate removal and therefore this step results in a loss of informative reads. However, the identification and correction of some of the larger fluctuations in coverage caused

by duplicates shows that this correction is necessary, removing a bias introduced in library preparation.

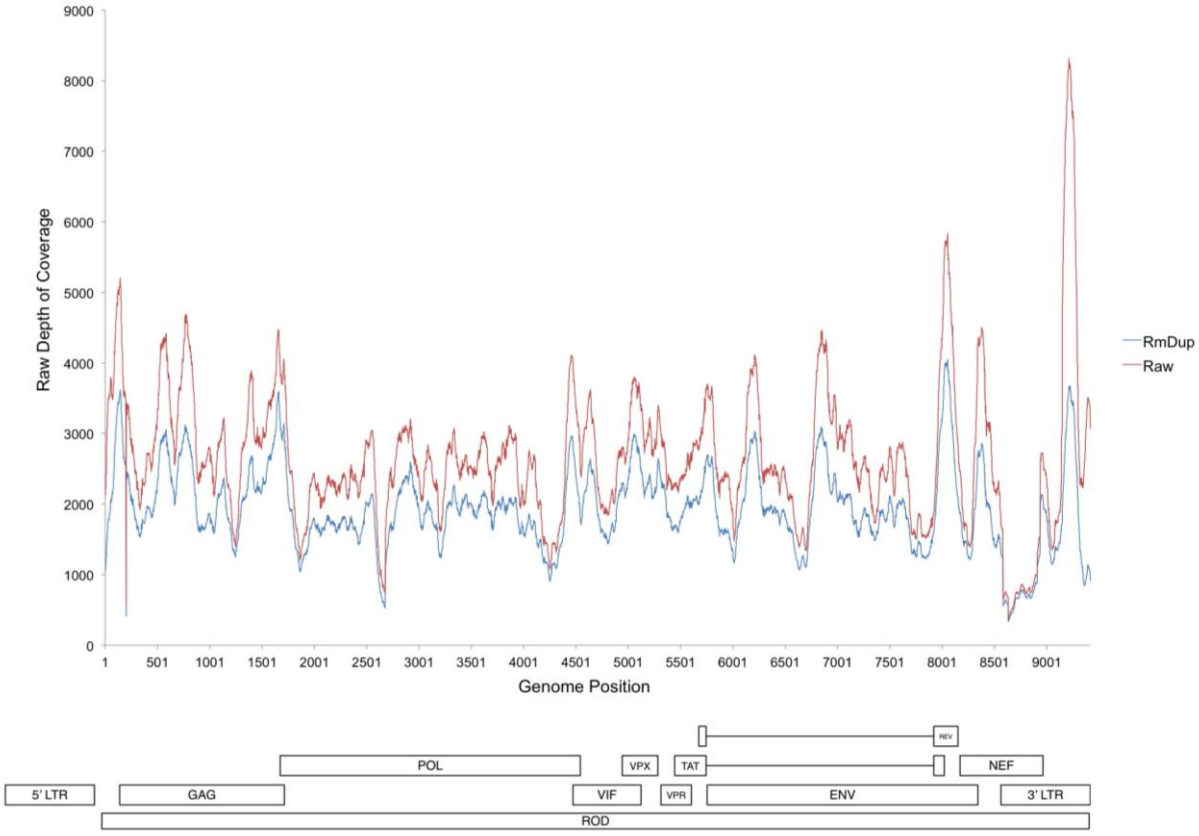



Figure 4.3: PCR and optical duplicates in assembled reads.
Whole genome was assembled using Bowtie2. Coverage according to position in the genome is shown for raw reads (red) and following the removal of PCR and optical duplicates (blue).

### 4.2.3 Assembly of reads to the reference genome

Reads were assembled to the known HIV-2 ROD reference genome sequence (AN: BD413542) using a panel of 4 commonly used alignment tools (**Table 4.1**). Performance was initially assessed based on 3 criteria; 1.) number of reads correctly assembled to the target genome, 2.) average depth of coverage over the whole genome, 3.) distribution of coverage over the whole genome (**Table 4.2**). Post alignment, sam files were filtered and sorted using the Samtools package and duplicates were flagged and removed using Picard[311]. Mean depth of coverage was calculated using the Samtools depth function, coverage at each position in the

genome was exported using the BedTools genomeCoverage function and coverage over the genome was assessed (**Figure 4.4**)[311].

Table 4.1: Alignment tools used in this analysis.

| Name | Mapping Algorithm | Reference |
|---|---|---|
| Bowtie2 | Burrows-Wheeler Transform (BWT) | [309] |
| BWA-SW | Burrows-Wheeler Transform (BWT)- Smith Waterman | [358] |
| GSNAP | Oligomer Hashing and Chaining | [359] |
| NovoAlign | Hash-based Aligner | [360] |

All tools are freely and publically available from the developer's website with the exception of NovoAlign which is licenced by Novocraft Inc.

The number of reads mapped and mean depth of coverage varied between aligners, however all aligners managed to assemble the entire coding region of HIV-2 ROD. In terms of reads mapped and depth of coverage, GSNAP was the best performing alignment tool. This is perhaps unsurprising as it has previously been demonstrated that GSNAP performs well when assembling reads containing sequence variation relative to the reference sequence[354].

Table 4.2: Summary of aligner performance.

| Aligner | Reads Mapped (% of total) | Mean Depth | % Genome Recovered | Range of Depth of Coverage |
|---|---|---|---|---|
| Bowtie2 | 165506/5154404 (3.2%) | 1924x | 100 | 336-4050x |
| BWA-SW | 152105/5154404 (2.9%) | 1794x | 100 | 320-3451x |
| GSNAP | 176885/5154404 (3.4%) | 2146x | 100 | 173-4921x |
| NovoAlign | 155696/5154404 (3.0%) | 1862x | 100 | 198-3769x |

Mean depth per locus is shown and was calculated by dividing the sum of the length of aligned reads by the length of the reference genome.

Coverage over the coding region of the genome was visualised by plotting the read depth per locus (**Figure 4.4**). Although the coverage is non-uniform, all aligners return coverage plots of remarkably similar shape, showing that even when different algorithms are used the underlying pattern of coverage is reliably recapitulated. Overall, there is a distinction when GSNAP is compared to the other aligners, and GSNAP gave a slightly higher mean depth of coverage over the majority of the genome.

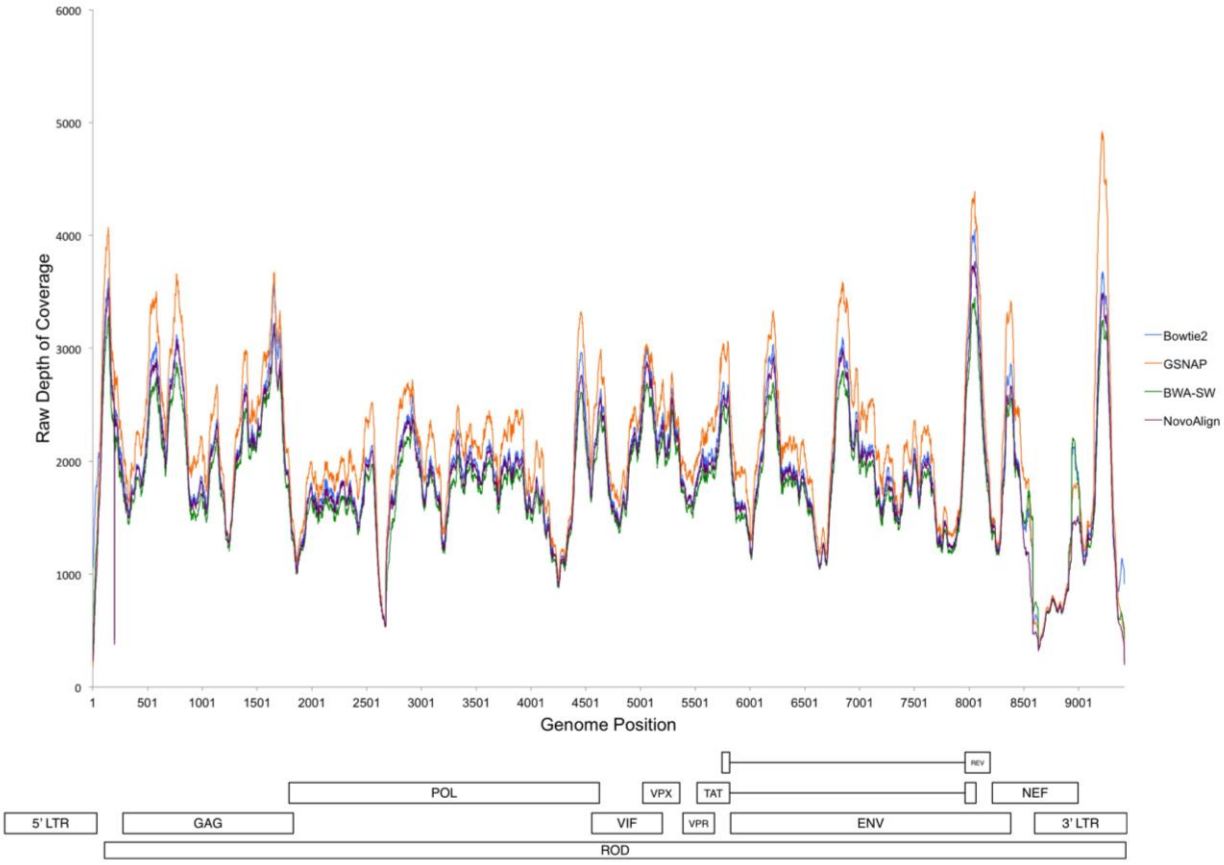



Figure 4.4: Distribution of assembled reads across the genome.
Raw values of depth of coverage were plotted across the length of genome for each aligner.

### 4.2.4 Quantification of biases

In order to assess whether variable coverage across the genome was due to biases in library preparation (e.g. RH bias or RNA degradation) or due to biased read assembly, factors that are known to influence coverage were assessed. A major source of bias in Illumina sequencing is GC content of the genome studied. For example, the malaria parasite *Plasmodium falciparum* has an extremely GC poor genome with a mean GC content of 25%. A study using 36bp Illumina paired end reads showed that coverage is biased towards the GC balanced regions of the genome, with GC-poor regions showing little or no coverage[361]. In contrast, the genome of HIV-2 ROD has a mean GC content of 45%, similar to the human genome that has a mean GC content of 46%[362]. GC content and mean depths were calculated for assemblies from all aligners in a sliding window of 50bp with a step size of 20bp across the length of the genome (**Figure 4.5**).

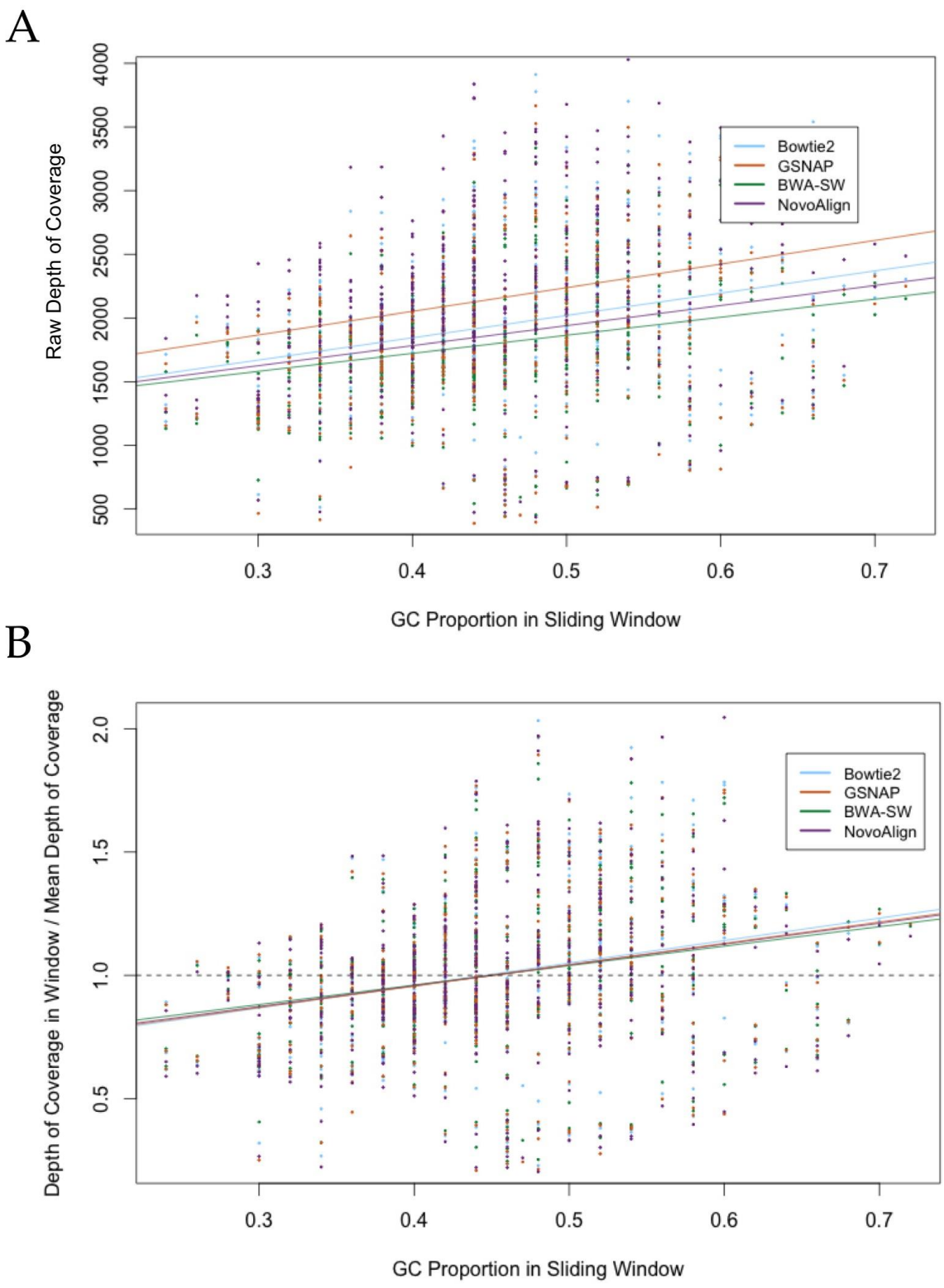



Figure 4.5: GC bias in assembled reads.
Plots show GC proportion in a sliding window of 50bp against mean coverage (A) or coverage normalised to mean whole genome coverage (B). Points and regression lines are coloured by aligner. Dashed line in plot B denotes expected regression in the absence of a GC bias.

In order to compare GC bias between different aligners, depth was normalised as a proportion of the mean depth over the whole genome. A linear regression line was fitted to the data and the degree of GC bias was defined by the slope of the line. All assemblies showed a slight positive GC bias and more GC-rich regions tended to have higher coverage. The slopes of the 4 lines were very similar (0.79-0.91), implying that the assembly algorithm used does not affect the GC-bias. Overall, the GC bias in the HIV-2 ROD RNA-Seq data was lower than has been reported from other pathogen sequencing projects (e.g. E.Coli K-12, GC = 50.8%, slope = -1.9)[355] and the magnitude of the slope implies that it should not introduce fluctuations of more than 0.45-fold in normalised coverage at the most extreme.

Divergence of the sample from the reference sequence could also have an impact on the number of reads that are correctly mapped to each genomic region. Although HIV-2 ROD is derived from a single clonal isolate, it is probable that mutations have arisen during subsequent *in vitro* culture, leading to differences between reference and sampled population. In order to assess whether divergence from the reference could have an impact on assembly quality, individual normalised coverage was calculated for each gene of HIV-2 ROD (**Figure 4.6**).

Normalised depth of coverage was relatively constant across all the genes. Different aligners returned similar normalised coverage for each gene, again showing all aligners were performing comparatively well on this data set. When the GC bias was accounted for, 3 genes showed notable deviation from the expected level of coverage. *Rev* and *vpx* showed higher coverage than expected while *nef* showed low coverage. Elevated coverage could be caused by a closer relationship to the reference sequence as assembly algorithms lose power when mapping to highly polymorphic regions. However, *rev* and *vpx* are not thought to be particularly highly conserved and the same pattern is not seen in *pol*, which is known to be the most conserved HIV-2 gene. Equally *env* is highly variable but shows coverage consistent with the genome average. Whilst some *in vitro* mutations would be expected in HIV-2 ROD, this strain has been grown in a single

cell-line culture and therefore many of the selection pressures that act to shape HIV-2 evolution *in vivo* will be absent in this system.

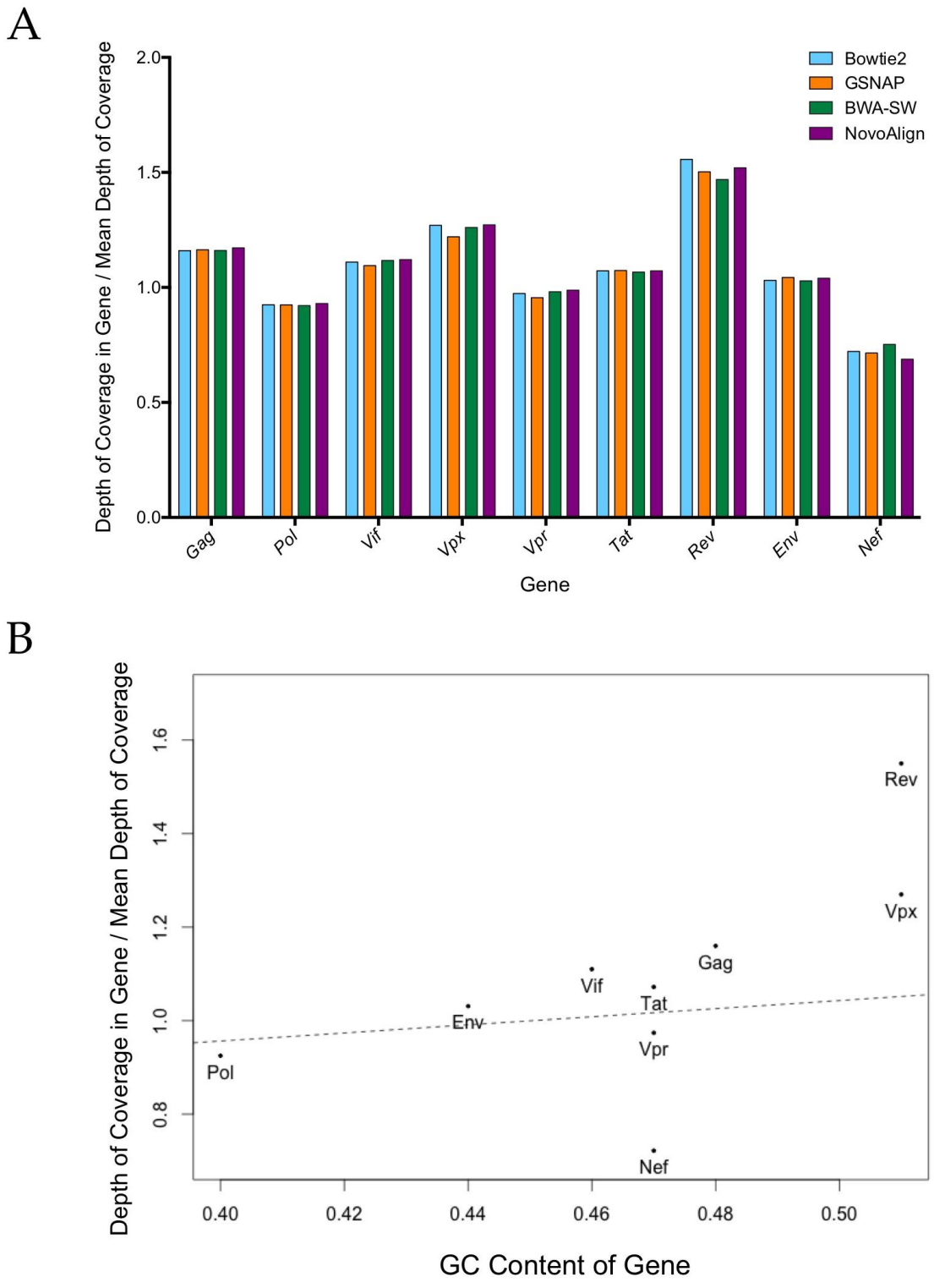



Figure 4.6: Coverage partitioned by gene.
Coverage was normalised to the mean coverage over the whole genome and calculated for each aligner (A). GC bias was accounted for by comparing GC content of the gene and mean normalised coverage (B), dashed line shows estimated GC bias.

Therefore, it seems divergence from the reference is not large enough to affect coverage in this study. The most plausible explanation is that the elevated coverage seen in *rev* and *vpx* is caused by the random hexamer bias and that this bias is more likely to be seen in shorter genes (*rev* = 232bp, *vpx* = 339) where the total length of the gene is similar to the mean library insert length (~450bp). In longer genes, where the length of the gene is many times longer than the average insert, the effect of the bias will be less obvious.

In contrast the low coverage seen over *nef* is readily explained by its genomic position. *Nef* is found at the 3′ end of the HIV-2 genome and the final 404bp of *nef* overlap with the 3′ LTR. The HIV-2 genome contains 2 semi-palindromic LTRs and in the viral ssRNA genome there is homology between the repeat (R) in the 3′ and 5′ LTRs, which makes assigning reads to the correct LTR problematic. Additionally, sequence information is often lacking at the 3′ and 5′ ends of RNA molecules and in this study the drop in coverage over the LTRs can be seen in **Figure 4.4**[363]. In order to assess the impact of 3′LTR overlap on coverage in *nef*, the gene was partitioned into non-LTR and LTR regions and normalised coverage was calculated for both regions (**Figure 4.7**).

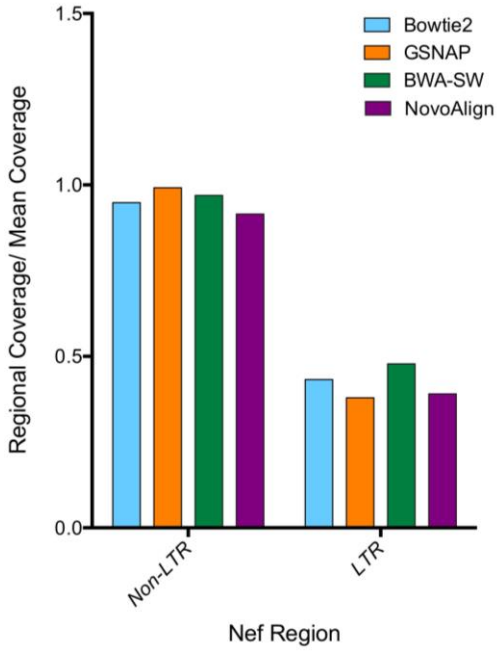



Figure 4.7: Coverage in *nef* partitioned by genomic region.
Normalised coverage was partitioned by genomic locations inside or outside the LTR in *nef*.

A significant drop in coverage can be seen when *nef* is partitioned into non-LTR and LTR regions. In the non-LTR portion of the gene, coverage is close to the expected value of 1.0 whereas in the LTR depth falls to less than 0.5 of the expected level. This drop is not explained by the GC bias and so can be attributed to the terminal location and homology between the LTRs and the associated drop in coverage. The reduction in read depth can be seen in the **Figure 4.4** at position 8565 in the genome.

### 4.2.5 Visualisation of assembled genomes and SNP calling

In order to assess the mapping quality of each aligner, the genome builds were visualised using IGV to allow identification of any conflicts between alignments (**Figure 4.9**). SNP calling between the different aligners was reproducible and therefore SNPs can confidently be called between the sample and reference HIV-2 ROD sequences. There were two major calling conflicts at positions 3892 and 8933, both involving indels (**Figure 4.8**).

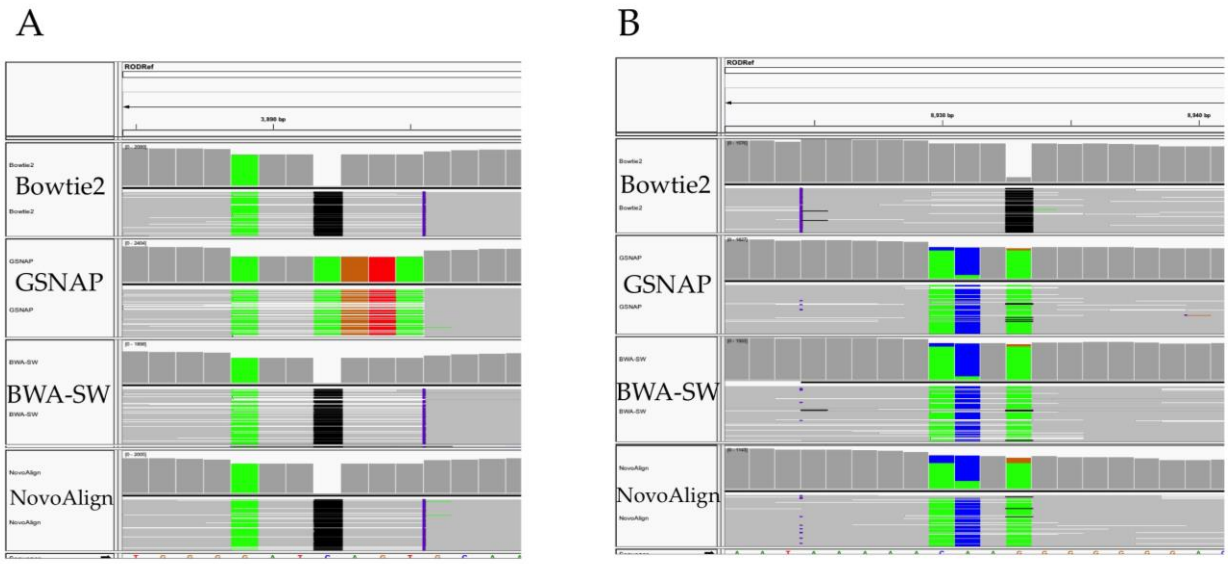



Figure 4.8: IGV visualization of assembly conflicts. Matches are shown in grey, deletions in black, insertions in purple and mismatches are coloured according to nucleotide

At position 3892 (**Figure 4.8A**), all aligners apart from GSNAP called a deletion at position 3892, followed by an adenine insertion 3 positions downstream, restoring the reading frame. GSNAP did not model the deletion and instead

showed four consecutive mismatches from the reference until the putative adenine insertion, which brings the region back into sync with the reference reading frame. At position 8933 the pictures is a little more complex (**Figure 4.8B**). GSNAP, BWA-SW and NovoAlign called 3 mismatched positions, whereas Bowtie2 interpreted the divergence as an adenine insertion followed by a deletion 8 bases downstream. The differences in assemblies are caused by the alternative scoring systems of the aligners. In one sense there is no 'correct' answer, as each assembly can be thought of as a hypothesis of the true structure, based on an explicit model of the biological cost of each genome modification. However, the genome context of each conflict can be useful in determining which model to believe. Position 3892 is located in the coding region of *pol*, a conserved gene that is crucial to viral replication. Therefore, it is more plausible that any deletion in *pol* causing a shift in reading frame would lead to high selective pressure for an insertion (or vice versa) that would restore the reading frame and lead to the production of full length proteins. Position 8933 is located in the non-coding 5' LTR. However, the LTR is vital for correct reverse transcription of the HIV-2 genome and therefore it is unclear whether a deletion and compensatory insertion or multiple SNPs would be better tolerated.

Conflicts in genome assembly between aligners show the importance of considering multiple assembly strategies when working with highly divergent pathogen genomes, where many more genome modifications may be tolerated than are typically seen in mammalian genomes. However, there is good agreement between different aligners for the majority of loci in the HIV-2 ROD genome and therefore we can confidently call the assembly a reliable representation of the underlying sequence variation in the sample.

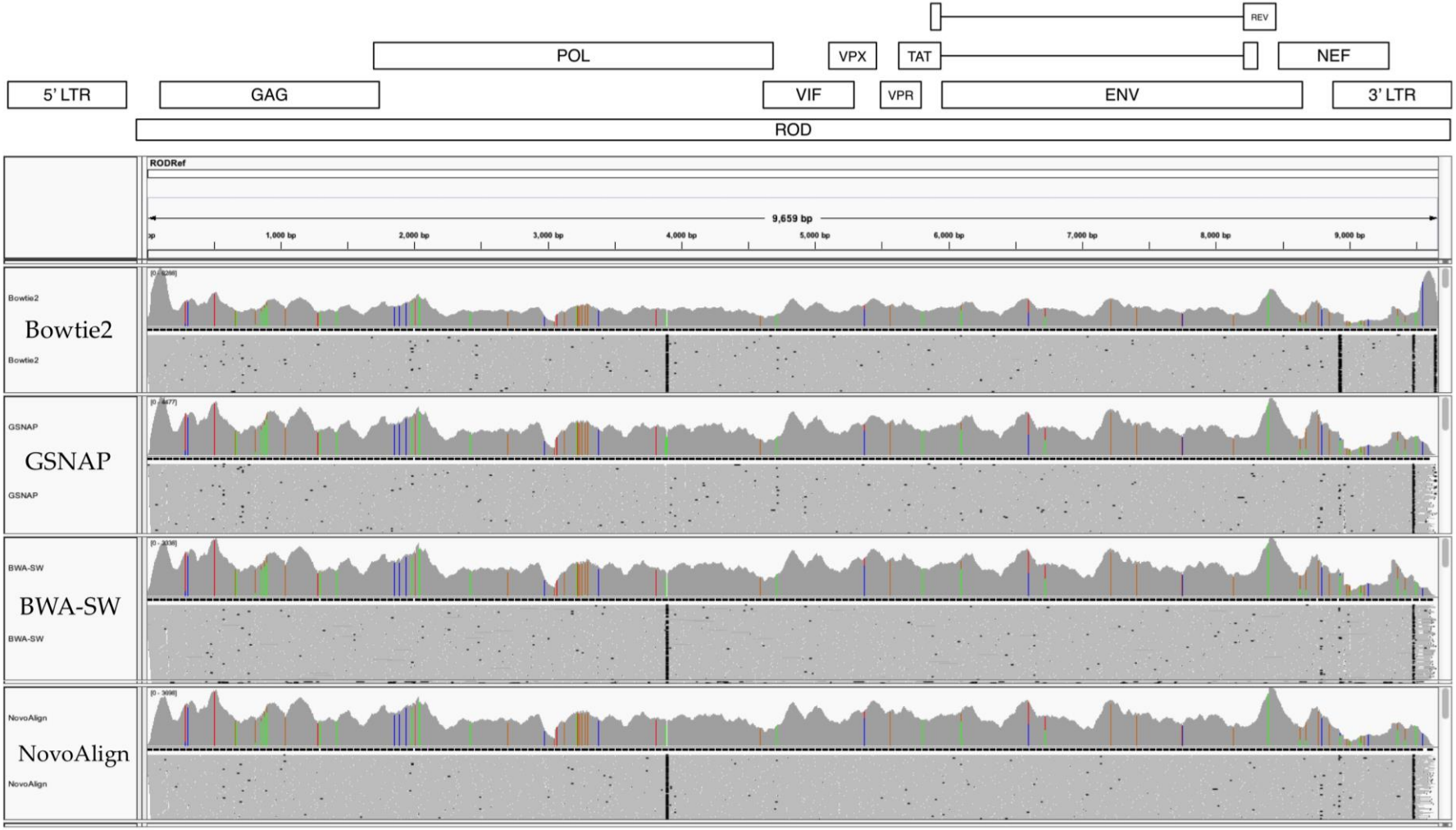



Figure 4.9: IGV visualisation of genome builds.
Tracks per aligner show depth of coverage (top) and mapped reads (bottom) are shown for all 4 aligners. Sites of mismatch with the reference are coloured according to nucleotide and gaps are shown in black

### 4.2.6 Divergence from the published reference sequence

One of the most important factors when designing *in vitro* studies of viral infection is the choice of viral strain used. HIV-2 is normally derived from two sources, either novel virus is isolated by co-culturing infected PBMCs from the patient of interest with healthy PBMCs or a previously isolated reference strain is used. Reference strains often have the advantage of available whole genome sequences. One of the implicit assumptions when using an HIV-2 reference strain and the publically available genome sequence is that any variation in the sequence introduced during prior propagation will not have an impact on the experimental design or results. Often reference strains are passed from lab to lab through collaborations and therefore it is not always apparent how many passages through cells a reference strain has undergone.

In order to assess how divergent the HIV-2 ROD propagated for this study was from the published reference sequence (AN: BD413542), polymorphisms that were fixed at a frequency of more than 0.95 in the sample population were annotated as SNPs. The BWA-SW build was used for this analysis as it agreed with the majority consensus at each site of conflict. A total of 70 SNPs were identified in 8/9 HIV-2 ROD genes (**Figure 4.10**).

The majority of SNPs were seen in *gag*, *pol* and *nef*. When a correction was applied to account for the gene length, *nef* showed the greatest contribution to divergence from the reference genome. This may be explained by the function of *nef*. *Nef* is a viral infectivity factor and it is often thought to be dispensable for successful HIV infection. One of the major functions of *nef* is the evasion of host-immune responses through down regulation cell-surface expressed of HLA molecules and T-Cell receptors. *In vitro* culture of HIV-2 can be thought of as 'survival of the most', where the short generation times between infection and assay may favour viruses that rapidly replicate, producing high titres and overtaking more slowly replicating variants. The absence of the complex interplay of the host immune system *in vitro* also leads to less constraint on accessory genes (such as *nef*) that are

not critical for viral replication and whose functions are targeted towards evasion of host immune responses.

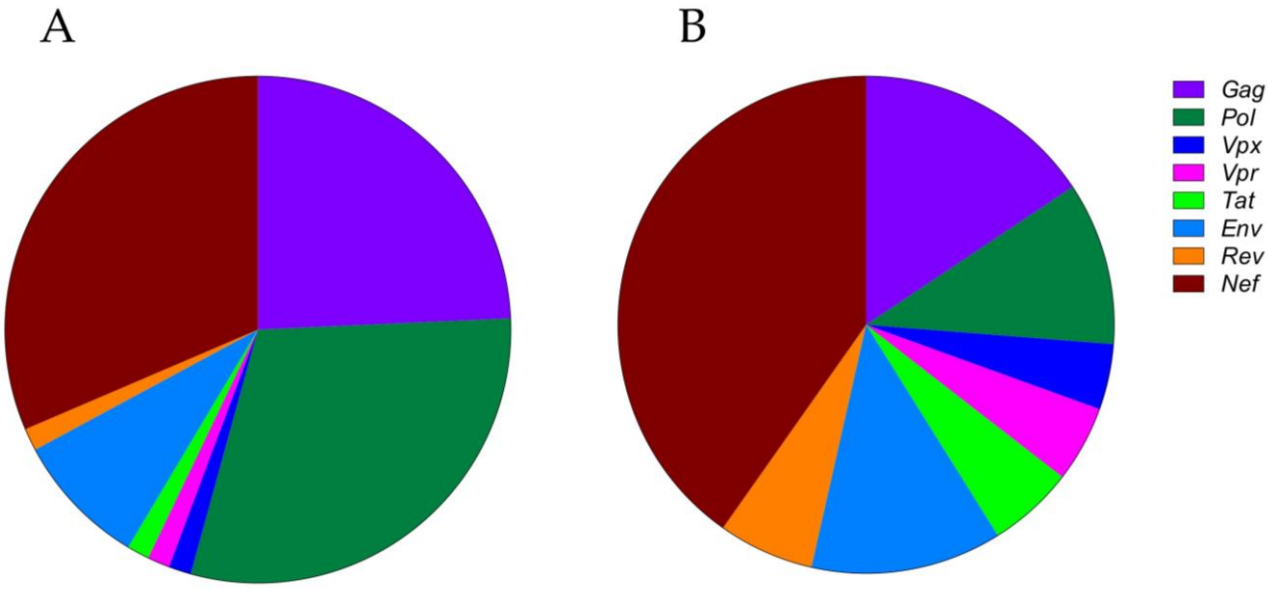



Figure 4.10: Distribution of SNPs partitioned by gene.
The proportion of all SNPs per gene (A) and the number of SNPs as a function of the gene length (B) are shown.

Synonymous and non-synonymous SNPs were partitioned by gene (**Table 4.3**). Some caution must be applied when looking at a small number of SNPs over an unknown time period with variable selection pressures caused by different HIV propagation methods. However, there is some evidence of overall positive selection of HIV-2 ROD in culture (dN/dS = 1.6) with *pol, nef* and *env* all showing evidence of positive selection. The only gene that appeared to be under purifying selection was *gag*, suggesting that there may be selection for conservation of virion structure.

Table 4.3: dN/dS ratios partitioned by gene.

| | *Gag* | *Pol* | *Vif* | *Vpx* | *Vpr* | *Tat* | *Env* | *Rev* | *Nef* | *Overall* |
|---|---|---|---|---|---|---|---|---|---|---|
| All SNPs | 17 | 21 | 0 | 1 | 1 | 1 | 6 | 1 | 22 | 70 |
| dN | 8 | 17 | 0 | 0 | 0 | 0 | 4 | 1 | 12 | 43 |
| dS | 9 | 4 | 0 | 1 | 1 | 1 | 2 | 0 | 10 | 27 |
| dN/dS | 0.88 | 4.25 | N/A | N/A | N/A | 1 | 2 | 1 | 1.2 | 1.6 |

Total number of SNPs, synonymous SNPs (dS), non-synonymous SNPs (dN) and the dN/dS ratio were calculated for each gene.

## 4.3 Discussion

The first aim of this study was to assess the feasibility of using RNA-Seq to sequence a propagated HIV-2 reference strain and to assess the biases in the resulting data. Sequencing reads were generated using a single run of the Illumina MiSeq platform[364]. MiSeq is a 'desktop' sequencing platform that is inexpensive and relatively common in molecular biology institutes. Mean depth of coverage over the whole genome was in the region of 2000x, varying slightly with assembly method and the total percentage of reads mapped was between 2.9-3.4%. The high level of depth seen using a modified publically available library preparation kit shows that RNA-Seq is a powerful tool to assess the variation in cultured HIV-2 strains. Previous *in vitro* HIV evolution studies have relied on PCR amplification of the genomic region of interest followed by sequencing of the amplicon. This study shows that RNA-Seq offers a feasible method for generating high depth sequence data over the whole HIV-2 coding region from a cultured reference strain without the need for prior amplification or primer design. As has been seen in other studies using RNA-Seq, a random hexamer bias was observed, resulting from differential binding affinities of the random hexamer pool[356]. Although this bias introduces fluctuations in coverage, the entire coding region of HIV-2 ROD was sequenced to high depth and the biased base composition did not extend past the first 13bp of each read. Therefore, random hexamer priming of cDNA synthesis is preferable to reverse transcription using primers targeted to the region of interest, which relies on accurate prior sequence knowledge and may be biased by mutations in primer binding sites[347]. cDNA synthesis can also be achieved through oligo (dT) priming, however this is heavily biased towards the 3' end of the RNA molecule and would result in a lack of coverage in the middle of the HIV-2 genome, where many functionally important genes are located.

Depth of coverage showed a slight positive GC bias. This is a common feature of all Illumina sequencing studies and the extent of the bias varies between genomes sequenced[355]. In this study, the GC bias did not greatly impact the results. The HIV-2 genome has a moderate GC content (45%) and does not contain large GC-rich or

GC-poor stretches. Therefore, GC bias would not be expected to affect the results of subsequent RNA-Seq studies using HIV-2. It is also worth noting that GC bias is a feature of all Illumina sequencing runs and not a unique challenge when using RNA-Seq for library preparation rather than commonly used DNA kits, such as NexteraXT.

Depth of coverage was partitioned by genomic region to assess whether read mapping could be influenced by divergence from the reference genome. Depth of coverage was consistent with two exceptions. Shorter genes were more prone to fluctuations in depth of coverage, probably as a result of the underlying random hexamer bias and insert size. There was also a significant drop in coverage over the LTR regions of the genome, resulting in lower mean coverage for *nef*, which is partially located in the LTR. Whilst RNA-Seq provided high coverage over the majority of the coding region, studies aiming to sequence short genes or *nef* may be affected by the fluctuations in coverage and be better suited to alternative sequencing strategies.

The second aim of this study was to assess the performance of a panel of commonly used alignment tools. Overall, assemblies and depth of coverage were not markedly different between the aligners. Comparisons of whole genome assemblies showed only two major conflicts, and on both occasions the different solutions were plausible. Interestingly, in one conflict the two BWA aligners gave different solutions, showing that even when the assembly algorithm is similar, tools can produce different results. The question of a 'true' assembly remains hard to define as each algorithm contains many assumptions and cost matrices. This study shows that a multifaceted approach to genome assembly, using multiple aligners is the most appropriate for the analysis of HIV-2 NGS data.

Genome assemblies were used to quantify the divergence of HIV-2 ROD *in vitro* from the published reference genome[365]. There was some evidence of positive selection acting on HIV-2 ROD and non-synonymous mutations were seen in 6 of the 9 HIV-2 genes. Although the number of mutations was small in the context of the genome length, this study clearly shows that the assumption of homogeneity

between published reference and lab strains may not be a valid one. Care must be taken when comparing different HIV-2 isolates, especially in light of the non-synonymous mutations identified in *gag* and *nef*, which are known to be crucial in early HIV-2 infection[153,366]. Therefore, routine whole genome RNA sequencing of HIV isolate stocks would be beneficial to *in vitro* assays, ensuring that results are consistent with experimental set up, rather than pre-existing and unidentified mutations.

The data presented in this study show that RNA-Seq is a powerful tool for generating whole genome sequence data of high depth from cultured HIV-2 isolates without the need for prior amplification or target enrichment. Biases that must be accounted for include random hexamer and GC biases. Whilst RH bias is unique to RNA-Seq, the data presented in this study suggest that the resulting fluctuations in coverage should not prevent the assembly of whole genome sequences in future studies. A slight positive GC bias was also seen, which can be corrected for in analyses by comparing the mean GC content of a region of interest with the resulting coverage. The major drawback of using RNA-Seq for HIV-2 isolates is the paucity of coverage in the LTRs. Whether or not accurate high-depth sequencing of the LTR is needed is dependent on the study conducted and therefore alternative sequencing strategies may be needed for detailed analysis of the LTR. The absence of an initial PCR step removes prior PCR bias from the library preparation and although there is a PCR involved in the library enrichment step, biases can be corrected in downstream sequence analysis. The bias induced by PCR target enrichment is harder to correct for and may have a large effect on the resulting genetic heterogeneity of the sequence data. It is worth noting that the viral load of cultured HIV-2 samples is much higher than is seen in primary patient samples and the applicability of RNA-Seq to primary HIV-1 or HIV-2 samples has yet to be demonstrated. Therefore, this study demonstrates that RNA-Seq offers a powerful tool for generating whole genome sequences for HIV-2 isolates, avoiding many of the biases inherent in other methods and therefore more accurately capturing the genetic variation of the viral population.

# Chapter 5: Generating Whole Genome HIV-2 Sequences From Primary Patient Plasma Samples Without the Need for Prior Target Enrichment

## 5.1 Introduction

Next Generation Sequencing (NGS) of viral genomes without the need for prior target enrichment is one of the most promising applications of emerging sequencing technologies. As discussed in Chapter 4, target enrichment has several limitations such as a need for detailed *a priori* knowledge about the genetic variation of a population to allow robust primer design. Traditional target enrichment methods employing PCR may distort signals of variation, potentially limiting the conclusions that can be drawn from sequencing projects[345]. Therefore, library preparation methods that negate the need for enrichment are powerful tools in the study of rapidly evolving and genetically diverse populations, such as the RNA viruses.

Recent studies have demonstrated the feasibility of using RNA-Seq library preparation methods to generate whole genome sequence data from a range of RNA viruses. Hepatitis C Virus (HCV), an RNA virus of the *Flaviviridae* family, can cause cirrhosis of the liver and between 150-200 million people are thought to be living with chronic hepatitis C infection worldwide[367]. Ninomiya *et al* utilised RNA-Seq to generate nearly full-length genome sequences from two chronically infected HCV patients[368]. Libraries were prepared from 800μL of serum obtained from patients with log viral loads between 6.8-7.0 IU/mL. Using Illumina short read sequencing (76bp reads), the authors were able to generate sequence data from more than 99% of the coding region and the mean depth of coverage was between 50-70x for both experiments. Batty *et al* also used RNA-Seq to generate whole genome sequences for HCV and Norovirus[285]. HCV genome assemblies showed similar results to Ninomiya *et al*, and for the two patients studied depth of coverage was in the region of 10-300x following RNA extraction and shotgun sequencing from 100ng total patient RNA. The authors also studied Norovirus, a

virus with a single stranded RNA genome that is responsible for 'Winter Vomiting Disease' with outbreaks common in semi-enclosed environments such as hospitals and schools[369]. Norovirus is transmitted through faecally contaminated food and person-to-person contact and therefore tracing the source of Norovirus outbreaks through inference of relationships allows the identification of risk factors underlying large-scale outbreaks[370]. Batty *et al* showed that high-throughput RNA-Seq could be employed to generate whole genome sequences rapidly from multiple samples at a mean depth of 100x. Comparisons of coverage obtained using RNA-Seq and sequencing of overlapping amplicons showed that RNA-Seq provided more consistent coverage over the whole genome, as coverage was low at both ends of the amplicons, even when an amplicon overlap was incorporated into the experimental design. Additionally, from a total of 77 faecal samples sequenced, the failure rate for RNA-Seq was 1%, which was considerably lower than the failure rate for amplicon sequencing, where 71% of samples failed to generate a complete set of genome-spanning amplicons.

Zoonotic transfer of retroviruses from non-human primates (NHP) to human populations has been responsible for the emergence of both HIV-1 and HIV-2[371]. The importance of HIV-1 and HIV-2 as human pathogens highlights the need for detailed knowledge of the radiation and nature of NHP viruses in the context of continuing public health surveillance in remote locations. Whilst there is some debate over the age and origins of the SIVs, it is clear that there is a great level of underlying variation that has yet to be fully described[13]. Lauck *et al* used RNA-Seq to identify two novel and highly divergent SIVs in black-and-white colobus monkeys[372]. Analysis showed that the novel SIV genomes (termed SIVkcol-1 and SIVkcol-2) assembled *de novo* from Illumina MiSeq short reads (150bp), formed a lineage with SIVcol but were distinct species with no evidence of co-infection. Therefore, this study highlights the benefit of using RNA-Seq to sequence and characterise novel SIVs that may have been missed using traditional methods due to divergence from related strains and therefore failure of PCR amplification.

Whilst the feasibility of using RNA-Seq has been demonstrated for HCV and Norovirus, its application for other human RNA viruses such as HIV-1 and HIV-2 has yet to be fully described. Malboeuf *et al* employed a method of random hexamer primed linear amplification of HIV-1 RNA prior to NGS to circumvent the need for sequence-specific amplification[284]. Linear amplification was performed using single primer isothermal amplification (SPIA) technology (NuGEN) that uses a hybrid DNA-RNA primer followed by RNAse digestion of the RNA portion of the primer binding site in the amplified target, ensuring amplification only occurs from the initial template[373]. By combining viral isolation through ultra-centrifugation and linear amplification, the authors were able to generate high depth coverage (385-38725x), recapitulating 8/9 of the genes of the HIV-1 genome from inputs as low as 100 copies of the viral genome. Additionally, linear amplification of viral RNA has been used to generate whole genome sequences from samples taken during the recent West African Ebola Virus outbreak[374]. Analysis showed that the most recent outbreak has been driven by human-to-human transmission following a single zoonotic transfer, rather than being caused by continued exposure to the natural reservoir. Accurate information about the factors driving the Ebola virus outbreak gained from NGS can be used to inform public health interventions, aimed at curbing the number of new infections. Linear amplification circumvents many of the biases involved in PCR target enrichment and negates the reliance on *a priori* sequence knowledge. However, linear amplification adds a significant cost to library preparation (>£100/sample) and the inherent bias introduced by random hexamer priming of cDNA synthesis may be enhanced when random hexamers are used to prime amplification, something that studies using linear amplification have yet to address.

The ability to generate whole genome HIV-1 or HIV-2 sequences without prior amplification has yet to be demonstrated. Whilst previous studies have shown it is feasible for HCV and Norovirus, the lower viral loads typically associated with HIV-2 infection (normally at least 1-2 logs lower than seen in chronic HCV infection) mean a detailed study of the absolute limitations of applying this

technique to HIV is needed. Therefore this study aimed to apply RNA-Seq to primary patient plasma samples, collected from our cohort and subsequently stored at -80$^{o}$C, to address the limitations and biases involved in RNA-Seq. This study also aimed to characterise the genetic diversity over different genomic regions using data generated through low bias & target-enrichment free sequencing.

## 5.2 Methods

### 5.2.1 Sample handling and RNA extraction

Patient plasma samples were collected in 2010 by Dr Thushan De Silva, stored at -80°C in the MRC laboratories in The Gambia and transported to Oxford shortly after collection in a liquid Nitrogen dry shipper. All patients were ART-naïve and HIV-2 mono-infected. When working on tropical diseases such as HIV-2, usability of primary patient samples is often limited by collection methods and inadequate storage. Therefore, the importance of the diligent care taken in the collection and transportation of these samples from a rural site in Guinea-Bissau should not be under-estimated. As HIV-2 has an RNA genome, samples are extremely sensitive to degradation by repeated free-thaw cycles[375]. All samples used in this study had a detailed record of freeze-thaw conditions, and samples chosen for RNA-Seq had no recorded occurrences of previous freeze-thaw cycles. This condition was chosen to maximize the quality of viral RNA used in library preparation, ensuring that sensitivity assessments were due to the laboratory methods chosen, rather than prior degradation of viral RNA.

Viral RNA extraction was conducted as outlined in Chapter 2.3.4 of this thesis. An input volume of 500µl cell-free plasma was chosen for two reasons. The first is that previous RNA-Seq studies have used plasma input volumes in this region (Ninomiya *et al* - 800µl, Lauck *et al* – 1mL), suggesting that sufficient total RNA for library preparation should be obtainable from this starting volume. Secondly, access to, and volumes of, primary patient material are often extremely limited, especially in the context of tropical diseases in developing countries. Therefore, the volume of plasma needed for successful RNA-Seq should not exceed the typical volume obtained from a standard bleed of adult patients (40mL of whole blood giving ~ 10mL of plasma). Additional allowances must be made for concurrent assays that are likely to be performed on primary plasma samples (e.g. viral load quantification or sequence verification using PCR). Therefore, we sought to use an input volume of patient plasma that was both in line with previous studies and

what could be reasonably expected to be available for future studies on additional cohorts.

Pilot sequencing runs were used to generate an optimum multiplexing number of 6 samples per HiSeq run (data not shown). The patients chosen for the run were selected on a number of criteria. The first was the presence of more than 2ml of plasma, allowing for additional downstream sequencing reactions if needed. Secondly, the freeze-thaw history of the samples was assessed and patients were shortlisted based on the availability of samples with no previously recorded freeze-thaw cycles. Thirdly, patients were selected to represent a range of HIV-2 disease statuses. The rational behind this was to allow the identification of a cut-off sensitivity limit for this method, allowing future RNA-Seq studies to be focussed on patient groups with a high probability of successful sequencing. HIV-2 disease progression is highly correlated to viral load and patients are normally partitioned into three groups; those with a viral load below the limit of detection, those with a viral load of less than 10,000 copies/mL and patients with a viral load in excess of 10,000 copies/mL. Patients were selected to represent all 3 groups, patient TD006 had a viral load below the limit of detection, patient TD013 had a viral load between 100 and 10,000 copies/mL and four patients were chosen with a viral load in excess of 10,000 copies/mL. All patients were HIV-2 mono-infected and were still actively recruited into the cohort, allowing additional samples to be collected in the future if needed.

Total nucleic acid was extracted using the Viral UltraSens Kit (QIAGEN, UK) following control runs using the electron microscope counted HIV-2 strain NIH-Z that indicated this method offered the highest sensitivity. Linear acrylamide was used in place of carrier RNA as the nucleic acid co-precipitant (as outlined in Chapter 2.3.4) and final elution was in 12μL $H_2O$. Following extraction, DNA was removed from the samples to ensure that reads were derived solely from RNA. RNA concentration was estimated using the QuBit RNA assay (as outlined in Chapter 2.3.6). Libraries were prepared using the NEBNext Ultra RNA Library Prep Kit for Illumina and library quality was assessed with a DNA HS

chip/Bioanalyser assay (as outlined in Chapter 2.3.8). RNA input for some samples was lower than the recommended minimum of 10ng. However, the addition of a library quality control step between preparation and sequencing allowed the identification of successful preparation protocols, even in the absence of 'sufficient' starting material. Libraries prepared from 6 patient samples were pooled and run on a single lane of the Illumina HISeq platform, generating 2x100bp paired end reads (**Table 5.1**). Prior to downstream analysis, reads were trimmed according to quality (Q<30) and adaptor sequences were removed (as outlined in Chapter 2.2.10).

Table 5.1: RNA-Seq samples.

| Sample | Viral Load | Estimated [RNA] (ng/μL) | RNA in 5μL (ng) |
|---|---|---|---|
| TD003 | 82005 | 1.74 | 8.7 |
| TD006 | <100 | 1.53 | 7.65 |
| TD013 | 1632 | 6.8 | 34 |
| TD024 | 10560 | 0.46 | 2.3 |
| TD031 | 107183 | 0.62 | 3.1 |
| TD062 | 139519 | 0.57 | 2.85 |

Estimated RNA concentration is shown in ng/μL as well as the estimated total RNA used for library preparation. Samples were chosen on the basis of library quality (Chapter 2.3.6) and to represent a range of viral loads

## 5.3 Results

### 5.3.1 *De novo* genome assembly using VICUNA

Assembling reads from a diverse viral population is non-trivial and highly dependent on the reference sequence used in the assembly. Ideally, a robust and closely related reference sequence would be used in order to assess depth of coverage over the whole genome, assuming there are no genomic regions where the divergence between reference and reads is too great to allow accurate mapping. Using a reference sequence that is too genetically distant from the sampled population can lead to inaccurate alignments, introducing erroneous diversity and resulting in a substantial loss of data from unaligned reads[344].

In the absence of a candidate reference genome, there are several methods that may be employed. One of these is using existing sequence data to create a consensus reference generated from an alignment of multiple candidate reference genomes. Whilst this method works well when there is an expectation of large regions of shared homology between the references and sample of interest, it is much less robust in the context of highly diverse HIV-2 strains from multiple patients. This is evidenced by the great effort needed to design primers to amplify regions of the HIV-1 and HIV-2 genomes, a problem that is magnified in the context of NGS read mapping as a clear idea of the whole genome is needed instead of the identification of 20 or so conserved bases. *De novo* genome assembly circumvents the need for a reference genome by assembling sequencing reads based on homology to each other rather than homology to a reference (**Figure 5.1**)[376].

De novo assembly normally takes place under one of two major frameworks[377]. Overlap-layout-consensus methods divide reads into contigs on the basis of good suffix-prefix alignments[378]. Multiple sequence alignment is then performed to generate a consensus sequence for each contig. Contigs are orientated and placed on the genome build according to information gathered from read mates that are located in different contigs. De Bruijn graph-based methods construct a graph by

splitting the input reads into *k*mers (denoted as vertices) and creating directed edges between pairs of vertices where the last (*k*-1) bases of the source match the first (*k*-1) of the target vertex[379]. Graphs are collapsed to shorten paths with a unique exit and entry. Small bubbles present in the graph are normally attributed to sequencing errors and removed. The graph is traversed by finding a Eulerian path, incorporating data from paired reads, allowing a reconstruction of the genome sequence.

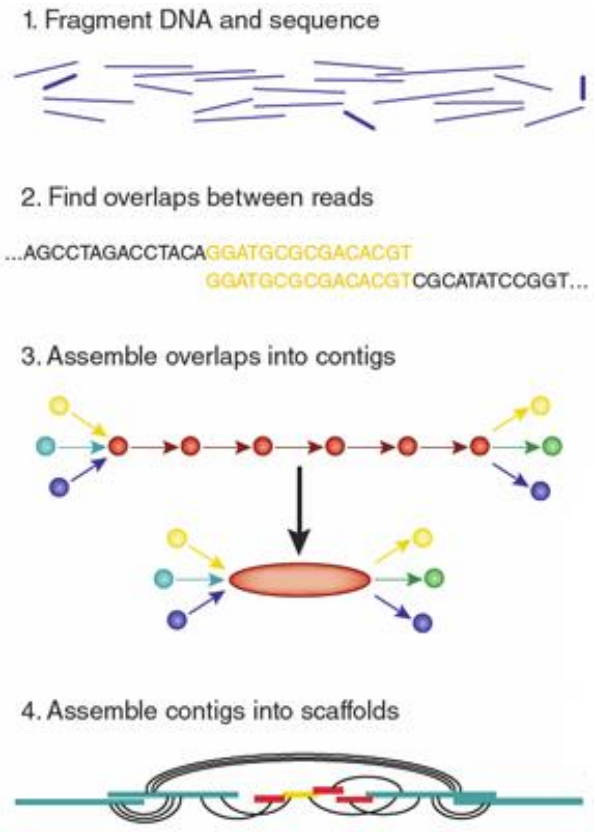



Figure 5.1: *De novo* genome assembly.
Schematic diagram showing *de novo* genome assembly (adapted from [376])

In order to assess whether the RNA-Seq libraries had successfully captured any reads originating from HIV-2, we utilised the *de novo* assembly algorithm VICUNA[353]. VICUNA uses an overlap-layout-consensus based method and is designed specifically for the assembly of variable coverage short read data from diverse populations, such as HIV. VICUNA assembly consists of 4 key steps and 2 optional steps (**Figure 5.2**). Assemblies of HIV-2 genomes were performed using all key and optional steps in the assembly. Reads were trimmed to remove any

adaptor sequences or low complexity reads that may be the result of library preparation artefacts. Due to the fact that libraries were prepared from total RNA without any prior target enrichment, we can assume that there is a high level of contamination from human RNA or co-infecting RNA viruses. Therefore we included the optional contamination removal step. VICUNA identifies contaminating reads by splitting a multiple sequence alignment (MSA) of reference sequences into multiple *k*mer bins and comparing the *k*mer composition of the target reads to the MSA *k*mer bin, identifying reads derived from the target sequence according to a sufficient number of similar *k*mers.

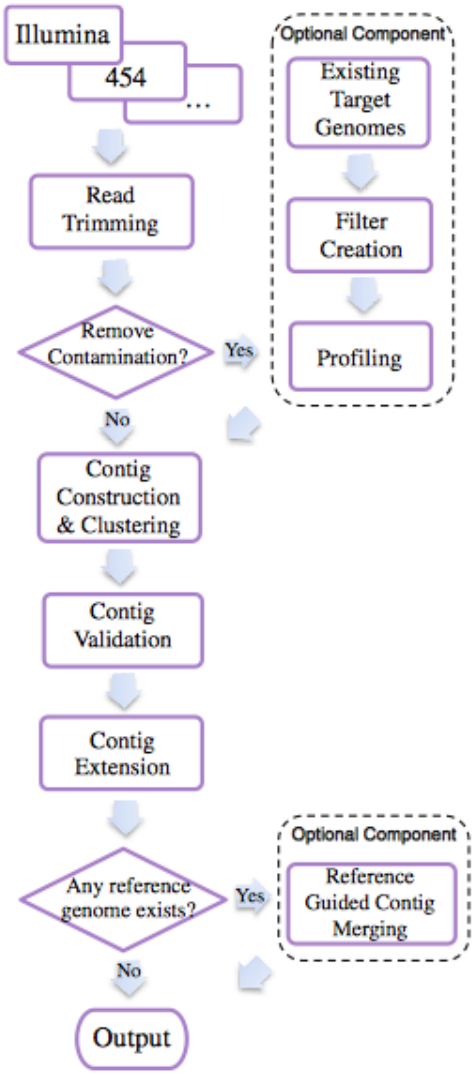



Figure 5.2: Schematic diagram of VICUNA assembly algorithm.
The key steps in the assembly as shown in squares and optional steps in diamonds. (Adapted from [353])

Following the removal of contaminating reads, VICUNA then looks for common min hashes in the *k*mers to identify reads with similarity. Hashing scores for each read are combined with a seed-based approach, which identifies good suffix-prefix overlaps between reads in a sliding window and reads are assembled into contigs. Contigs are then aligned to the consensus for verification and contigs that satisfy the validation are used to update the consensus. Contigs are extended using the longest contig as a seed and looking for good suffix-prefix with the contig that shares the largest number of read mates with the seed and the two are merged. This is repeated until no further *de novo* merges can be performed and contigs are then compared to the reference MSA, to verify merges and calculate final contig compositions.

In order to ensure that any failures in read mapping were due to an absence of reads rather than a failure of mapping, data from the HIV-2 reference strain CBL-20 was also included in the analysis. HIV-2 CBL-20 is a laboratory adapted HIV-2 strain that was grown in H9 cells, creating a positive control with a high viral titre. An RNA-Seq library was prepared from RNA extracted from cultured HIV-2 CBL-20 and sequenced on the MiSeq platform, creating 2x150bp paired end reads. Therefore HIV-2 CBL-20 was assumed to have an equivalent number of reads of viral origin as the HIV-2 ROD sample used in Chapter 4. Unlike HIV-2 ROD, CBL-20 does not have a published whole genome sequence and so VICUNA was used to generate a consensus for CBL-20 in parallel to the patient samples.

As VICUNA requires the initial input of an MSA of reference sequences, different inputs were used and the contigs generated and the level of genome capture in the *de novo* assemblies were assessed (**Table 5.2**). Eight different MSA inputs were tested. MCR35, MCN13 and ISY-SBL are whole genome sequences from The Gambia, CAM2 is an HIV-2 whole genome sequence from Guinea-Bissau and ROD is from Senegal. Additionally these 5 sequences were aligned and used as a single MSA for assembly (GM/GW). The last two conditions tested were an alignment of all available HIV-2 group A whole genome sequences ('All WG', n=20, AN in **Table 3.1**) and an alignment of all available HIV-2 group A sequences in LANL

('Pan-HIV2 A' n=4251, lengths = 82-10359bp). These two inputs were treated separately because although Pan-HIV2 A is more information-rich than All WG, the vast number of sequences of varying lengths and sampling locations means that the underlying structure of the HIV-2 group A genome is more obscured as the alignment contains many more indels than would be expected in a viral quasi-species from a single patient. Where *de novo* genome assembly was successful, the results obtained using each MSA were consistent between patients and therefore only patient TD024 will be discussed in detail (**Table 5.2**).

Assemblies using a single reference genome were relatively consistent, showing genome coverage in the range of 65-79% at a depth of 41-48x. However, the number of genes where the complete coding sequence was present in the *de novo* assembly was variable (3/9 – 7/9 genes) and generally low. This is most likely to be attributable to genomic regions where the genetic distance between sample and reference is too divergent, leading to the rejection of valid contigs in the final validation. Assemblies using the MSA of HIV-2 A sequences from The Gambia and Guinea-Bissau performed slightly better, however, there was still incomplete capture of the coding region of HIV-2 A, showing that there is not enough information in the MSA to capture the complete diversity in the primary patient samples. The MSAs of all HIV-2 A whole genomes and all HIV-2 A sequences showed identical results and successfully captured all 9 HIV-2 genes for 3 patients. Therefore, the indels in the Pan-HIV2 A alignment appear to not have affected the ability to generate the complete coding sequence. For all samples the Pan-HIV2 A MSA was used, as it is the most information-rich and should minimise data loss caused by genetic distance between reference and sample.

The overall success rate of HIV-2 RNA capture was assessed using the *de novo* assembly results from VICUNA (**Table 5.2**). For the positive control HIV-2 CBL-20, a total of 930072 reads were used to generate a set of contigs that captured all 9 genes of HIV-2 and 87% of the total genome. Therefore, VICUNA is a suitable tool for *de novo* generation of HIV-2 genomes from RNA-Seq data in the absence of a known reference sequence.

Table 5.2: Summary of *de novo* genome assembly.

**A**

| Sample | MSA | Contig lengths (bp) | Reads Assembled | Genome Coverage (%) | Depth | Genes Recovered |
|---|---|---|---|---|---|---|
| TD024 | MCR35 | 1713, 1203, 951, 795, 635, 545, 506, 471, 423, 397, 368, 327, 312 | 3288 | 77 | 41x | 3 |
| TD024 | ISYSBL | 1583, 1189, 1130, 647, 548, 454, 429, 433, 374, 362, 306 | 2700 | 53 | 42x | 5 |
| TD024 | ROD | 1353, 1335, 1309, 1297, 1183, 755, 567, 385, 351 | 3502 | 79 | 43x | 7 |
| TD024 | CAM2 | 1010, 979, 834, 795, 741, 713, 549, 539, 452, 443, 441, 312 | 3078 | 68 | 44x | 4 |
| TD024 | MCN13 | 2353, 1224, 945, 853, 537, 522, 435, 380, 354, 345, 327, 313 | 3287 | 65 | 48x | 5 |
| TD024 | GM/GW | 1540, 1335, 1200, 1199, 1192, 1110, 840, 800, 321 | 4340 | 87 | 47x | 8 |
| TD024 | All WG | 1546, 1335, 1200, 1038, 800, 772, 751, 611, 537, 523, 515, 380, 338, 321 | 4998 | 93 | 52x | 9 |
| TD024 | Pan-HIV2 A | 1546, 1335, 1200, 1038, 800, 772, 751, 611, 537, 523, 515, 380, 338, 321 | 4998 | 93 | 52x | 9 |

**B**

| Sample | Viral Copies | MSA | Contig lengths (bp) | Reads Assembled | Genome Recovered (%) | Genes Recovered | Length (bp) |
|---|---|---|---|---|---|---|---|
| CBL-20 | $>10^8$ | Pan-HIV2 A | 4708, 2327, 1427, 955, 493, 378, 372, 327, 324 | 930072 | 87 | 9 | 9885 |
| TD003 | 41002 | Pan-HIV2 A | 0 | 0 | 0 | 0 | 0 |
| TD006 | <50 | Pan-HIV2 A | 0 | 0 | 0 | 0 | 0 |
| TD013 | 816 | Pan-HIV2 A | 0 | 0 | 0 | 0 | 0 |
| TD024 | 5280 | Pan-HIV2 A | 1546, 1335, 1200, 1038, 800, 772, 751, 611, 537, 523, 515, 380, 338, 321 | 4998 | 93 | 9 | 9531 |
| TD031 | 53591 | Pan-HIV2 A | 3000, 2514, 2155, 1019, 683, 486, 352 | 9065 | 90 | 9 | 9397 |
| TD062 | 69759 | Pan-HIV2 A | 1648, 1421, 1308, 1173, 1136, 747, 756, 543, 514, 486, 481, 394, 359, 300 | 13304 | 87 | 9 | 9776 |

Contig lengths, reads assembled and depth of coverage are shown for each multiple sequence alignment input for patient TD024 (A). Results are shown for all RNA-Seq libraries, using the Pan HIV-2 A MSA (B).

### 5.3.2 Capture of the HIV-2 coding region by *de novo* genome assembly

VICUNA *de novo* genome assembly successfully generated contigs spanning the entire coding region of HIV-2 for CBL-20 and patient samples TD024, TD031 and TD062. In order to ascertain whether the consensus genome sequences were unique to the patient and to exclude cross-contamination during library preparation, an MCCT of RNA-Seq and reference whole genome sequences was generated using Bayesian MCMC inference under a GTR+I+G model of nucleotide substitution using 100,000,000 iterations with trees sampled every 10,000 generations. Following a burn-in corresponding to 10% of the samples, the MCCT was visualised and annotated (**Figure 5.3**).

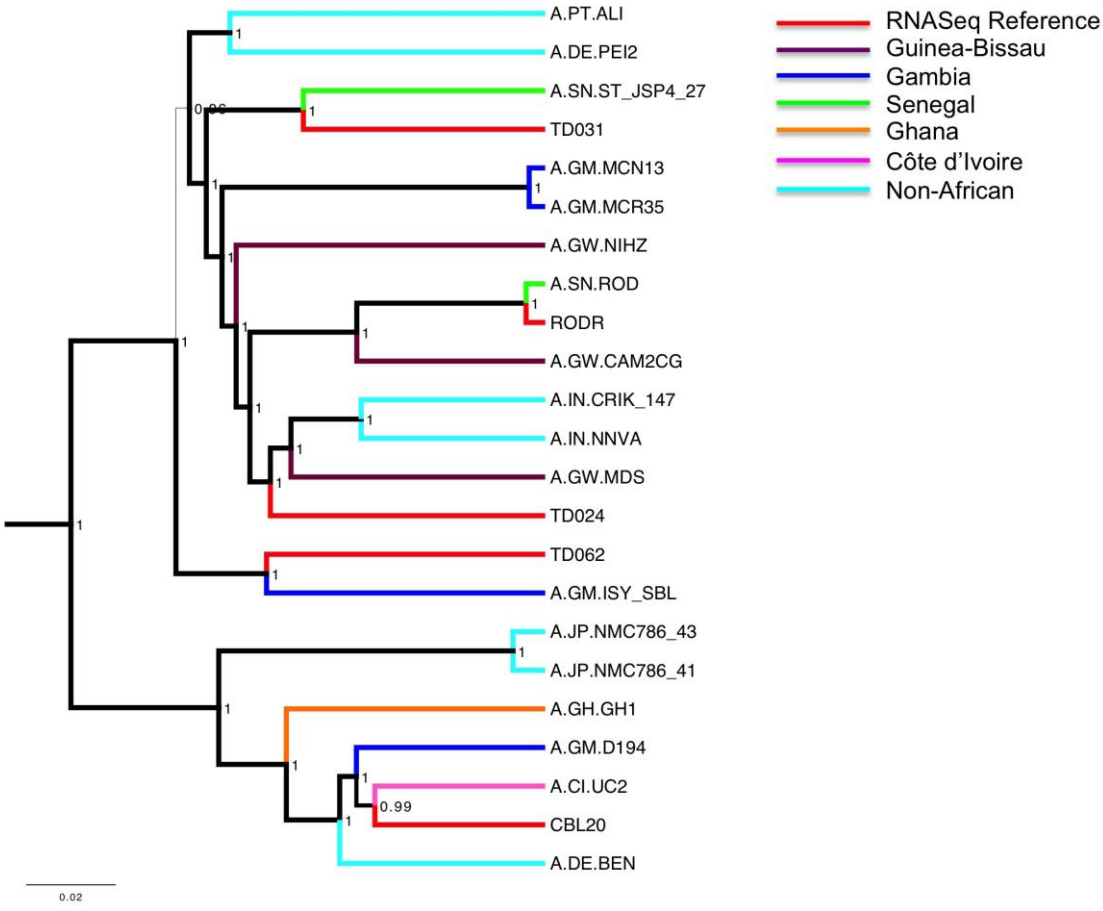



Figure 5.3: MCCT of RNA-Seq consensus and reference sequences. Braches are weighted by posterior probability and internal nodes are annotated with posterior probability. Sequences are coloured according to country or origin, excluding the RNA-Seq consensus sequences that are in red. Scale bar shows substitutions per site.

As was observed in chapter 3, HIV-2 A sequences from Caió do not form a monophyletic cluster and are interspersed within the general radiation of HIV-2 A sequences. The consensus sequence generated for HIV-2 ROD in chapter 4 was included in the analysis and was closely related to the reference sequence in LANL (posterior probability=1). None of the patient sequences clustered closely with CBL-20, ROD or each other, excluding the possibility of cross contamination during library preparation. Therefore, VICUNA has generated a unique consensus sequence for each patient, which is distinct from any other known reference sequences. This shows the benefit of generating a patient-specific consensus prior to read mapping, ensuring the genetic distance between reads and reference is minimised.

In order to assess the genomic regions captured by *de novo* assembly of reads, the consensus genomes were annotated with genes and full-length proviral LTRs, according to homology with annotated genomes in LANL (**Figure 5.4**).

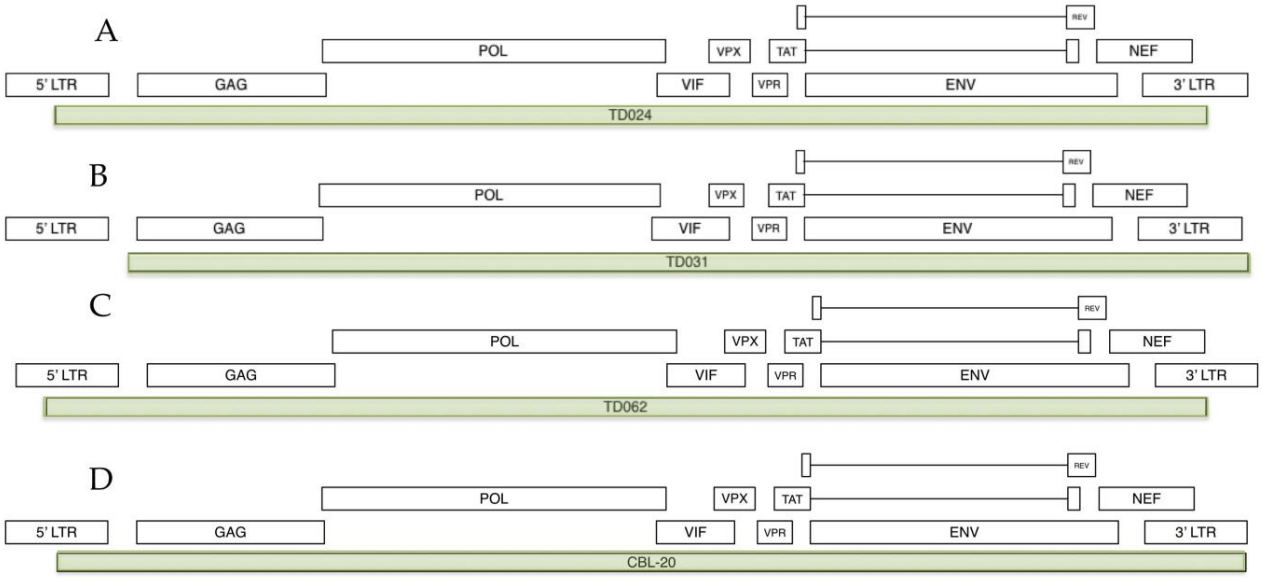



Figure 5.4: Annotated RNA-Seq genome consensus sequences.
*De novo* genome is shown in green for TD024 (A), TD031 (B), TD062 (C) and CBL-20 (D). Genes are shown as open boxed and the *tat* and *rev* introns are denoted as a straight line.

Consensus genome sequences covered the entire coding region of all 9 genes of HIV-2. Coverage at the distal ends of the genome in the LTRs was more variable, in line with what was observed in Chapter 4. The LTRs of HIV-2 are semi-palindromic and are made up of 3 distinct regions. The LTRs only exist as 'repeats'

in the context of viral DNA[380]. In the ssRNA genome, the 5′ LTR is expected to consist of the repeat (R) and untranslated 5′ (U5) regions and the 3′ LTR is expected to contain the untranslated 3′ (U3) and repeat (R) regions. The U3 region is approximately 500bp in length and comprises important regulators of transcription including the NF-κβ and TATA box binding domains as well as three tandemly arranged binding sites for the transcription factor Sp1. Additionally HIV-2 U3 contains the purine-rich PuB1 and PuB2 sites that are bound by members of the Ets family of proto-onco genes. The HIV-2 accessory gene *nef* overlaps with the U3 region, sharing approximately 400bp. The R region is approximately 90bp in length and is found in both the 5′ and 3′ LTRs, as a homologous inverted repeat, allowing transcription from the 5′ end of the genome in the absence of the TF binding sites of the U3 in viral ssRNA genomes. The junction between the U3 and R regions is the transcriptional start site of the HIV-2 genome. U5 is an 80bp region found only in the 5′ LTR of viral RNA genomes that contains Ap-3-like and DBF-1 promoter/enhancer sequences.

Analysis of the *de novo* genome sequences showed consistent misplacing of sequences in the LTR (**Table 5.3**).

Table 5.3: Summary of LTR assembly.

| Sample | 5' LTR | 3' LTR |
|---|---|---|
| TD024 | + 143bp U3 | - 29bp U3 - 90bp R |
| TD031 | - Entire LTR -177bp *Gag* Leader | + 80bp U5 |
| TD062 | + 319bp U3 | - 90bp R |
| CBL-20 | + 117bp U3 | + 80bp U5 |

Incomplete capture of the expected LTR region is shown by '-' and erroneous addition of LTR regions is shown by '+'. The size of each misplaced sequence fragment is shown in bp.

Complete LTR capture was achieved in all instances apart from sample TD031, which was missing ~420bp of the U5 region and 177bp of the *gag* leader sequence. In the other samples there was consistent erroneous placing of reads from U3 in the 5′ LTR. At the 3′ LTR there was a split between complete deletion of R and erroneous addition of U5.

Incorrect LTR assembly is a factor of the library insert size and the homology between the 5' and 3' LTRs. Reads mapping to the R region could be placed in either LTR and in this instance the information from the read mate is used to identify the true location of the read. Library fragments were between 200-350bp and so the vast majority of mates would also fall in the LTR, making correct positioning problematic. A larger insert size would circumvent this problem but in the context of a 10kb genome, there is a limit on the maximum insert size than can be used before there is significant loss of coverage from the distal regions. Complete LTR sequences were captured for 2 patients and CBL-20, therefore incorrect positioning of LTR regions is less problematic in light of the homology between the two LTRs.

### 5.3.2 Assessment of factors affecting successful capture of HIV-2 derived reads

*De novo* genome assembly was successful for the positive control CBL-20 and 3 of the 6 patient libraries (**Table 5.2B**). It remains possible that the *de novo* assembly strategy employed failed to capture reads that were from HIV-2 in the sequenced libraries. However, VICUNA is optimised for the assembly of RNA viral genomes and all known HIV-2 sequence data was used in the assignment of 'HIV-2 like' reads. It is more likely that the failure was due to a lack of sensitivity in the library preparation. Equally, even if there were reads derived from HIV-2 in the failed samples, they are obviously at too low a frequency to be biologically informative. In order to assess the limitations of using RNA-Seq on HIV-2 samples, the expected mass of viral RNA as a percentage of the total RNA was calculated. Although nucleic acid extraction methods never attain 100% efficiency it is assumed to be relatively constant under the same extraction conditions. HIV-2 virions contain two copies of a single stranded 10kb genome and the estimated weight per 10kb RNA molecule of approximately $5x10^{-18}$g was used to calculate the estimated mass of viral RNA, assuming 100% efficiency of extraction (**Table 5.4**).

Table 5.4: HIV-2 derived RNA in sequencing libraries.

| Sample | Viral Load (copies/mL) | Expected HIV-2 RNA (ng) | Total RNA (ng) | % HIV-2 RNA Expected |
|---|---|---|---|---|
| TD003 | 82005 | $2x10^{-4}$ | 8.7 | 0.0023 |
| TD006 | 100 | $2.5x10^{-7}$ | 7.65 | 0.0000033 |
| TD013 | 1632 | $4x10^{-6}$ | 34 | 0.000012 |
| TD024 | 10560 | $2.6x10^{-5}$ | 2.3 | 0.0011 |
| TD031 | 107183 | $2.7x10^{-4}$ | 3.1 | 0.0087 |
| TD062 | 139519 | $3.5x10^{-4}$ | 2.85 | 0.012 |

For each patient, the expected HIV-2 RNA, total input RNA and expected % representation of HIV-2 in the final library is shown

Successful RNA-Seq libraries all had an expected HIV-2 RNA% of more than 0.001%, suggesting this is an absolute cut-off for sequencing. However, sample TD003 had an estimated input of greater than 0.001% and still failed to show any HIV-2 derived reads. The reasons for the failure of sample TD003 remain unclear. Sample TD003 had a high plasma VL (82005 copies/mL) which was greater than sample TD024 for which RNA-Seq was successful. The other samples that failed had low viral loads (TD006 <100 copies/mL, TD013 1632 copies/mL) and so in these instances the failure can be attributed to a low VL, leading to an insufficient percentage of HIV-2 RNA in the final library preparation. The depth of coverage for the successful samples was significantly lower than the plasma VL, showing that RNA-Seq is not able to pick up every RNA molecule present in the sample. Possible explanations for the failure of sample TD003 include degradation of RNA during the preparation of plasma from whole blood or an un-noted freeze-thaw cycle leading to degradation of viral RNA. All plasma samples were extracted under the same experimental conditions and therefore RNA extraction is unlikely to have lead to a loss of HIV-2 RNA in this instance. The high RNA yield from sample TD013 (34ng in 5µl) can be attributed to incomplete removal of host cells during plasma preparation. All plasma samples were spun at 1,500xg for 15 minutes prior to RNA extraction, to ensure removal of cells and cell-derived debris. However, contamination of cell-derived RNA could have occurred prior to this point, resulting from lysis of host cells and the release of RNA molecules

contained in the cytoplasm. Therefore, when the minimum requirements of a VL in excess of 10,000 copies/mL and an estimation of at least 0.001% HIV-2 RNA in the total RNA input are met, RNA-Seq showed a success rate of 75%, capturing all genes of HIV-2 in all successful samples.

### 5.3.4 Error rate per sequencing cycle

Error rates were estimated for each sample using the GATK package. This gave an estimation of error rate per cycle for each of the 100 sequencing cycles for the first read set of each pair (**Figure 5.5**). Error rates are estimated using the number of mismatches seen during each cycle and the recalibrated quality score of each position.

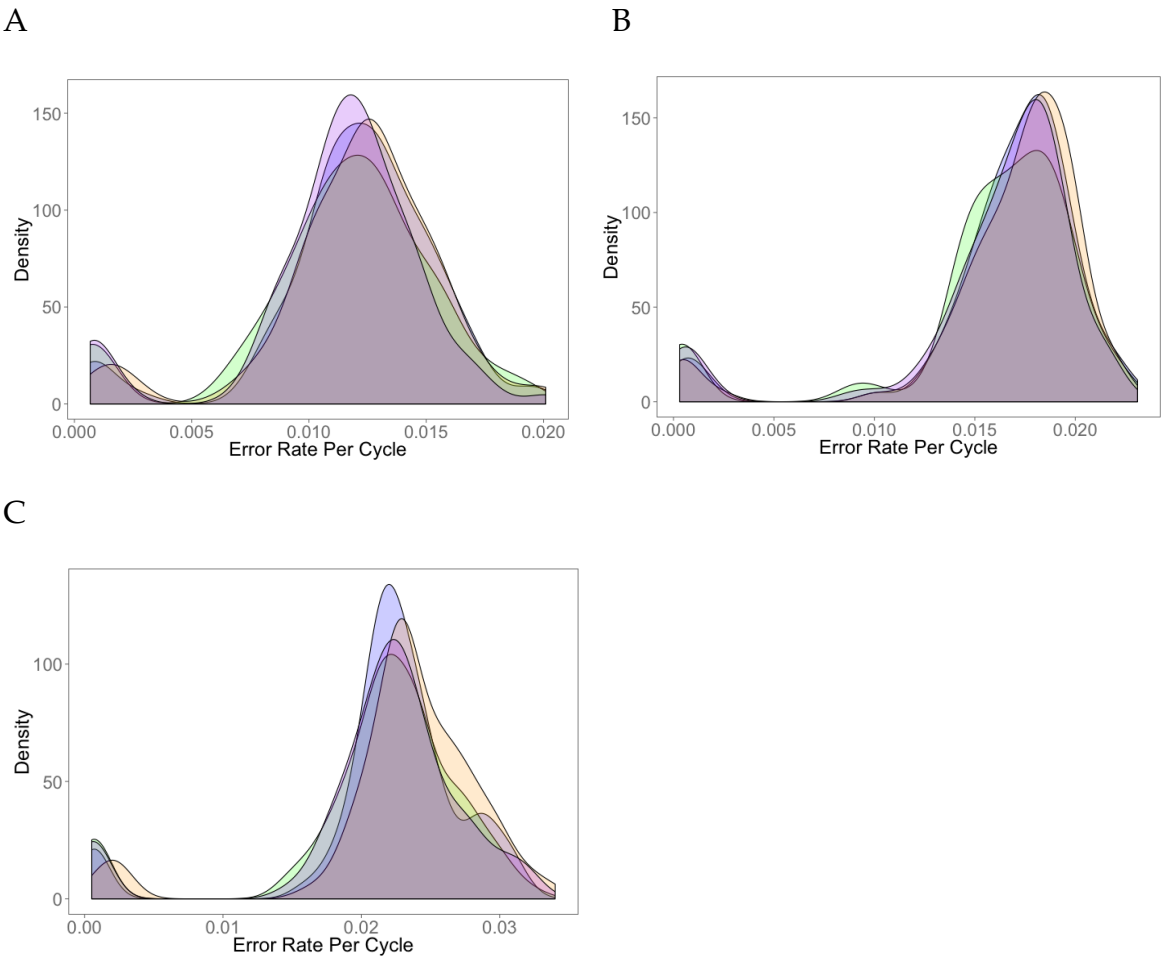



Figure 5.5: Error rates per sequencing cycle.
Error rates per cycle are plotted for samples TD024 (A), TD031 (B) and TD062 (C) for each aligner. Mean error rate per cycle was plotted against the probability density for all 100 cycles

Generally, error rates were consistent over cycles and there was no evidence of a drop in accuracy over the length of the read. Estimations of mean error rates were a little on the high side (TD024 = 1.3%, TD031=1.7%, TD062= 2.1%), which is caused by the underlying variability of HIV-2. GATK is not able to distinguish between low frequency variants in the viral population and sequencing errors. However, predicted mean error rates are in line with what would be expected and are informative when choosing a cut-off frequency for reliable SNP calling.

### 5.3.5 Identification of optical and PCR duplicates

As is described in Chapter 4, RNA-Seq library preparation involves library enrichment using PCR, which can introduce PCR duplicates. Assembled .bam files were assessed for optical and PCR duplicates using Picard. The percentage of reads identified as duplicates varied between alignments and BWA-SW assemblies had a noticeably high number of reads marked as duplicates when compared to the other assemblies (**Table 5.5**).

Table 5.5: PCR and optical duplicates in sequenced libraries.

| | % Reads Removed As Duplicates | | | |
|---|---|---|---|---|
| Sample | Bowtie2 | BWA-SW | GSNAP | NovoAlign |
| TD024 | 46.7 | 56.2 | 46.7 | 47.1 |
| TD031 | 39.8 | 58.6 | 39.3 | 39.3 |
| TD062 | 66.5 | 73.9 | 66.4 | 66.5 |

The percentage of reads identified as duplicates by Picard is shown for each aligner.

Sample TD062 had a much higher level of duplicates than would be expected. As explained in Chapter 4, this is probably due to the length of the HIV-2 genome, where reads derived from different templates are mis-assigned as duplicates due to their identical start positions. Overall, when using RNA-Seq to minimise biases introduced during library preparation it is important to account for PCR and optical duplicates, even if the algorithms for duplicate removal are overly stringent than is required, resulting in some data loss.

### 5.3.6 Assembly of reads to *de novo* patient specific references

Reads were assembled to the patient-specific consensus sequences using a panel of assembly tools (as outlined in Chapter 4). Overall, the mean depth and number of reads aligned was similar between the different aligners (**Table 5.6**). All aligners managed to assemble all 9 of the genes of HIV-2.

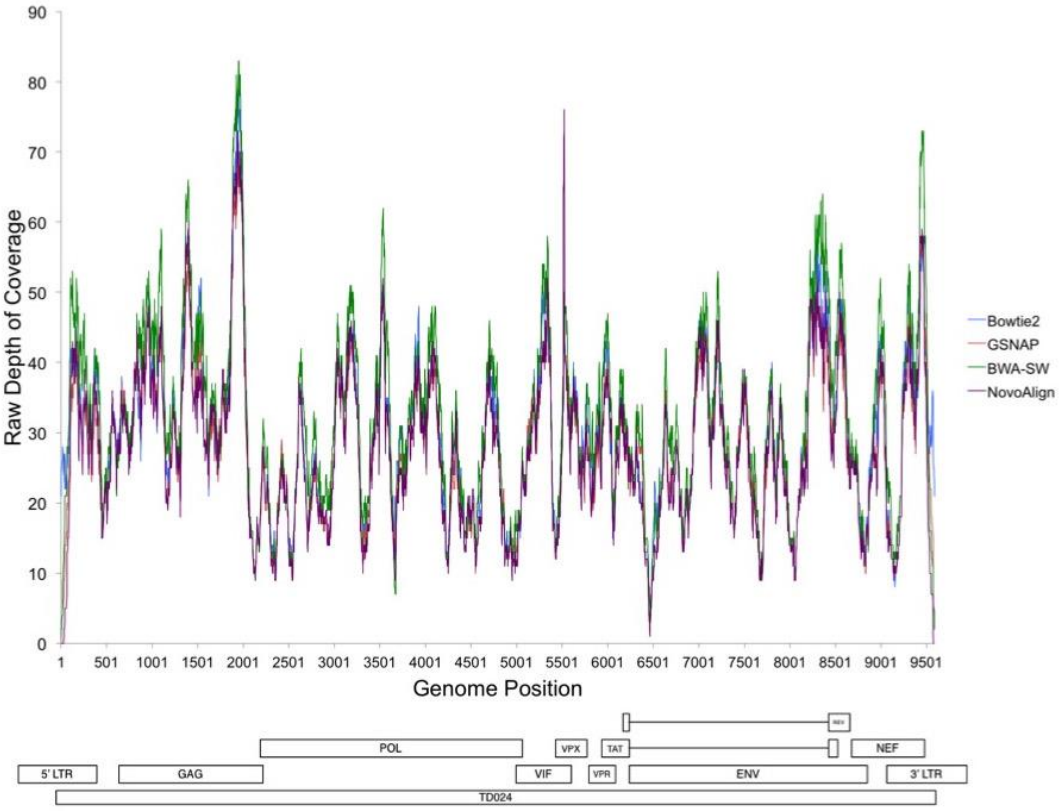
Table 5.6: Assembly of reads to patient-specific reference sequences.

| Sample | Aligner | Viral Load (Copies/mL) | Range of Depth | Mean Depth | Reads Aligning | % Reads Aligning | Genes Assembled |
|---|---|---|---|---|---|---|---|
| TD024 | Bowtie2 | 10560 | 12-79x | 28.53x | 3709 | 0.015035818 | 9 |
| TD024 | BWA-SW | 10560 | 14-83x | 31.90x | 3988 | 0.016166849 | 9 |
| TD024 | GSNAP | 10560 | 13-70x | 27.69x | 3426 | 0.013888572 | 9 |
| TD024 | NovoAlign | 10560 | 12-76x | 27.79x | 3463 | 0.013930326 | 9 |
| TD031 | Bowtie2 | 107183 | 19-163x | 62.33x | 7658 | 0.034577177 | 9 |
| TD031 | BWA-SW | 107183 | 12-188x | 67.23x | 8044 | 0.036320033 | 9 |
| TD031 | GSNAP | 107183 | 14-160x | 60.33x | 7172 | 0.032382804 | 9 |
| TD031 | NovoAlign | 107183 | 18-162x | 60.50x | 7267 | 0.032811745 | 9 |
| TD062 | Bowtie2 | 139519 | 12-115x | 50.01x | 6617 | 0.022561886 | 9 |
| TD062 | BWA-SW | 139519 | 9-99x | 59.61x | 7468 | 0.025463528 | 9 |
| TD062 | GSNAP | 139519 | 13-141x | 45.92x | 5658 | 0.019291998 | 9 |
| TD062 | NovoAlign | 139519 | 17-106x | 46.64x | 5751 | 0.019609099 | 9 |

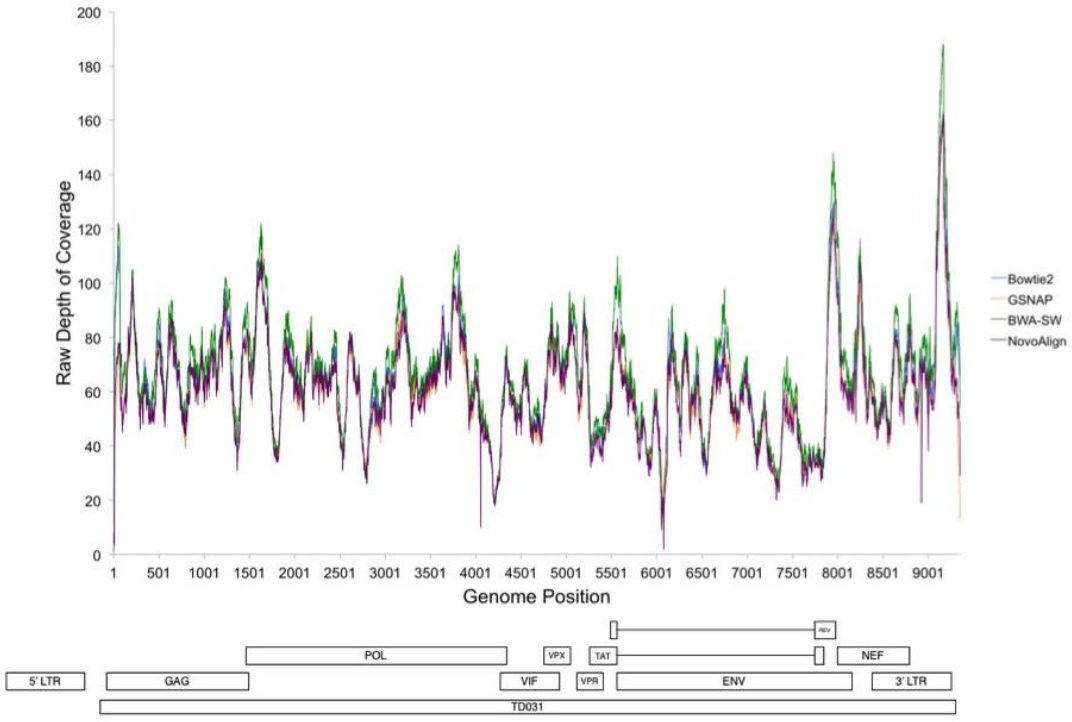
Results are shown for each aligner individually where *de novo* genome assembly was successful.

Mean depth of coverage was between 27.79x-67.23x, which is similar to the depth achieved in previous studies using RNA-Seq to sequence HCV. The overall number of reads aligning was in the region of 0.01-0.04%. Whilst this represents a high level of contamination from non-HIV-2 RNA, sufficient reads were aligned to generate whole genome sequences at relatively high depth. There was a slight increase in depth when samples TD031 and TD062 were compared with the lower VL sample TD024. However, the overly stringent removal of PCR duplicates probably leads to a loss of sequence data with increasing depth of coverage, leading to no clear correlation between VL and mean depth of coverage.

# A: TD024




# B: TD031



C: TD062

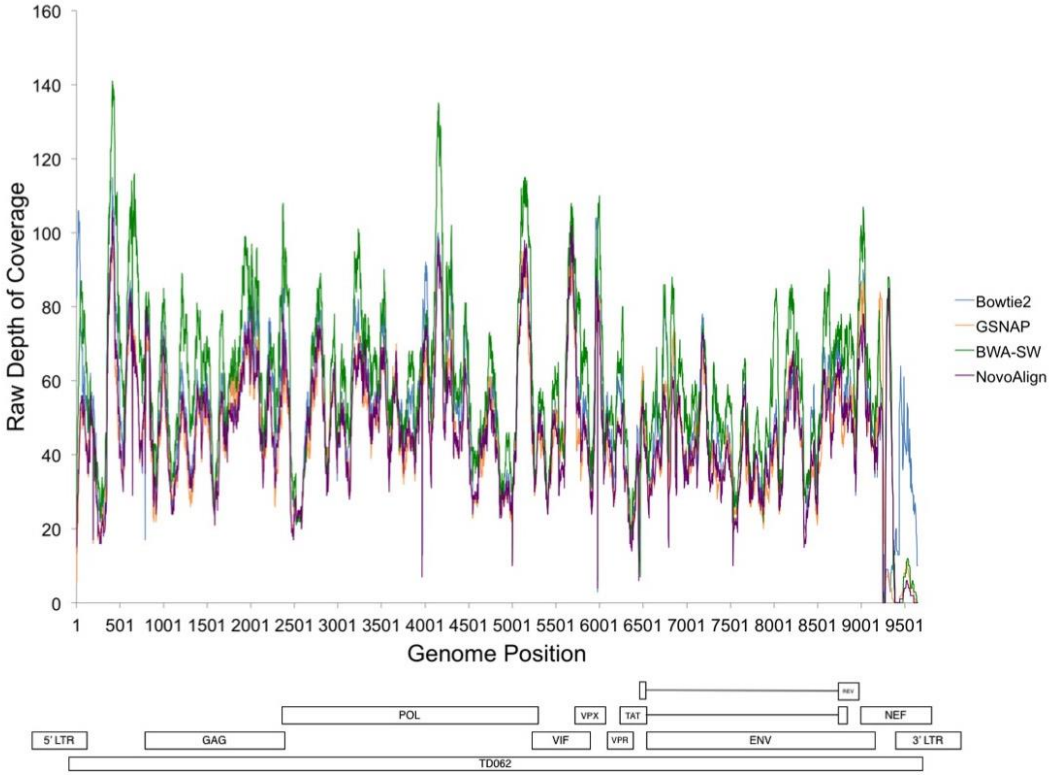



Figure 5.6 Depth of coverage across the HIV-2 genome.
Coverage is plotted per locus and coloured by aligner for samples TD024 (A), TD031 (B) and TD062 (C). Locations of genes are shown below each plot and the genomic region covered by the *de novo* consensus is shown as an open box annotated with the sample name.

Coverage over the length of the genome was visualised for all patients (**Figure 5.6**). Generally, the different aligners returned similar coverage plots, showing that the fluctuations are probably due to underlying biases in the library preparation or sequencing methods rather than being artefacts of assembly. For all patients there was no clear 'best' aligner. Therefore, the generation of a patient-specific consensus sequence *de novo* prior to read assembly allows robust and reproducible mapping of reads from diverse viral quasi-species.

### 5.3.7 Quantification of random hexamer bias

In order to assess the extent of the RH bias in patient samples, base composition per position in the total unmapped reads was visualised for each patient library in FASTQC (**Figure 5.7**). As has been previously described, there was evidence of a

random hexamer bias in the nucleotide composition of the first 13bp of each read[356]. The pattern of the bias was remarkably similar between the different samples, suggesting that there may be preferential binding to the same motifs in all samples. The effect of the bias did not extend past the first 13bp of each read and the nucleotide composition at each position stabilised after this point. The high GC content of the reads is indicative of contamination with structural RNA species such as rRNA and tRNA, which have a higher GC content, promoting stability of functionally important secondary structures[381].

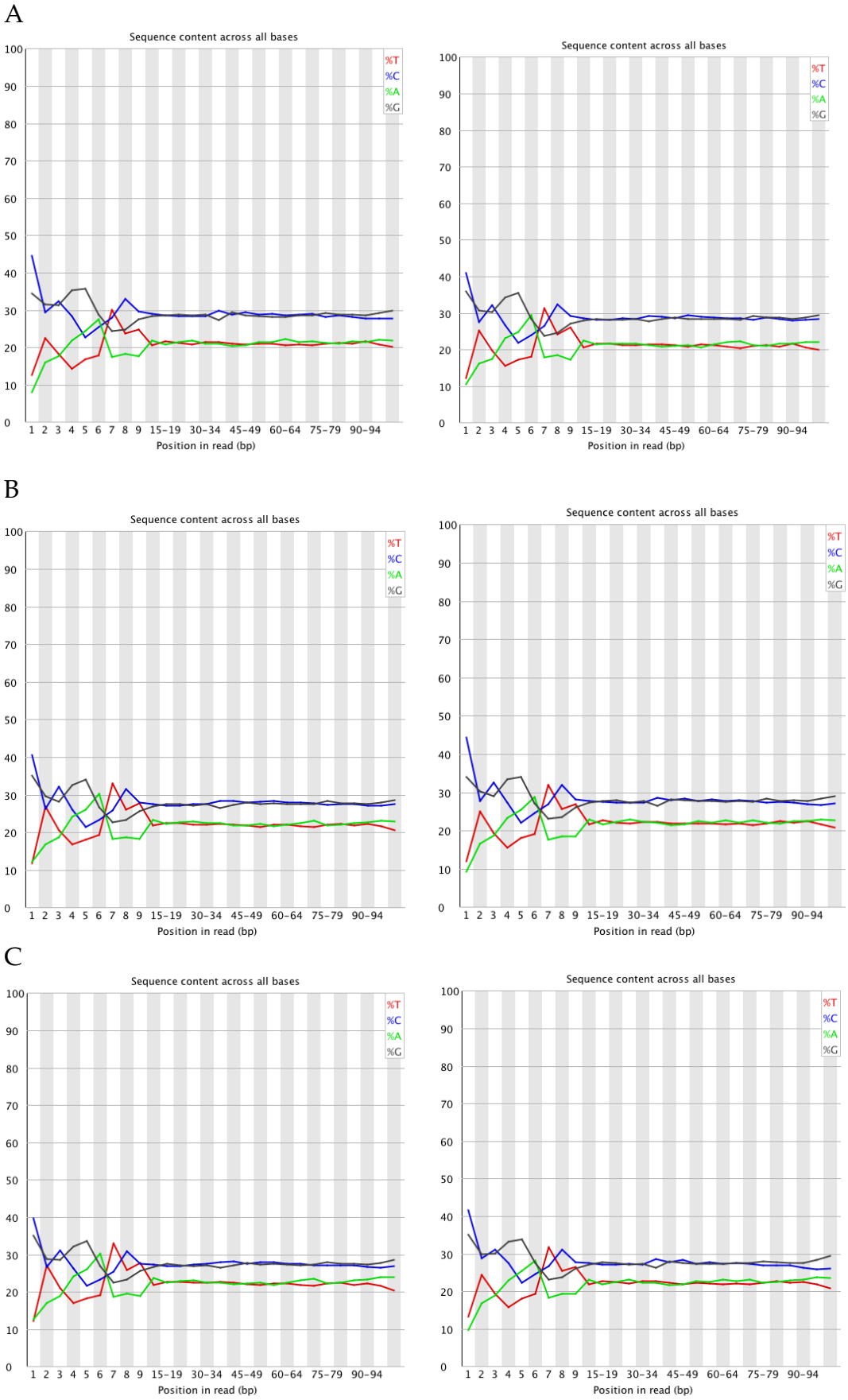



Figure 5.7: FastQC plots showing random hexamer bias
Mean nucleotide composition per read position in forward and reverse reads is shown for TD024 (A), TD031 (B) and TD062 (C).

### 5.3.8 Quantification of GC-bias in assembled reads

Assembled reads for each patient samples were assessed for the presence of a GC bias. As outlined in Chapter 4, GC bias was assessed using a 50bp-sliding window with a step size of 20bp along the length of the assembled genome. GC content in the window was plotted against mean depth of coverage in the window normalised by mean depth of coverage over the whole genome for each aligner (**Figure 5.8**).

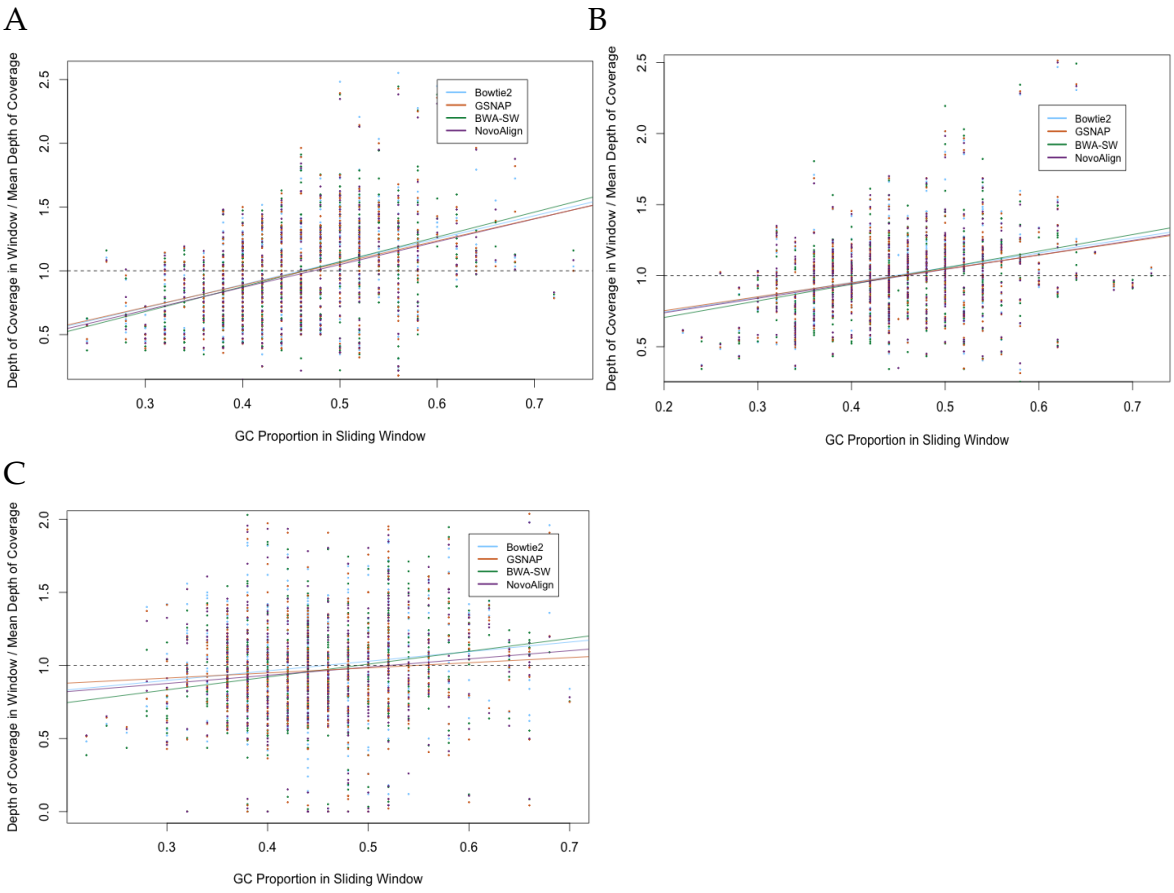



Figure 5.8: Visualisation of the GC-bias in assembled reads.
GC-bias was assessed in a 50bp-sliding window and points are coloured by aligner for TD024 (A), TD031 (B) and TD062 (C).

The slope of the line is an indicator of the extent of the GC bias and the deviation from the expectation in the presence of no GC bias (slope = 0). The magnitude of the slope was calculated for each sample and aligner (**Table 5.7**).

Table 5.7: Quantification of the GC-bias in assembled reads.

| | Bowtie2 | | BWA-SW | | GSNAP | | NovoAlign | |
|---|---|---|---|---|---|---|---|---|
| Sample | Slope | Intercept | Slope | Intercept | Slope | Intercept | Slope | Intercept |
| TD024 | 1.8 | 0.17 | 1.95 | 0.0955 | 1.74 | 0.19 | 1.79 | 0.15 |
| TD031 | 1.04 | 0.54 | 1.17 | 0.47 | 0.97 | 0.56 | 1.02 | 0.54 |
| TD062 | 0.66 | 0.7 | 0.88 | 0.57 | 0.35 | 0.81 | 0.56 | 0.71 |

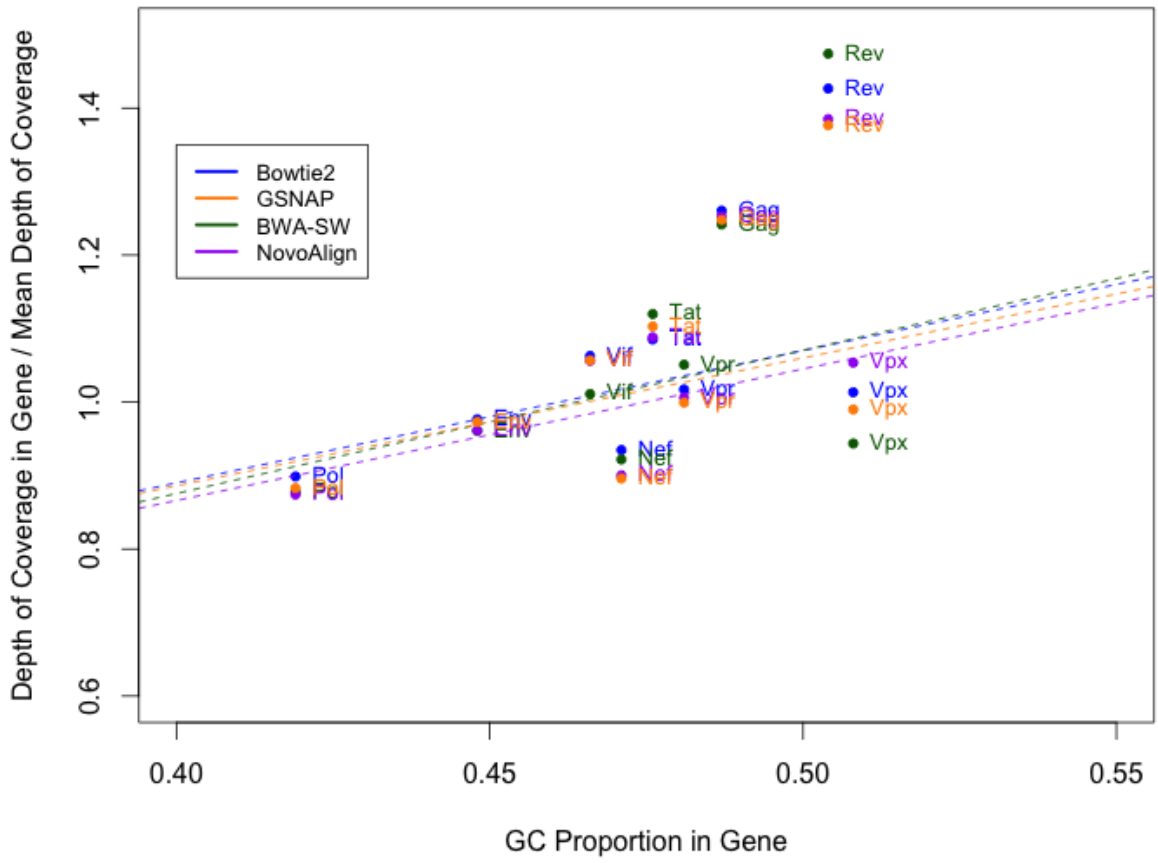
GC bias was quantified by assessing the slope of the linear regression line and values of slope and y-intercept are shown for TD024 (A), TD031 (B) and TD062 (C).

Samples TD024 and TD031 showed evidence of similar biases when each aligner was assessed individually. The effect of the bias was much stronger in sample TD024, with a slope of 1.74-1.95. This could lead to greater fluctuations in coverage as a result of the GC content of the region examined. Sample TD062 showed the lowest GC bias (0.35-0.88) and the greatest variability between different alignment tools. In all instances there was a positive GC bias, suggesting that GC rich regions will have a higher depth of coverage than GC poor regions. The results of these analyses showed GC bias in this study in keeping with reported biases from previous sequencing studies[355]. All samples had coverage of the complete coding sequencing, suggesting that the effect of the bias did not lead to regions of no assembly, as has been previously shown.

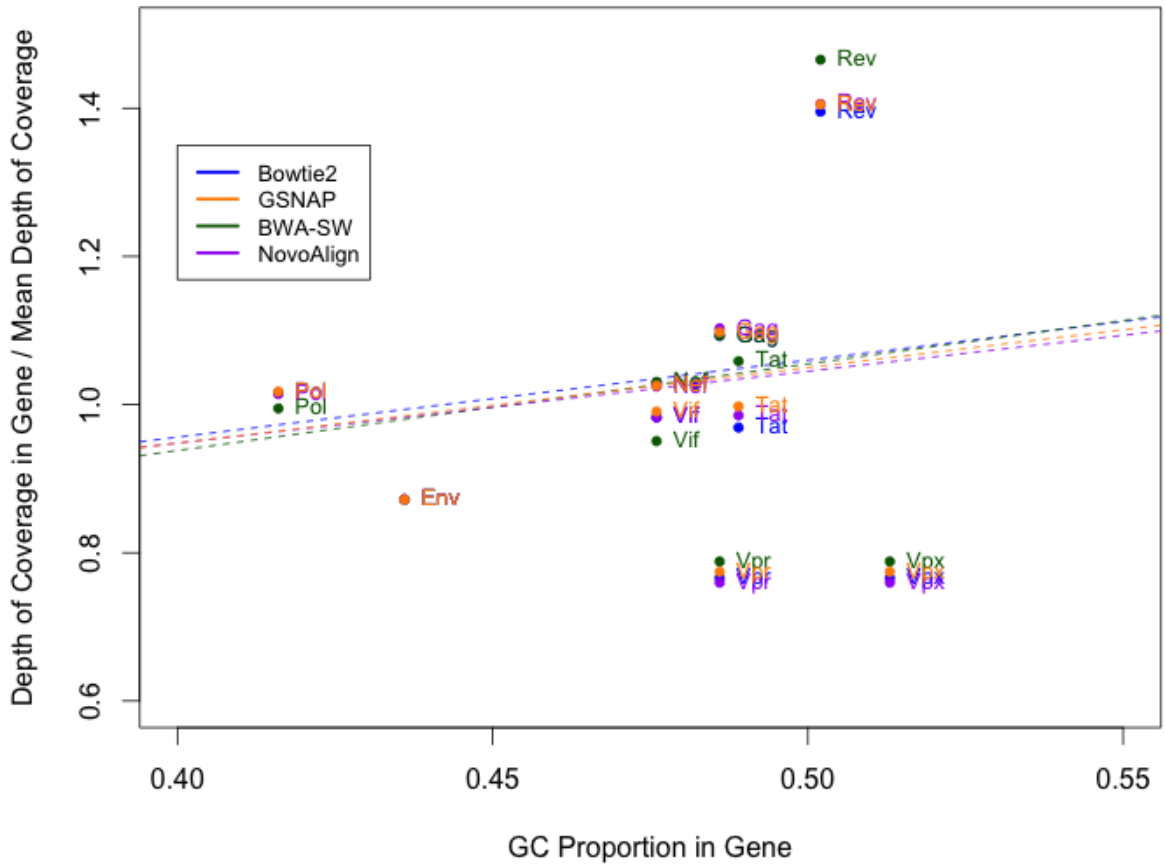
### 5.3.9 Depth of coverage as a function of genomic context

In order to assess whether genomic context had an effect on the number of reads aligned, the normalised mean depth of coverage for each gene was plotted as a function of the GC proportion (**Figure 5.9**). All aligners showed a similar pattern when coverage was partitioned by gene. As was observed in Chapter 4, more fluctuation in coverage is seen in shorter genes. There was better capture of *nef* than was seen in Chapter 4 and the most consistent genes were *pol* and *env*. Overall, there is no evidence that genomic context has an effect on depth of coverage and the differences between each patient are likely to be due to other factors such as RH bias, rather than divergence between reads and the population consensus sequence.

A


Depth of Coverage in Gene / Mean Depth of Coverage
GC Proportion in Gene
Bowtie2
GSNAP
BWA-SW
NovoAlign
0.6
0.8
1.0
1.2
1.4
0.40
0.45
0.50
0.55
Rev
Gag
Tat
Vif
Vpr
Vpx
Env
Nef
Pol

B


Depth of Coverage in Gene / Mean Depth of Coverage
GC Proportion in Gene
Bowtie2
GSNAP
BWA-SW
NovoAlign
0.6
0.8
1.0
1.2
1.4
0.40
0.45
0.50
0.55
Rev
Gag
Tat
Nef
Pol
Vif
Env
Vpr
Vpx

C

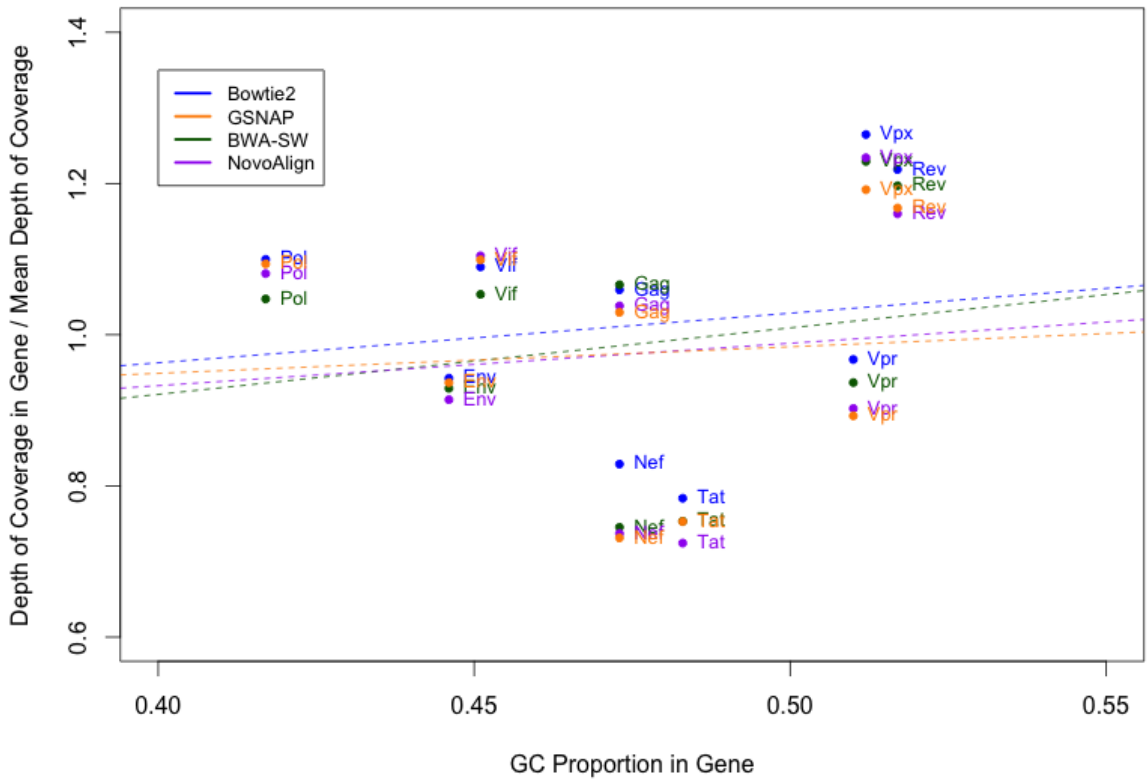



Figure 5.9: Depth of coverage partitioned by gene.
Depth of coverage was assessed for each gene individually for TD024 (A), TD031 (B) and TD062 (C). Points and lines are coloured by aligner. GC-bias is shown by the dashed lines.

### 5.3.10 Estimation of genetic diversity

One of the most powerful applications of RNA-Seq is the ability to estimate genetic diversity in the absence of a PCR bias introduced through prior target enrichment. In order to assess how nucleotide pairwise diversity ($\pi$) varied between patients and between genes in the same patient, read assemblies were used to estimate $\pi$ for each patient (**Figure 5.10**). Variants were filtered to remove variants at a frequency of less than 5%, ensuring that sequencing errors were unlikely to be called as SNPs. Sample TD062 had proviral *vpx* sequences available from 2010, which were generated by either clonal expansion and Sanger sequencing or MiSeq sequencing of PCR amplicons and these were also included in the analysis.

Raw estimates of $\pi$ were compared between patients over the whole genome and then for each gene individually. Estimates of overall nucleotide pairwise

diversity varied between patients (**Table 5.8**). Previous studies have shown that increased genetic diversity is a hallmark of disease progression in HIV-1 and in HIV-1 and HIV-2 dual infection increased genetic diversity is seen in rapid progressors. In this study, the highest diversity was seen in patient TD062, which was also the patient with the highest VL. The lowest diversity was seen in patient TD031, who had a similarly high VL of 107183 copies/mL. The intermediate estimate of diversity for patient TD024, who had the lowest VL by an order of magnitude, suggests that increased diversity is not necessarily a function of increasing VL in HIV-2 infection. The link between VL and disease progression in HIV-2 is not straightforward. Whilst the maintenance of an undetectable VL in the absence of treatment is indicative of a similar life expectancy to the uninfected population, and a VL of over 10,000 copies/mL is associated with progression to AIDS, the subtleties of VL measurements over 10,000 copies/mL still have to be investigated for HIV-2[87]. Our study suggests that diversity is not linked to VL, albeit in a small patient sample size.

Inclusion of the *vpx* diversity estimates made in Chapter 3 of this thesis allowed comparisons of three different sequencing strategies. Whilst the estimates derived in Chapter 3 were from amplified proviral DNA and estimates in this chapter are from RNA-Seq data that is derived from free plasma virions, the large discrepancy that was seen between the two methods means a comparison is still informative. The raw diversity estimate for *vpx* was highest for the clonal Sanger sequences, intermediate for RNA-Seq and lowest for MiSeq. As discussed in Chapter 3, the partitioning of variant frequencies into strict frequency bins could lead to an over-estimation of diversity from clonal Sanger sequences. This study further adds to the picture by suggesting that the raw diversity seen in a low-bias sequencing method is higher than that seen when NGS is applied to amplified genomic fragments. This is almost certainly due to the creation of PCR duplicates during the initial amplification, which lower the diversity present in the sample and are hard to correct for in post-sequencing analysis pipelines.

Partitioning of diversity by gene allowed comparisons of intra-patient diversity to be made. Raw diversity estimates were normalized using the estimate of diversity over the whole genome to ease comparison between the different samples. As would be expected from studies on HIV-1, the highest diversity was seen in *env* and *nef* for 2 of the 3 patients[382],[76]. *Env* contains 5 variable and constant regions and mutations in the variable regions are associated with Nab escape (although the evidence for Nab escape in HIV-2 is limited). *Nef* is involved in evasion of the host immune response through the down regulation of TCR and HLA expression and is therefore under a high selective pressure, which could lead to increased diversity in this gene[366]. Interestingly, all three patients showed a high level of diversity in *pol*. This is somewhat unexpected as *pol* is the most conserved gene of HIV and has been shown to be the least diverse genomic region in a NGS study of HIV-1 when sequences from different strains were compared[283]. However, studies have also shown the evolution of drug-resistance associated mutations in HIV-2 *pol* [383] and a recent study showed greater variation in *pol* compared to *env*, following vertical transmission of HIV-1[384]. The observation of high diversity in *pol* from three independent samples leads to the conclusion that *pol* maybe more diverse in HIV-2 than HIV-1. The lowest diversity was seen in *vpx* and *rev*. As discussed in Chapter 3 of this thesis, more sites in *vpx* are under purifying selection than positive selection in viraemic patients, which may explain the low diversity in *vpx* when compared to other genes.

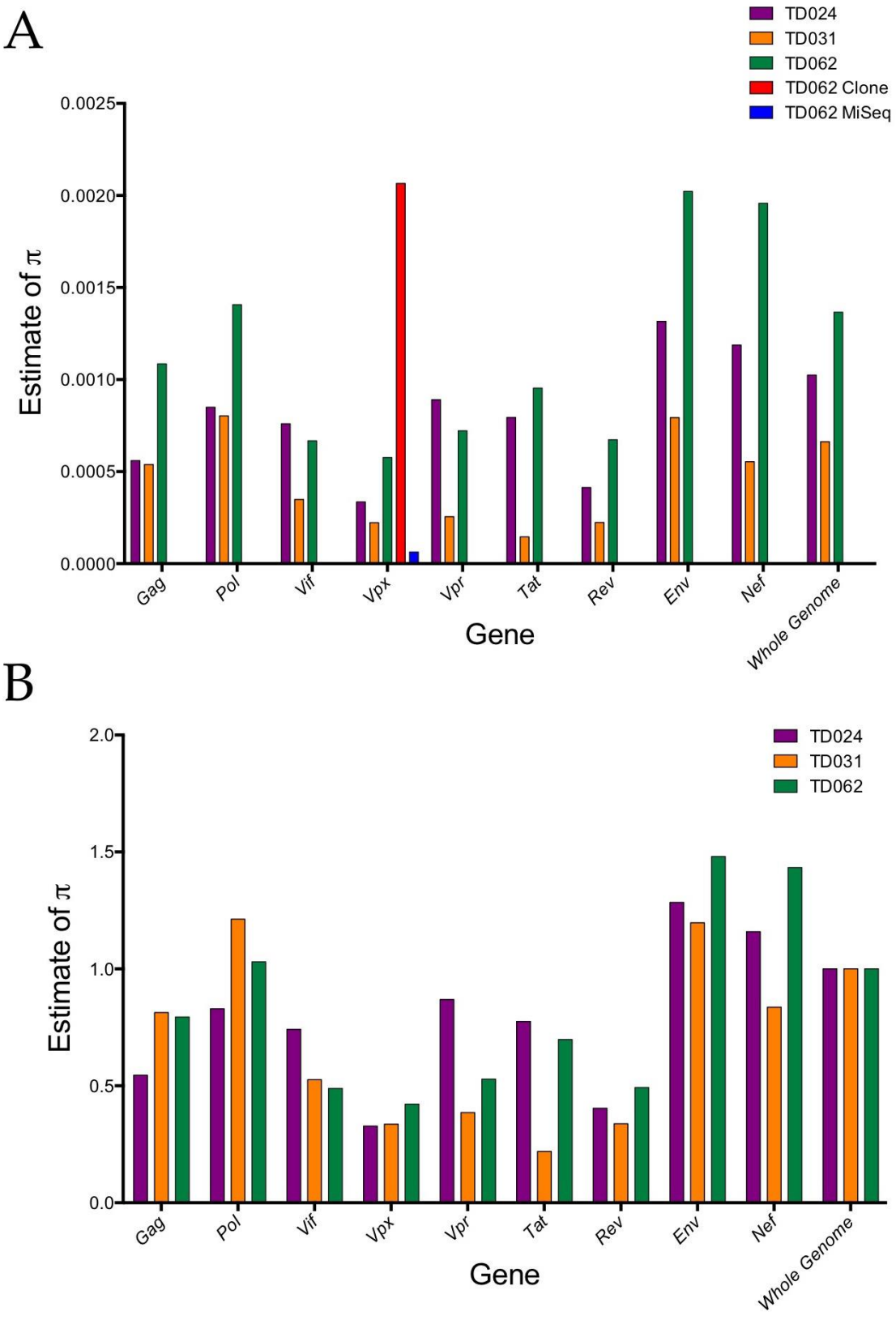



Figure 5.10: Gene-specific estimates of nucleotide pairwise diversity ($\pi$). Estimates of $\pi$ were made for each gene and coloured according the patient (A) and estimates were normalised to the genome wide average for comparison (B).

Table 5.8: Raw and normalised estimates of nucleotide pairwise diversity.

| Patient | π (raw) | π (normalised) |
|---|---|---|
| TD024 | | |
| Gag | 0.00055923 | 0.54577128 |
| Pol | 0.00084959 | 0.829143326 |
| Vif | 0.00075958 | 0.741299553 |
| Vpx | 0.00033575 | 0.327669666 |
| Vpr | 0.00089062 | 0.869185876 |
| Tat | 0.00079436 | 0.775242519 |
| Rev | 0.00041412 | 0.404153573 |
| Env | 0.00131616 | 1.28448461 |
| Nef | 0.0011875 | 1.158921008 |
| Whole Genome | 0.00102466 | 1 |
| TD031 | | |
| Gag | 0.00053886 | 0.813115843 |
| Pol | 0.00080342 | 1.21232515 |
| Vif | 0.00034875 | 0.526248284 |
| Vpx | 0.00022261 | 0.335908618 |
| Vpr | 0.00025555 | 0.385613617 |
| Tat | 0.00014548 | 0.219522868 |
| Rev | 0.00022374 | 0.337613738 |
| Env | 0.00079342 | 1.197235593 |
| Nef | 0.000554 | 0.835961431 |
| Whole Genome | 0.00066271 | 1 |
| TD062 | | |
| Gag | 0.00108515 | 0.794434602 |
| Pol | 0.00140708 | 1.030118453 |
| Vif | 0.00066743 | 0.488623219 |
| Vpx | 0.00057661 | 0.422134208 |
| Vpr | 0.00072223 | 0.528742112 |
| Tat | 0.00095343 | 0.698002841 |
| Rev | 0.00067348 | 0.493052403 |
| Env | 0.0020225 | 1.480665329 |
| Nef | 0.00195765 | 1.433188866 |
| Whole Genome | 0.00136594 | 1 |

## 5.4 Discussion

The primary aim of this study was assess the feasibility of using shotgun RNA sequencing (RNA-Seq) to generate whole genome sequences from primary patient plasma samples taken from HIV-2 infected members of the Caió community cohort. RNA-Seq has been shown to be a powerful tool for the study of RNA viruses such as HCV and Norovirus[285]. However, the application of this technology to HIV had not previously been demonstrated[344]. The first part of this chapter is concerned with the absolute limits of using RNA-Seq on plasma samples from HIV-2 mono-infected individuals. The data presented shows an absolute cut off value of a viral load of >10,000 copies/mL and a total HIV-2 RNA representation of at least 0.001% of the total RNA input (by estimated mass). Using these criteria, this study shows a success rate of 75% over 4 patient samples, demonstrating the successful generation of whole genome HIV-2 sequences from a total starting volume of 500µl patient plasma, without the need for prior target amplification. The ability to generate whole genome HIV-2 sequences without the need for prior sequence knowledge or target amplification shows that RNA-Seq is a powerful technique when applied to appropriate patient material.

*De novo* genome assembly using VICUNA allowed the generation of patient-specific reference sequences prior to read assembly[353]. The assessment of this tool on a range of input multiple sequence alignments (MSAs) demonstrates how prior sequence knowledge can affect the ability to recover complete genome sequences, even when they may be present in the library. Phylogenetic analysis of the resulting consensus sequences showed that patient-specific sequences were scattered throughout the radiation of HIV-2 group A whole genome sequences. Inclusion of the consensus sequence generated for HIV-2 ROD and the published reference sequence demonstrated how closely related a reference might need to be in order to be reliably used for read mapping. Additionally, phylogenetic analysis allowed the exclusion of cross-contamination during library preparation, showing that sequences generated were indeed patient-specific and derived from free

plasma virions in each sample. The divergence between patient-specific *de novo* consensus sequences and published HIV-2 genomes shows the importance of including an initial *de novo* assembly step, as the identification of an appropriate reference for all three patients is not possible, even when the genome sequences are known. Another method for read assembly in the presence of high levels of host RNA contamination is known as 'digital subtraction'[385]. This method first aligns reads to the reference human genome and then removes these reads prior to assembly. Whilst digital subtraction would remove a large proportion of the contaminating reads, co-infection with other RNA viruses and unassembled reads of human origin could lead to contaminants remaining post subtraction. This study showed that the initial contamination removal step in VICUNA negates the need for digital subtraction and allows robust identification of HIV-2 derived reads. Assembled reads were converted into .fasta format and a search was conducted in the BLAST nucleotide database, which showed that all assembled reads were identified as HIV-2 with high probability (data not shown).

As was described in Chapter 4, all assemblies were conducted using a panel of 4 commonly used alignment tools. There was no significant difference between the performance of different aligners, suggesting that read mapping to the *de novo* consensus is robust and repeatable, independent of the aligner used. Mean depth of coverage was in keeping with previous studies using RNA-Seq in spite of the lower viral loads seen in HIV-2[368]. Mean depth ranged from 27.69x-67.23x, generated by assembly of 0.013-0.036% of the total reads. Although the overall percentage of reads aligning is very small, Illumina platforms allow a large amount of sequence data to be generated at a low cost/base compared to other sequencing platforms[281]. All 6 samples were multiplexed on a single lane of the Illumina HiSeq platform and although HiSeq machines are less common than MiSeq, they are commercially accessible for most groups. The depth of coverage was lower than would be seen following target amplification. However, the absence of prior amplification and reduction in number of amplification cycles samples go through before sequencing leads to a significant reduction in PCR

bias[348]. PCR bias is extremely difficult to detect and correct for in post-sequencing analysis and so RNA-Seq offers the most low-bias form of HIV sequencing currently available.

Analysis of the biases involved in RNA-Seq showed the presence of both a GC and random hexamer bias. As has been seen in previous studies and Chapter 4 of this thesis, the random hexamer bias can be seen in the biased nucleotide composition of the first 13bp of each read but does not persist into the rest of the read. RH bias is expected to be a contributing factor in the non-uniform coverage seen over the genome post-assembly. GC bias was shown to vary between patient and in all cases was positive, leading to a modest increase in depth over GC-rich regions. However, the moderate GC content of the HIV-2 genome and the lack of large CpG islands means this is not a major consideration in this study and did not result in regions of no coverage. GC bias may be a more important factor to consider when genomes with a greater GC skew are considered.

All genes of the HIV-2 genome were sequenced to a good depth of coverage. The major region where coverage suffered was in the 3' and 5' LTR. As was discussed, the reasons for this are two-fold: 1.) There is a known drop of coverage at the ends of RNA molecules in RNA-Seq 2.) Homology between the LTRs leads to incorrect read mapping[380]. In 2 of the samples the complete sequence of the LTR was recovered, albeit in the wrong genomic location. For sample TD031, part of the U5 region and the *gag* leader sequence was missing. The importance of the sequence of the LTR depends on the nature of the study and as this study aimed to assess the diversity over the genes of HIV-2, it was not a major obstacle. Error rates per cycle were also assessed and were slightly higher than values previously reported for Illumina platforms (1.3-2.1%)[361]. These error rates are probably over-inflated due to the variable nature of HIV-2 quasi-species. However, the estimates of error rate did not vary significantly over the length of the read and therefore were used to define a SNP cut-off frequency of 5% in diversity analyses.

Whole genome assemblies were used to compare the nucleotide pairwise diversity ($\pi$) across the genome and for each gene. Mean $\pi$ estimates were similar

between patients and there was no correlation between π and VL. When the estimates of π were made for each gene individually, clear differences were seen. The most diverse genes were *nef* and *env*, which have previously been shown to be the most diverse genes in HIV-1. Interestingly, a high level of diversity was seen in *pol* in all three samples. Higher diversity has been described for *pol* when compared to *gag* following HIV-1 vertical transmission and this study also shows a high level of diversity in *pol* for HIV-2[384]. The use of RNA-Seq to generate these sequences has removed many of the biases typically involved in HIV sequencing and therefore the estimates of π presented in this study are accurate representations of the true diversity in the viral population. A low level of diversity was seen in *vpx*, which may be linked to the signals of purifying selection observed in Chapter 3, suggesting strong functional constraint acting on *vpx* in viraemic patients. However, signals of positive selection were also observed in viraemic patients, suggesting HIV-2 is not under global negative selection.

This study therefore demonstrates the feasibility of using RNA-Seq to generate low-bias whole genome HIV-2 sequences and describes in detail the biases involved and methods of down-stream correction. The major limitation of RNA-Seq remains the relatively high viral load needed for successful sequencing, limiting the application to viraemic HIV-2 infected patients in this cohort.

# Chapter 6: General Discussion

In the 30 years since the initial isolation and characterisation of HIV-1 and HIV-2, the two causative agents of AIDS, a great deal of progress has been made in understanding the complex host and viral factors which contribute to disease outcome following infection. Critical amongst these are the ability of the adaptive immune system to launch HIV-specific CTL responses, the ability of the virus to switch co-receptor usage, the restriction on viral replication inferred by host restriction factors and the effect that HIV-1 and HIV-2 dual infection may play in modulating HIV-1 disease progression. Whilst technology has progressed and new methods of investigating HIV infection have evolved, the viruses have caused global devastation, with the current estimates placing the total number of infections at more than 75 million[6]. There remains no effective HIV vaccine or cure, and whilst the development of successful combined ART has transformed HIV into a manageable condition for many, access to these medications is not universal and the cumulative toxicity of ART is being recognised as an increasingly prominent concern in ageing HIV-infected populations[386]. The different natural histories and global dispersion of HIV-1 and HIV-2 show the importance of fully investigating host and viral factors involved in disease progression. The reasons for the global dominance of HIV-1 group M remain unclear, however recent work has highlighted the role iatrogenic interventions and transport networks in Central Africa played in the early dissemination of this virus[371].

The presence of the accessory gene *vpx* in HIV-2 and SIVsmm allows degradation of the host restriction factor SAMHD1 and the absence of *vpx* in HIV-1 is the largest difference in genomic architecture between these two viruses. The effects of SAMHD1 antagonism on HIV-2 disease progression are unknown, however, the lack of an apparent alternative SAMHD1 antagonist in HIV-1 suggests SAMHD1 antagonism may be beneficial for the host, allowing early viral sensing by the innate immune system. The data presented in this thesis analysed *vpx* sequences from primary patient samples taken from a well-characterised and

longitudinally studied cohort. No statistically significant differences were seen in the frequencies of non-synonymous mutations when the patients were partitioned according to progression status. This observation is in line with previous studies and again concludes that overall variation in *vpx* does not appear to be related to HIV-2 disease progression. However, several differences were noted between the two groups. The most striking of these was the presence of significantly more evidence of positive selection in the viraemic patients. Signals of positive selection were almost entirely absent from non-viraemic patients. This study also observed more sites showing evidence of negative selection in viraemic patients, suggesting that *vpx* is under more selection in viraemic rather than non-viraemic patients. The source of selective pressure on *vpx* is almost certainly the host restriction factor SAMHD1 and *in vitro* analysis of the effect of point mutations in the sites showing strong evidence of selection would help to elucidate the source of the observed selection pressure. Additionally, several mutations were observed in only one progression group. One of these, Y69F was found exclusively in the non-viraemic patients and has previously been shown to have no effect on SAMHD1 degradation *in vitro*. *In vitro* analysis of the effect of mutations that are present only in viraemic patients (D3T, T28S, A31T, Q64R and K84G) would allow elucidation of the significance of this observation. This study also identified previously un-described mutations in *vpx*, showing the importance of using primary patient samples in capturing a complete picture of variation in *vpx*. The addition of *vpx* sequences derived from plasma virion RNA would also be beneficial. Although previous studies have shown that sequences from plasma and proviral compartments form a single monophyletic cluster, the differences in variant frequencies between the compartments are unknown in HIV-2 infection.

One of the most important considerations when interpreting results from HIV-2 genetic studies is how the sequences were generated. In recent years there has been a sequencing revolution and the advent of next generation sequencing technologies has significantly increased the potential amount of sequencing data that can be derived from a single experiment. In chapter 3, a comparison was made

between using MiSeq and cloning and Sanger sequencing to assess variant frequencies and diversity in the study of a PCR-amplified genomic fragment encompassing a single gene. Whilst the two methods showed high concordance on some levels, there were significant differences that need to be addressed in future sequencing studies. The ability to call variant frequencies was similar between the two methods; however, the accuracy of doing so was inversely proportional to (log) VL of the samples. Therefore, the estimated size of the viral population is an important factor to consider when using cloning and Sanger sequencing to assess diversity in a viral quasi-species and this study predicts a loss of power to call variants accurately as a function of increasing VL. Additionally, the estimates of nucleotide pairwise diversity were significantly different between the two methods, although there was a high correlation between the estimates for each patient. This suggests that diversity comparisons may only be valid when the data have been generated using exactly the same sequencing method, and makes meta-analysis of studies of HIV genetic diversity problematic. The high estimates of diversity from clonal amplification and Sanger sequencing are concerning and raise the possibility that polymerase errors could contribute more significantly to estimates of diversity than has previously been reported. This study did not address the PCR bias that is likely to be present in the amplified genomic fragments. PCR bias (or template resampling) can lead to over-estimation of the genetic homogeneity in a population and the stochastic nature of the bias makes post-sequencing correction impossible without the incorporation of a control such as primer ID tagging into the sequencing reaction.

The generation of an accurate and bias-free picture of the diversity present in the HIV quasi-species remains an appealing tool in the study of HIV infection[344]. Whilst corrections can be applied to remove some of the biases introduced during target enrichment, the ability to sequence viral RNA directly remains lacking. Chapters 4 and 5 of this thesis focus on a novel application of RNA-Seq to the sequencing of HIV. To the best of our knowledge, this study reports for the first time successful generation of whole genome HIV sequences without the need for

prior target enrichment. The lack of prior PCR has many benefits including a reduction in the PCR bias and the ability to generate whole genome sequences without the need for detailed *a priori* sequence knowledge. The *de novo* assembly algorithm employed, VICUNA, required input of HIV-2 sequence data, however, this was used for the identification of HIV-2 'like' reads post-sequencing and allowed an increase in the accuracy of contig validation. Viral assembly strategies that do not require any prior sequence knowledge, such as digital subtraction, were not assessed in this study but remain powerful tools for the assembly of genomes from novel and uncharacterised viruses. The generation of a *de novo* patient consensus sequence prior to read mapping allowed the assembly of whole genome sequences and analysis of the consensus sequences generated showed that identification of an appropriate reference sequence is challenging. Therefore the multi-step approach to read mapping presented here provides a repeatable pipeline, as shown by the similar results using different alignment algorithms. In line with previous RNA-Seq studies, the genomes assembled in this study showed both random hexamer and GC biases. The effects of both biases were thoroughly assessed and whilst there was an expectation of variable coverage resulting from these biases, the whole coding region of HIV-2 was successfully captured, suggesting these biases should not affect future studies of HIV using RNA-Seq. Additionally the estimated error rate was evaluated and shown to be within the expected limits for Illumina platforms, suggesting that the initial low RNA input did not adversely affect the downstream sequencing reaction.

The absolute limitations of using RNA-Seq on primary patient samples were assessed in Chapter 5 and this study suggests that a minimum viral load of 10,000 copies/mL is needed for successful sequencing. This cut-off point implies that RNA-Seq should also be feasible for samples from HIV-1 patients who are not on successful ART. RNA-Seq has also been used in work related to this project but outside of the remit of this thesis to generate whole genome sequences from samples taken from vertically infected Kenyan infants and their mothers, showing that RNA-Seq can also be successfully employed to sequence HIV-1. There are two

major drawbacks to using RNA-Seq on HIV-2 samples. The limit of 10,000 copies/mL means that it is not possible to generate sequence data from LTNPs, which make up a large proportion of HIV-2 infected individuals. Additionally, the failure rate was estimated to be 75%, showing that RNA-Seq is not a completely reliable technique. Both the absolute cut off in terms of viral load and the reported failure rates were higher than have been reported when genome-spanning amplicons are used to generate whole genome HIV-1 sequences. However, in the absence of a set of robust 'pan-HIV-2' primers, RNA-Seq offers a feasible method for the generation of whole genome sequences.

Genome builds were used to assess nucleotide site diversity over the whole HIV-2 genome for the first time. As may have been predicted, diversity was high in the immunologically relevant genes *nef* and *gag*. However, a somewhat unexpectedly high level of diversity was observed in *pol*. High diversity in *pol* has recently been observed in acute infection following vertical transmission of HIV-1 and therefore, the immunodominance of HIV-2 *pol* may be closer to what is observed in acute rather than chronic HIV-1 infection. The low level of diversity in *vpx*, seen in all 3 patients, is intriguing. When combined with the results of Chapter 3, which showed evidence of more significant signals of selection acting on *vpx* in viraemic patients, this results suggests that the role of *vpx* in the lack of HIV-2 control maybe more important than previously reported. We hypothesise that the role of HIV-2 infection of professional antigen presenting cells, leading to a more robust immune response and control of viral replication leads to a selective pressure on *vpx*, favouring a reduction of antagonism in SAMHD1. Further investigation of the roles of the accessory genes of HIV-2 should yield important discoveries about the mechanisms of control of viraemia and disease progression following HIV-2 infection.

Overall, this thesis presents novel data on *vpx* diversity in primary patient samples and demonstrates for the first time the applicability of RNA-Seq to plasma samples from HIV infected individuals. Further expansion of this method to include more patient samples and repeats of those shown here will give an even

clearer picture of the broad applicability of this method. The ability to generate whole genome sequences without the need for prior target enrichment opens up exciting possibilities to gain an ever more accurate picture of the dynamics of HIV evolution following infection and through different stages of disease progression. As the HIV epidemic continues, the development of new sequencing technologies broadens the scope of HIV research and this study shows how application of techniques designed to address specific questions about human genetics can also be beneficial for the study of HIV. One of the major challenges that still remains is the ability to sequence whole genomes from low and non-viraemic patients. Whilst third generation sequencing technologies such as nanopore sequencing offer the potential for much greater sensitivity and may even eventually permit the direct sequencing of RNA without the need for reverse transcription, the error rates reported for these platforms mean they are currently unsuitable for HIV research, where rare variants in the viral quasi-species are informative. The application of RNA-Seq to HIV has many possible future uses and one that may be extremely promising is the ability to assess accurately the frequency of drug-resistance or CTL escape mutation *in vitro*, where the viral load is typically much higher than seen in patient plasma. Low-bias sequence data presents a novel view on HIV evolution and diversity and could be employed in future studies looking at responses of HIV to vaccine challenge or reservoir re-activation.

# Appendix

## Low-Bias HIV-2 Whole Genome Sequencing Through Shotgun RNA Sequencing (RNA-Seq).

James, K.L.[1,2], DeSilva T.[3], Brown, K.[4], Whittle, H.[5], Taylor, S.[6], McVean, G.[2], Rowland-Jones, S.L.[1]

1 - Nuffield Department of Medicine, University of Oxford, UK. 2 – Wellcome Trust Centre for Human Genetics, University of Oxford, UK. 3- University of Sheffield, UK. 4 – CGAT, University of Oxford, UK. 5 – MRC Laboratories, The Gambia. 6 – Weatherall Institute of Molecular Medicine, University of Oxford, UK.

**Abstract:**

Accurate estimation of the extent of genetic diversity present in the HIV quasi-species is critical for the development of a preventative vaccine. To date, traditional and next generation sequencing methods have relied on target amplification prior to sequencing, introducing biases that may obscure the true signals of diversity in the viral population. Additionally, target-enrichment through PCR requires a large amount of *a priori* sequence knowledge, which is lacking for HIV-2. Therefore, a target enrichment free method of library preparation is desirable. We addressed this issue by applying an RNA shotgun sequencing (RNA-Seq) method to cultured viral stocks and primary patient plasma samples taken from HIV-2 infected individuals to show that RNA-Seq offers a feasible method for low-bias HIV-2 whole genome sequencing. Libraries generated from total plasma RNA were analysed with a 2-step pipeline (*de novo* genome assembly followed by read re-mapping), generating whole genome sequences with a mean depth of coverage of 28-67x. Assembled reads showed a low level of GC-bias and comparison of the diversity over the genome allowed the observation of low diversity in the accessory gene *vpx* in all patients.

This study demonstrates for the first time that RNA-Seq is a feasible sequencing method for plasma samples collected from HIV-2 infected individuals, resulting in high depth coverage of the complete HIV-2 genome in the context of reduced bias.

**Importance:**

Next generation sequencing technologies offer the potential to allow high-resolution capture of the complex patterns of genetic diversity seen in the HIV quasi-species at the patient and population level. An accurate picture of viral genetic diversity is a critical factor in the development of a globally effective HIV vaccine. However, sequencing strategies are often complicated by target enrichment prior to sequencing, introducing biases that can distort variant frequencies, which are not easily corrected for in downstream analyses. Additionally, detailed *a priori* sequence knowledge is needed to inform robust primer design when employing PCR amplification, a factor that is often lacking when working with tropical diseases localised in developing countries. Recent work has demonstrated that direct RNA shotgun sequencing (RNA-Seq) can be used to circumvent these issues for HCV and Norovirus. Whilst RNA-Seq has also been proposed for HIV-1 and HIV-2, it has yet to be demonstrated for either virus. In this study we applied shotgun RNA sequencing (RNA-Seq) to total RNA extracted from *ex vitro* and *ex vivo* HIV-2 samples, demonstrating the applicability of this technique to HIV for the first time and allowing us to generate a dynamic picture of genetic diversity over the whole genome of HIV-2 in the context of low-bias sequencing.

**Introduction:**

Human Immunodeficiency Viruses types 1 and 2 (HIV-1 and HIV-2), the two causative viruses of acquired immunodeficiency syndrome (AIDS), are human pathogens of high importance [1]. Following the introduction of HIV-1 and HIV-2 into human populations through zoonotic transmission of SIVs infecting several species of non-human primates,

HIV-1 and HIV-2 are estimated to have infected more than 75 million people worldwide, resulting in over 40 million deaths [2].

Whilst HIV-1 and HIV-2 share some common features, a major difference between the two viruses is the typical viral load associated with chronic infection. In patients infected with HIV-2, viral load is strongly correlated with disease progression and a large proportion (~37% in the study cohort) maintain undetectable viral loads and high CD4 counts in the absence of treatment, often for decades [3]. Additionally, lack of HIV-2 control is associated with lower viral loads when compared to HIV-1 disease stage matched patients, often in the order of a 1-2 log reduction [4]. Patients with a viral load of more than 10,000 copies/mL are defined as HIV-2 progressors and have a reduced survival probability that is similar to that seen in HIV-1 infected individuals in the absence of treatment [5].

One of the main barriers to the development of a globally protective HIV vaccine is the ability of HIV to rapidly evolve, introducing mutations that abrogate the binding of neutralising antibodies, rendering vaccine responses ineffective [6]. Therefore, a major focus of HIV research has been understanding the factors affecting viral evolution and identification of sites of high conservation in the HIV genome as potential vaccine targets [7]. Due to the small (~10kb) size of the HIV ssRNA genome, target enrichment is normally required prior to sequencing to generate sufficient DNA for downstream sequencing applications [8]. The most commonly employed method of target enrichment is PCR amplification [9]. This method has two major drawbacks; the first of these is the requirement for detailed *a priori* sequence knowledge to inform robust primer design that ensures the majority of variants in the viral quasispecies are captured [10]. A single pan-HIV-1 primer set has recently been described by Gall *et al*, allowing amplification of all HIV-1 groups with a sensitivity of 3,000copies/mL [11]. However, the sequence knowledge of HIV-2 is significantly lower than for HIV-1 and a robust pan-HIV-2 primer set has yet to be defined. Additionally, mutations in primer binding sites can reduce binding efficiency and therefore alter the proportion of specific variants in the final pool of amplicons, or in

the most extreme case may abrogate primer binding completely, resulting in the loss of that variant in the final analysis [12]. PCR is also stochastically biased, with amplicons from previous cycles acting as templates in the subsequent amplification cycles, further distorting the picture of the viral diversity [13].

Several methods have been proposed to circumvent these problems and reduce the biases introduced into sequencing data through target enrichment. Primer ID allows identification of reads derived from the same viral temple through incorporation of a unique 8-mer tag during the reverse transcription of viral RNA [14]. Down-stream reads can be pooled according to template and multiple reads from the same template can be used for error correction. A study using Primer ID observed biased diversity estimates between 2-100-fold when comparing to a library generated without any PCR bias correction, highlighting the importance of considering this factor when sequencing a highly diverse population, such as HIV. However, primer ID still relies on sufficient *a priori* sequence knowledge to allow robust primer design and the incorporation of the barcode into the 3' end of the cDNA molecule means it is not applicable to library preparation techniques involving random fragmentation of the target, such as those employed when using Illumina platforms.

Shotgun RNA sequencing (RNA-Seq) has recently been demonstrated as a powerful tool for the study of RNA viruses [15]. Library preparation is performed using random hexamer priming of the total RNA in a sample, negating the need for sequence-specific target enrichment [16]. This is particularly desirable for HIV-2, where the total sequence data available is significantly lower than for HIV-1. Two studies have applied RNA-Seq to human RNA viruses. Ninomiya *et al* applied RNA-Seq to plasma samples taken from two chronically HCV infected patients and demonstrated nearly full-length genome sequences with a mean depth of coverage between 50-70x for the two patients [17]. Batty *et al* further expanded this method, presenting a high-throughput method for Norovirus sequencing, allowing 77 faecal samples to be sequenced with a mean depth of coverage of 100x and a success rate of more than 99% [18]. The authors compared this with a PCR amplification

strategy and found that the success rate for whole genome amplification using PCR was 29%, representing a significant decrease in the performance when compared to RNA-Seq. RNA-Seq has also recently been used for the discovery of two novel SIVs. Lauck *et al* applied RNA-Seq to plasma samples taken from free-ranging colobus monkeys and identified two novel and highly SIVs (termed SIVkcol-1 and SIVkcol-2), demonstrating the power of sequence independent sequencing in viral discovery [19].

Whilst sequence-independent DNA amplification (linear isothermal amplification) has been successfully applied to HIV-1, the applicability of RNA-Seq to HIV has yet to be demonstrated [20]. Our study applied RNA-Seq library preparation methods to cultured lab adapted HIV-2 reference strains and patient plasma samples taken from a rural West African community cohort. We show that RNA-Seq is a feasible and powerful tool when applied to *in vitro* and *ex vivo* HIV-2 samples with a viral load of at least 10,000 copies/mL. Additionally we quantified the biases involved in using RNA-Seq, demonstrating that this method represents a novel, low-bias method of HIV sequencing. Finally, we computed estimates of nucleotide pairwise diversity for each gene of HIV-2, allowing an inter-patient comparison for the first time, and showing consistently low estimates of diversity in the accessory gene *vpx*, highlighting the importance of this HIV-2 specific gene in successful HIV-2 infection.

**Materials and Methods:**

**Patient sample collection and ethics statement:** All patient samples used in this study were collected from members of the Caió community cohort who had provided written and informed consent. Samples were collected prior to the start of this study and plasma was separated from whole blood through centrifugation (5000xg, 5 minutes, 4$^{o}$C) and filtration (0.45μM filter, Millipore, Billerica, MA, USA). Plasma samples were stored at -80$^{o}$C and transported to Oxford in a liquid nitrogen dry shipper. Ethical approval for this study was

granted by the Gambian Government/MRC joint ethics committee (#SCC1204) and the Oxford tropical research ethics committee (#170-12).

***In vitro* culture of lab adapted HIV-2 reference strains:** The lab-adapted HIV-2 strains HIV-2 ROD and HIV-2 CBL-20 were propagated *in vitro* in the lymphocyte cell line H9, a single cell clone derived from a HUT 78 cell line. Infection of $5x10^6$ cells was carried out with 200 µl of $9x10^3$ TCID/50mL of viral stock. Cells were removed through centrifugation at 250xg for 10 minutes and supernatant was collected on days 3, 5, 7, 9, 11, 13 and 15. HIV-2 concentration was assayed using a reverse transcriptase colorimetric assay (Roche). For each isolate, the supernatant sample with the highest reverse transcriptase concentration was selected for RNA-Seq.

**RNA extraction, RNA quantification and DNase treatment:** Total nucleic acid was extracted from 500µl patient plasma or purified supernatant using the QIAamp UltraSens Viral Kit (Qiagen). Extraction was performed according to the manufacturer's protocol with the substitution of carrier RNA with linear acrylamide (Ambion) as the nucleic acid co-precipitant. Final elution was performed in 12µl $H_2O$. DNA was removed from the samples through treatment with DNase I (Turbo DNase, Ambion) according to the manufacturer's protocol. RNA concentration was estimated using the QuBit RNA assay (Invitrogen).

**Library preparation and sequencing:** Sequencing libraries were prepared from 5µl of the eluted RNA using the NEBNext Ultra RNASeq Kit (New England Biolabs) according to the manufacturer's protocol. Sequencing libraries were multiplexed and sequenced using the Illumina HiSeq or MiSeq platforms (Illumina). Patient samples were multiplexed 6/lane (HiSeq), generating 2x100bp paired-end reads and lab adapted strains were multiplexed 2/lane (MiSeq), generating 2x150bp paired-end reads.

**De novo genome assembly and read re-mapping:** Sequence data were analysed using a custom pipeline. Reads were trimmed using sickle stipulating a median Q-score >30 and a read length > 40bp [21]. *De novo* genome assembly was performed using VICUNA [22],

with the addition of the optional contamination removal step. During contamination removal, HIV-2 derived reads were identified through similarity to a multiple sequence alignment containing all publically available HIV-2 group A sequence data (See SI). Overlapping contiguous sequences generated by VICUNA were assembled into whole genome sequences using the map to reference feature in Geneious v 6.1.6 [23] and manually inspected to derive a whole genome consensus sequence. Consensus genome sequences were manually inspected to ensure they contained intact open reading frames. Reads were re-mapped to the consensus genome sequence using Bowtie2 [24], BWA-SW [25], GSNAP [26] and NovoAlign [27] for each sample. Files containing assembled reads were manipulated using the SAMtools package [28] and downstream statistical analyses and data visualisations were performed using R [29] and the Interactive Genome Viewer [30]. Error rates were estimated using the ErrorRatePerCycle feature of GATK [31].

**Quantification of biases:** Random hexamer bias was assessed through visualisation of the base composition of reads using FASTQC [32]. GC bias was quantified using a custom Python script that scanned the genome using a 50bp sliding window with a step size of 20bp. Mean GC content and mean depth of coverage were computed for each window and GC bias was assessed by fitting a linear regression in R.

**Phylogenetic analysis:** A reference set of 18 HIV-2 group A whole genome sequences were obtained from the Los Alamos HIV Database [33] (Table S1). Reference sequences were aligned with consensus whole genome sequences using Muscle [34] and the alignment was manually inspected using Geneious v. 6.1.6. A Bayesian phylogeny was inferred using BEAST v 1.8.0 [35], under the general time reversible model of nucleotide substitution with a proportion of invariant sites and gamma-distributed rate heterogeneity, as determined by jModelTest2 [36]. The Markov Chain Monte Carlo algorithm was run using 100,000,000 iterations with samples taken from the posterior every 10,000 generations. Following a burn-in corresponding to 10% of the samples, the resulting maximum clade credibility tree (MCCT) was visualised using FigTree v 1.4.1[37].

**Estimation of genetic diversity:** Mpileup files were generated from assembled reads using the SAMtools package and variants were called using VarScan [38] with a cut-off frequency of 0.05. Nucleotide pairwise diversity ($\pi$) was estimated using the Nei and Li method [39] through a custom Python script, taking depth at each position as a proxy for population size and the product of frequency of alternative variants and depth as the number of pairwise differences between sequences. Estimates of diversity were generated for each individual gene and over the whole genome and estimates were normalised using the whole genome average to allow comparison between patients.

**Results:**

***De novo* genome assembly and factors influencing RNA-Seq success rate:**

We assessed the ability of standard RNA-Seq library preparation methods to capture HIV-2 using a panel of patients whose plasma viral loads represented the broad spectrum seen in natural HIV-2 infection (Table 1). In addition to patient plasma samples we also included two lab adapted HIV-2 strains (HIV-2 ROD and HIV-2 CBL-20) as positive controls. *De novo* genome assembly using VICUNA was performed on all patient samples and HIV-2 CBL-20, as this lab adapted isolate does not have a published whole genome sequence. Successful capture of the HIV-2 genome was defined using the contigs generated by VICUNA (Table 1). Contigs were assembled to the HIV-2 reference genome UC2 and merged to form a single patient-specific consensus sequence.

In total, whole genome assembly was successful for 3 of the 6 patient samples and the positive control, HIV-2 CBL-20. Successful patient samples showed complete capture of the coding region of HIV-2 and merged contigs ranged from 9397-9776bp in length. A merged contig spanning the complete coding region was also assembled for CBL-20, demonstrating the applicability of RNA-Seq to both *in vitro* and *ex vivo* samples. The results presented in this study suggest an absolute cut off in sensitivity of 10,000 copies/mL with an expectation of at least 0.001% HIV-derived RNA Table 1). When these limits are

considered, the success rate was 75%, which is much higher than has previously been described for genome-spanning PCR amplicons for other RNA viruses.

**Phylogenetic analysis of *de novo* genome sequences:**

These results highlight the need for the *de novo* assembly of a patient-specific reference prior to read mapping in the context of a highly variable and genetically diverse species, such as HIV-2. We used a panel of publically available HIV-2 sequences (Table S1) to validate whole genome consensus HIV-2 sequences generated in this study. Bayesian phylogenetic analysis of the consensus sequences showed that the newly generated sequences fell within the known existing radiation of HIV-2 group A and represented novel braches that were distinguishable from existing reference sequences (Figure 1). The HIV-2 ROD sequence generated in this study and the published reference sequence were also included in the analysis and, as would be expected, these two sequences were closely related and clustered together with a high posterior probability of 1. Therefore this analysis shows that a distinct consensus sequence existed for each patient and there was no cross-contamination during the library preparation. Additionally, as would be expected when considering samples were taken from a community cohort that is not thought to be the result of a narrow-source outbreak, this analysis shows that the selection of an appropriate 'best' reference sequence for assembly would be different for each patient.

**Capture of the HIV-2 genome by *de novo* genome assembly:**

In order to assess how well *de novo* assembly using VICUNA had captured the HIV-2 genome, consensus sequences were aligned to the reference sequence UC2 and annotated according to homology (Figure S1). Whilst patient consensus sequences contained all 9 genes of HIV-2 in intact reading frames, there was some variability in the assembly of the 3' and 5' LTR and the *Gag* leader sequence. This is summarised in Table S2. In the case of patient sample TD031, the loss of 177bp of the *Gag* leader sequence can be attributed to the failure of the RNA-Seq library preparation method to capture this region. The initial fragmentation step in library preparation can lead to the loss of distal regions of the RNA

molecule and this is the most probable cause of the lack of coverage in this genomic region. For patient samples TD024 and TD062 and the reference strain CBL-20, the lack of coverage can be explained by examining the nature of the LTRs in HIV-2. The 5' and 3' LTR regions only exist as true 990bp repeats in the proviral form of the virus, whereas in the RNA genome, the 5' LTR comprises the R and U5 regions and the 3' LTR is composed of the R and U3 regions [40]. As the MSA used during assembly contained HIV-2 sequence data from both cDNA and RNA HIV-2 genomes, assembly was conducted using 'complete' LTRs, both containing U5, R and U3. An attempt to correct for this was not feasible as the majority of HIV-2 sequences in LANL are derived from pro-virus and removal of these sequences would have resulted in a reduction in the power to detect HIV-2 derived reads. Ambiguous read mapping is normally resolved by using the location of the read mate to provide information on the most likely co-ordinates. In this study the insert size (250-350bp) and the nature of the LTRs meant problematic mapping in the case of reads mapping to the R region, as the read mate will also fall in the LTR. Therefore, it is not possible to resolve the correct orientation of the reads, resulting in the loss of coverage from one LTR. As this study was concerned with sequence diversity in the genes of HIV-2 this was not corrected for, however, it is worth noting for future studies applying RNA-Seq to HIV.

**Read re-mapping to the patient-specific consensus whole genome sequences:** In contrast to re-sequencing projects, where a high depth of coverage is required for error correction, deep sequencing of pathogen populations uses high depth of coverage to gain a picture of the diversity in the population as a whole [15]. Following *de novo* assembly of a patient-specific consensus genome sequence, we assessed the performance of 4 commonly used alignment tools when re-mapping reads to the patient-specific consensus (Table 2). Read re-mapping was performed using the total reads without prior HIV-2 enrichment or digital subtraction of human sequences to allow an assessment of how these tools perform in the context of a high level of contamination, which is likely to be a factor of all pathogen

sequencing strategies employing RNA-Seq. Mean depth and range of coverage were compared for each aligner (Figure 2). These results show consistent performance when 4 common aligners are used, with mean depth ranging from 27.69x – 67.23x for the 3 patient samples. This depth of coverage is in line with previous RNA-Seq studies, showing that RNA-Seq is a feasible tool for generating high-depth HIV-2 whole genome sequences. Additionally the similarity in the plots of coverage over the genome for the 4 aligners shows that following the *de novo* assembly of a patient-specific consensus sequence, read mapping is robust and repeatable, irrespective of which alignment algorithm is employed for read re-mapping.

**Assessment of the random hexamer bias:** A commonly recognised bias that is specific to RNA-Seq protocols is the random hexamer bias [41]. Hypothetical differential binding affinities between different random hexamers results in a biased nucleotide composition seen at the 3' end of the reads, normally spanning 7-13bp. As previously described, the data generated in this study showed evidence of a random hexamer bias affecting the first 13bp of the read (Figure S2). The pattern of the bias was remarkably similar in all three patient samples, suggesting that there may be preferential binding to the same motifs in all samples. This biased read composition can be attributed to random hexamer bias rather than low quality sequencing at the end of the reads as the median Q-score was constant over the length of the read. The effect of the bias did not extend past the first 13bp of each read and the nucleotide composition stabilised after this point. A correction was not applied to account for the biased nucleotide composition of the first 13bp, as removal of these positions does not remove the effects of this bias seen in down-stream analyses.

**Quantification of the GC% bias:** Depth of coverage in samples sequenced using Illumina short read chemistry can be affected by the local GC content of the genome [42]. We assessed the effect of local GC content on depth of coverage using a custom script which took a sliding window of 50bp, with a step size of 20bp, and calculated GC% and mean depth of coverage in each window. The extent of the GC bias was quantified using the

slope of the linear regression line and the bias was assessed for each aligner individually (Table 3). In order to allow a comparison between different aligners, the mean depth of coverage was normalised in each window using the genome-wide mean depth of coverage (Figure 3). All assemblies showed a slight, positive GC bias, suggesting that GC rich regions had depth of coverage that was higher than the mean. For patient samples TD024 and TD031, the magnitude of the slope was similar for all 4 aligners, suggesting a constant effect when different assembly algorithms were employed. In contrast, sample TD062 showed more fluctuation between aligners. However, the magnitude of the bias was lowest in this patient, suggesting that the overall effect of the GC bias would be reduced, in spite of the fluctuations. Therefore, this study showed evidence for a positive GC bias in HIV-2 samples sequenced using RNA-Seq, which may be a cause of variability in depth of coverage over the genome. However, the magnitude of the bias was in line with previous studies and did not show a loss of coverage of any genomic regions due to GC bias.

**Depth of coverage as a function of genomic context:** RNA-Seq is a robust method for capturing the complete coding region of HIV-2 at a high depth of coverage and fluctuations between patients are more likely to be attributable to the random hexamer or GC biases, rather than the diverse nature of the HIV-2 quasi-species. The HIV-2 genome contains 9 genes and the functional constraint and variability of each is highly influenced by its role in the HIV-2 life cycle [43]. For example, *env* encodes the envelope protein and consists of constant and variable regions, responsible for maintaining viral stability and facilitating escape from host immune responses respectively. Short read assembly can be influenced by genomic context and in highly variable regions, reads may fail to align due to divergence between the consensus and individual read sequences. In order to assess whether genomic context could affect depth of coverage, the HIV-2 genome was partitioned according to gene and mean depth of coverage was compared for each gene individually. The effect of genomic context on depth of coverage was visualised by plotting mean depth of coverage

as a function of GC content for each gene (Figure 4). All aligners showed a similar pattern of coverage and no consistent loss of coverage in any genomic region.

**Quantification of the divergence from the published reference sequence following *in vitro* expansion:** The use of lab-adapted HIV strains has revolutionised the field of HIV study. The ability to rapidly grow virus *in vitro* for infectivity and escape assays has allowed the elucidation of many of the host immune factors affecting clinical outcome following HIV infection. However, these studies are often based on the assumption that these stocks are clonal and that observed outcomes are as a result of differing experimental conditions rather than random mutations acquired by the virus in culture. In order to assess the divergence of HIV-2 ROD *in vitro* from the published reference sequence, variants at a frequency of >95% in the assembled reads were identified as SNPs using VarScan (Figure 5). A total of 70 SNPs were observed (Figure 7). The genomic location of SNPs was visualised and SNPs were seen in all genes except *vif*. The frequency of SNPs was also calculated and the majority located in *nef*. The function of *nef* as an accessory gene that aids escape from many host immune responses may explain this observation, as these factors are absent in an *in vitro* system, reducing the selective pressure acting on *nef* and allowing the accumulation of mutations without an associated fitness cost to the virus. This study shows that the assumption of a clonal viral population when using a lab-adapted HIV strain may not be a valid one and we present RNA-Seq as a cheap, quick and low-bias method for verification of the genetic make up of HIV populations *in vitro*.

**Genome-wide estimation of genetic diversity in HIV-2 in the context of low-bias sequencing:** To determine how the diversity varies over the HIV-2 genome, we estimated nucleotide pairwise diversity from assembled reads (Bowtie2 assembly) using a custom script. Raw estimates of diversity were normalised using the genome average to allow comparison between patients (figure 6). Overall, our analysis showed similar results between patients and as expected, the highest level of diversity was seen in *env* for all three patients. Interestingly, the diversity in *pol* was higher than for the accessory genes *vpx, vpr*

and *vif* for all three patients. Whilst *pol* is traditionally thought of as a highly conserved gene in HIV-1, a recent study has shown higher diversity in *pol* when compared to *gag* following vertical transmission of HIV-1 [44]. This study suggests that *pol* is also more diverse than *gag* in HIV-2 following horizontal transmission. The RNA-Seq method employed in this study, allowed a comparison of the estimates of diversity in each gene of HIV-2, from a single patient at a single time point in the context of low sequencing biases. The low level of diversity seen in *vpx* in all three patients is of particular interest as *vpx* is an HIV-2/SIVsmm specific accessory gene that is absent from HIV-1 [45]. The main role of *vpx* is antagonism of the host restriction factor SAMHD1, which blocks reverse transcription of viral RNA in slowly dividing cells such as macrophages and resting CD4+ T Cells. The impact of *vpx* antagonism on HIV-2 disease progression remains poorly defined, however, the results of this study suggest that *vpx* may play a critical role in the HIV-2 life cycle, as low diversity may be indicative of a strong purifying selective pressure.

**Discussion:**

Deep sequencing of HIV offers unparalleled opportunities to gain a high-resolution picture of the nature and diversity of the viral quasi-species in a single patient. This study presents a novel application of RNA-Seq, allowing the entire coding region of HIV-2 to be sequenced without the need for detailed *a priori* sequence knowledge, a factor that has previously prevented the description of a robust pan-HIV-2 whole genome amplification strategy. We showed the broad applicability of this method, presenting data from both *ex vitro* lab-adapted isolates and *ex vivo* patient plasma samples. For patient samples, we demonstrated a cut off in sensitivity at a viral load of 10,000 copies/mL and an expectation of at least 0.001% HIV RNA in the sample. When these conditions were fulfilled, we report a success rate of 75%, which is lower than previously reported by Batty *et al* when applying RNA-Seq to Norovirus. However, the lower HIV-2 plasma viral loads of the patient samples used in this study readily explain this reduced success rate. A cut off of

10,000 copies/mL restricts this method to viraemic HIV-2 patients, however, we anticipate that RNA-Seq can also be successfully applied to samples taken from un-treated HIV-1 patients, where the typical viral load is 10 to 1000 times higher than for HIV-2. Additionally, the pan HIV-1 amplification strategy developed by Gall *et al* showed a cut off in sensitivity at 3,000 copies/mL, demonstrating the difficulty of generating robust and high-depth sequence data from patients without detectable plasma viraemia.

Whilst RNA-Seq allows whole genome sequencing of HIV-2 without the need for detailed sequence knowledge, the lack of sequence-specific target amplification also leads to a reduction in the use of PCR and the resulting biases, allowing sequence data that is more representative of the true population frequencies. This study aimed to quantify the other biases known to be associated with RNA-Seq. We found evidence of a moderate positive GC bias, which varied between samples but was consistent when different aligners were used. We also showed evidence of a biased nucleotide composition in the first 13bp of the reads, suggesting the presence of non-random random hexamer priming. Although these biases are likely to be responsible for the fluctuations in coverage over the genome, we observed no correlation between genomic location and depth of coverage, suggesting these fluctuations were randomly distributed and not due to the varying diversity seen in different functional genomic sites.

The ability to sequence the whole HIV-2 genome in a single experiment allowed us to compare pairwise nucleotide site diversity between the different genes of HIV-2. There is limited data about the nature of genetic diversity in HIV-2, however, a recent study by Esbjornsson *et al* suggested that HIV-1 genetic diversity was reduced in the context of HIV-1 and HIV-2 dual infection, when compared to HIV-1 mono-infected disease matched individuals [46]. The authors did not examine HIV-2 genetic diversity, however, the reduced rate of disease progression for HIV dual infected individuals suggests there may be an epistatic interaction in the context of HIV dual infection. This study showed remarkably similar patterns of diversity across all three patients. The highest diversity was seen in *env*,

an observation that is in line with patterns seen in HIV-1. The high level of diversity in *env* is driven by the selective pressures exerted by the host immune system, driving escape of Nab binding and immune responses. Similarly, a high level of diversity was seen in *nef* in all three patients. *Nef* is an accessory gene that has a key role in the evasion of host immune responses, primarily through HLA and CD4 down-regulation, preventing the display and recognition of virally derived peptides at the cell surface. Interestingly, a high level of diversity was observed in *pol* in all patients. In HIV-1 infection, *pol* is thought to be a highly conserved and therefore typically shows low diversity. However, a recent study has shown a high level of diversity in *pol* following vertical HIV-1 transmission and this study shows a similar result for HIV-2. Therefore, these results could be attributable to similarities between HIV-1 vertical and HIV-2 transmission, possibly suggesting that in contrast to sexual HIV-1 transmission where the majority of infections are attributable to a single transmitted-founder virus, multiple transmitted founder viruses cause HIV-2 infection.

*Vpx* is an HIV-2 specific accessory gene that is entirely absent from the HIV-1/SIVcpz lineage. The role of *vpx* is antagonism of the host restriction factor SAMHD1, however, little is known about the implications of this antagonism on the course of HIV-2 disease progression. The observation of a consistently low level of diversity in *vpx* may be indicative of a high level of conservation in *vpx*, suggesting that *vpx* has a critical role in the maintenance of high level of HIV-2 viraemia. The different roles of *vpx* in HIV-2 infection remain to be clearly defined, but a recent study by Yu *et al* identified a SNP in a *vpx* allele derived from a viraemic patient that totally abrogated the ability of *vpx* to promote SAMHD1 degradation *in vitro* [47].

In conclusion, by applying RNA-Seq library preparation methods to HIV-2 *ex vitro* and *ex vivo* samples, we have demonstrated the applicability of this method to HIV samples for the first time. Resulting *de novo* genome assemblies captured the entire coding region of HIV-2 in intact open reading frames and read re-mapping allowed us to demonstrate the

importance of a two-step analysis pipeline. In the context of a highly diverse retrovirus, such as HIV-2, the selection or generation of an appropriate reference sequence is a critical first step, allowing robust and repeatable down-stream read mapping. We also demonstrated a low level of GC and random hexamer bias, and in the absence of sequence-specific target amplification, show that RNA-Seq offers a method of whole genome HIV-2 sequencing in a low bias context. The importance of developing novel and low-bias HIV sequencing protocols cannot be understated as the ability to gain a complete an accurate picture of genetic diversity in HIV is critical to the development of a globally effective preventative HIV vaccine.

**References**


[1] De Silva TI, Cotten M, Rowland-Jones SL. HIV-2: the forgotten AIDS virus. Trends Microbiol 2008;16:588–95. doi:10.1016/j.tim.2008.09.003.

[2] Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, et al. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. Science 2014;346:56–61. doi:10.1126/science.1256739.

[3] Berry N, Jaffar S, Schim van der Loeff M, Ariyoshi K, Harding E, N'Gom PT, et al. Low level viremia and high CD4% predict normal survival in a cohort of HIV type-2-infected villagers. AIDS Res Hum Retroviruses 2002;18:1167–73. doi:10.1089/08892220260387904.

[4] Popper SJ, Sarr AD, Travers KU, Guèye-Ndiaye A, Mboup S, Essex ME, et al. Lower human immunodeficiency virus (HIV) type 2 viral load reflects the difference in pathogenicity of HIV-1 and HIV-2. J Infect Dis 1999;180:1116–21. doi:10.1086/315010.

[5] Hansmann A, Schim van der Loeff MF, Kaye S, Awasana AA, Sarge-Njie R, O'Donovan D, et al. Baseline plasma viral load and CD4 cell percentage predict survival in HIV-1- and HIV-2-infected women in a community-based cohort in The Gambia. J Acquir Immune Defic Syndr 1999 2005;38:335–41.

[6] Kawashima Y, Pfafferott K, Frater J, Matthews P, Payne R, Addo M, et al. Adaptation of HIV-1 to human leukocyte antigen class I. Nature 2009;458:641–5. doi:10.1038/nature07746.

[7] Barouch DH. Challenges in the development of an HIV-1 vaccine. Nature 2008;455:613–9. doi:10.1038/nature07352.

[8] Archer J, Baillie G, Watson SJ, Kellam P, Rambaut A, Robertson DL. Analysis of high-depth sequence data for studying viral diversity: a comparison of next generation sequencing platforms using Segminator II. BMC Bioinformatics 2012;13:47. doi:10.1186/1471-2105-13-47.

[9] Gao F. Amplification and cloning of near full-length HIV-2 genomes. Methods Mol Biol Clifton NJ 2005;304:399–407. doi:10.1385/1-59259-907-9:399.

[10] Pan W, Byrne-Steele M, Wang C, Lu S, Clemmons S, Zahorchak RJ, et al. DNA polymerase preference determines PCR priming efficiency. BMC Biotechnol 2014;14:10. doi:10.1186/1472-6750-14-10.

[11] Gall A, Ferns B, Morris C, Watson S, Cotten M, Robinson M, et al. Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. J Clin Microbiol 2012;50:3838–44. doi:10.1128/JCM.01516-12.

[12] Pinto AJ, Raskin L. PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. PloS One 2012;7:e43093. doi:10.1371/journal.pone.0043093.

[13] Smith EN, Jepsen K, Khosroheidari M, Rassenti LZ, D Antonio M, Ghia EM, et al. Biased estimates of clonal evolution and subclonal heterogeneity can arise from PCR duplicates in deep sequencing experiments. Genome Biol 2014;15:420. doi:10.1186/PREACCEPT-1251182501124451.

[14] Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proc Natl Acad Sci U S A 2011;108:20166–71. doi:10.1073/pnas.1110064108.

[15] McElroy K, Thomas T, Luciani F. Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. Microb Inform Exp 2014;4:1. doi:10.1186/2042-5783-4-1.

[16] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5:621–8. doi:10.1038/nmeth.1226.

[17] Ninomiya M, Ueno Y, Funayama R, Nagashima T, Nishida Y, Kondo Y, et al. Use of illumina deep sequencing technology to differentiate hepatitis C virus variants. J Clin Microbiol 2012;50:857–66. doi:10.1128/JCM.05715-11.

[18] Batty EM, Wong THN, Trebes A, Argoud K, Attar M, Buck D, et al. A Modified RNA-Seq Approach for Whole Genome Sequencing of RNA Viruses from Faecal and Blood Samples. PLoS ONE 2013;8:e66129. doi:10.1371/journal.pone.0066129.

[19] Lauck M, Switzer WM, Sibley SD, Hyeroba D, Tumukunde A, Weny G, et al. Discovery and full genome characterization of two highly divergent simian immunodeficiency viruses infecting black-and-white colobus monkeys (Colobus guereza) in Kibale National Park, Uganda. Retrovirology 2013;10:107. doi:10.1186/1742-4690-10-107.

[20] Kurn N, Chen P, Heath JD, Kopf-Sill A, Stephens KM, Wang S. Novel isothermal, linear nucleic acid amplification systems for highly multiplexed applications. Clin Chem 2005;51:1973–81. doi:10.1373/clinchem.2005.053694.

[21] JN. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files n.d.

[22] Yang X, Charlebois P, Gnerre S, Coole MG, Lennon NJ, Levin JZ, et al. De novo assembly of highly diverse viral populations. BMC Genomics 2012;13:1–13. doi:10.1186/1471-2164-13-475.

[23] Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: An integrated and extendable desktop software platform for the organization

and analysis of sequence data. Bioinformatics 2012;28:1647–9. doi:10.1093/bioinformatics/bts199.

[24] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–9. doi:10.1038/nmeth.1923.

[25] Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinforma Oxf Engl 2010;26:589–95. doi:10.1093/bioinformatics/btp698.

[26] Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 2005;21:1859–75. doi:10.1093/bioinformatics/bti310.

[27] Novocraft.com: Novoalign short read mapper (http://www.novocraft.com/main/downloadpage.php) n.d.

[28] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinforma Oxf Engl 2009;25:2078–9. doi:10.1093/bioinformatics/btp352.

[29] R Core Team. R: A Language and Environment for Statistical Computing 2013.

[30] Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013;14:178–92. doi:10.1093/bib/bbs017.

[31] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303. doi:10.1101/gr.107524.110.

[32] Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data. Http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ n.d. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed September 19, 2014).

[33] Los Alamos Laboratories. HIV Sequence Databases n.d.

[34] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–7. doi:10.1093/nar/gkh340.

[35] Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 2007;7:214. doi:10.1186/1471-2148-7-214.

[36] Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 2012;9:772–772. doi:10.1038/nmeth.2109.

[37] FigTree n.d. http://beast.bio.ed.ac.uk/figtree (accessed October 1, 2014).

[38] Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinforma Oxf Engl 2009;25:2283–5. doi:10.1093/bioinformatics/btp373.

[39] Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci 1979;76:5269–73.

[40] Knipe DM, Howley PM. Fields' Virology. Lippincott Williams & Wilkins; 2007.

[41] Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res 2010;38:e131. doi:10.1093/nar/gkq224.

[42] Chen Y-C, Liu T, Yu C-H, Chiang T-Y, Hwang C-C. Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. PLoS ONE 2013;8:e62856. doi:10.1371/journal.pone.0062856.

[43] Barroso H, Taveira N. Evidence for negative selective pressure in HIV-2 evolution in vivo. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis 2005;5:239–46. doi:10.1016/j.meegid.2004.07.008.

[44] Lipscomb JT, Switzer WM, Li J-F, Masciotra S, Owen SM, Johnson JA. HIV Reverse-Transcriptase Drug Resistance Mutations During Early Infection Reveal Greater Transmission Diversity Than in Envelope Sequences. J Infect Dis 2014. doi:10.1093/infdis/jiu333.

[45] Laguette N, Sobhian B, Casartelli N, Ringeard M, Chable-Bessia C, Ségéral E, et al. SAMHD1 is the dendritic- and myeloid-cell-specific HIV-1 restriction factor counteracted by Vpx. Nature 2011;474:654–7. doi:10.1038/nature10117.

[46] Esbjörnsson J, Månsson F, Kvist A, Isberg P-E, Nowroozalizadeh S, Biague AJ, et al. Inhibition of HIV-1 disease progression by contemporaneous HIV-2 infection. N Engl J Med 2012;367:224–32. doi:10.1056/NEJMoa1113244.

[47] Yu H, Usmani SM, Borch A, Krämer J, Stürzel CM, Khalid M, et al. The efficiency of Vpx-mediated SAMHD1 antagonism does not correlate with the potency of viral control in HIV-2-infected individuals. Retrovirology 2013;10:27. doi:10.1186/1742-4690-10-27.

**Figures:**

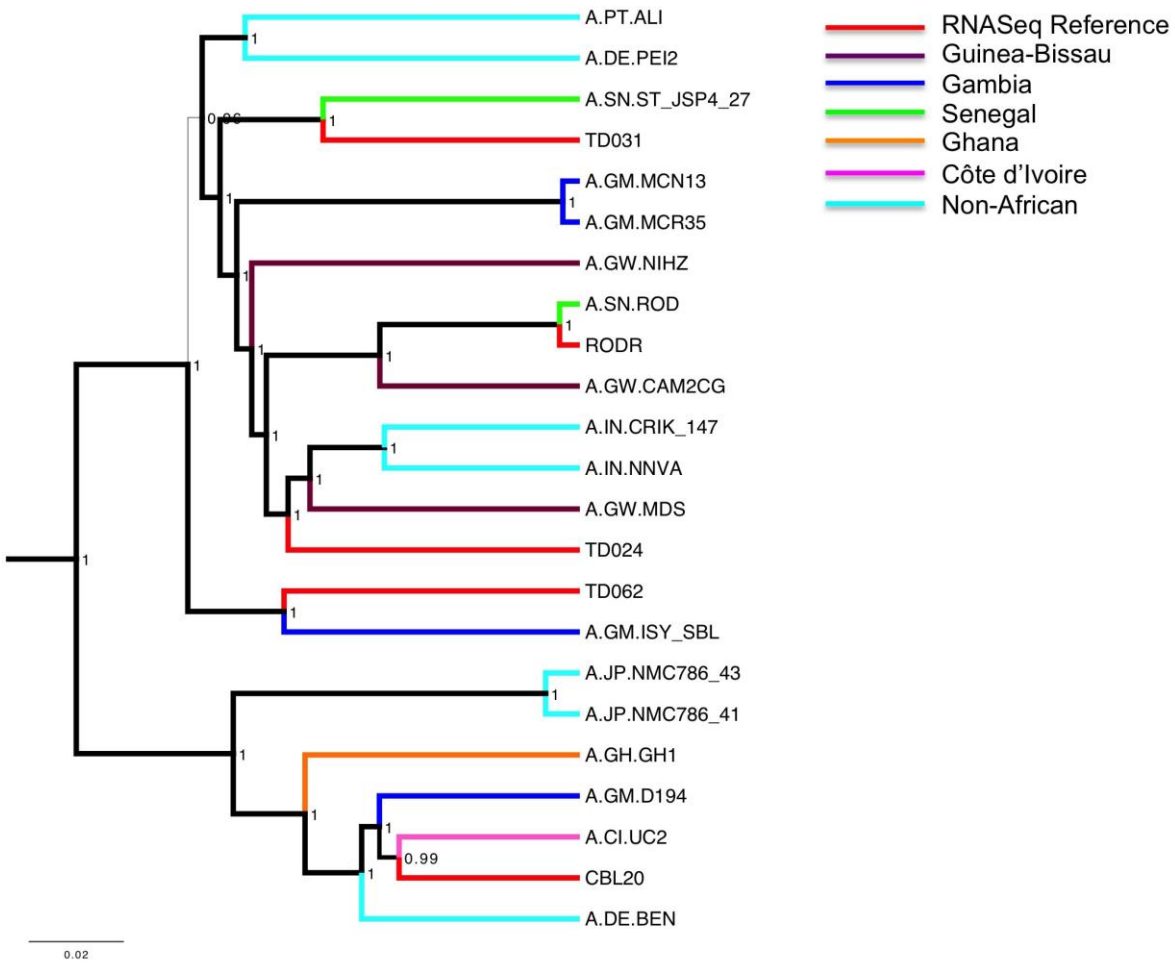



**Fig 1:** Bayesian phylogeny of HIV-2 genome sequences generated in this study. 18 whole genome HIV-2 group A sequences were included as a reference set (table S1). Reference sequences are coloured according to country of origin and sequences generated in this study are shown in red. Bayesian posterior probabilities are included on the corresponding nodes and the scale bar represents the number of nucleotide substitutions per site.

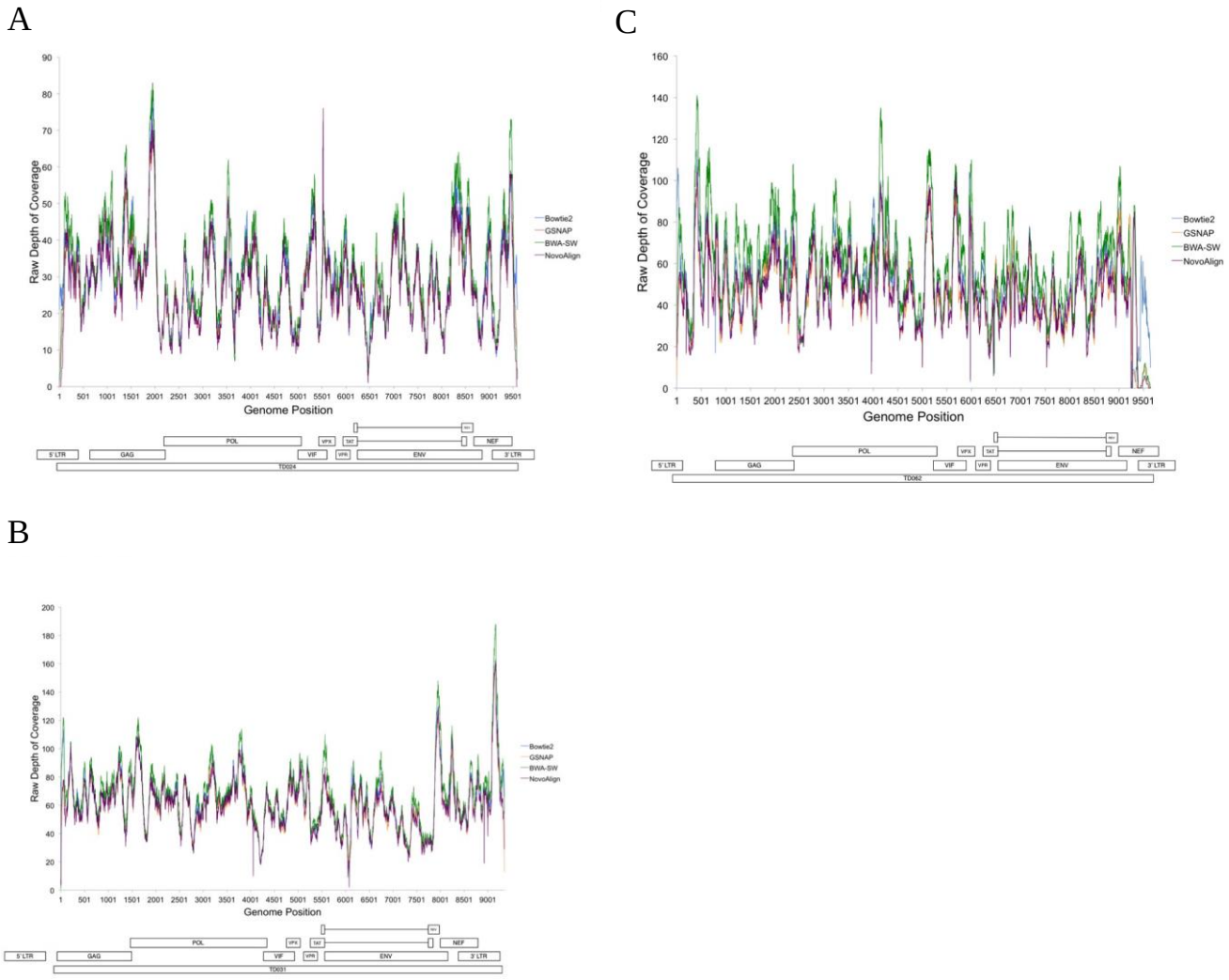



**Fig 2:** Depth of coverage for each locus was plotted for TD024 (A), TD031 (B) and TD062 (A). Open rectangles represent the locations of HIV-2 genes and the position of the longest merged contig is also shown for each sample. Coverage plots are shown for each of the four aligners, Bowtie2 (blue), GSNP (orange), BWA-SW (green) and NovoAlign (purple). Coverage was plotted as raw depth, showing the number of reads mapping to each locus.

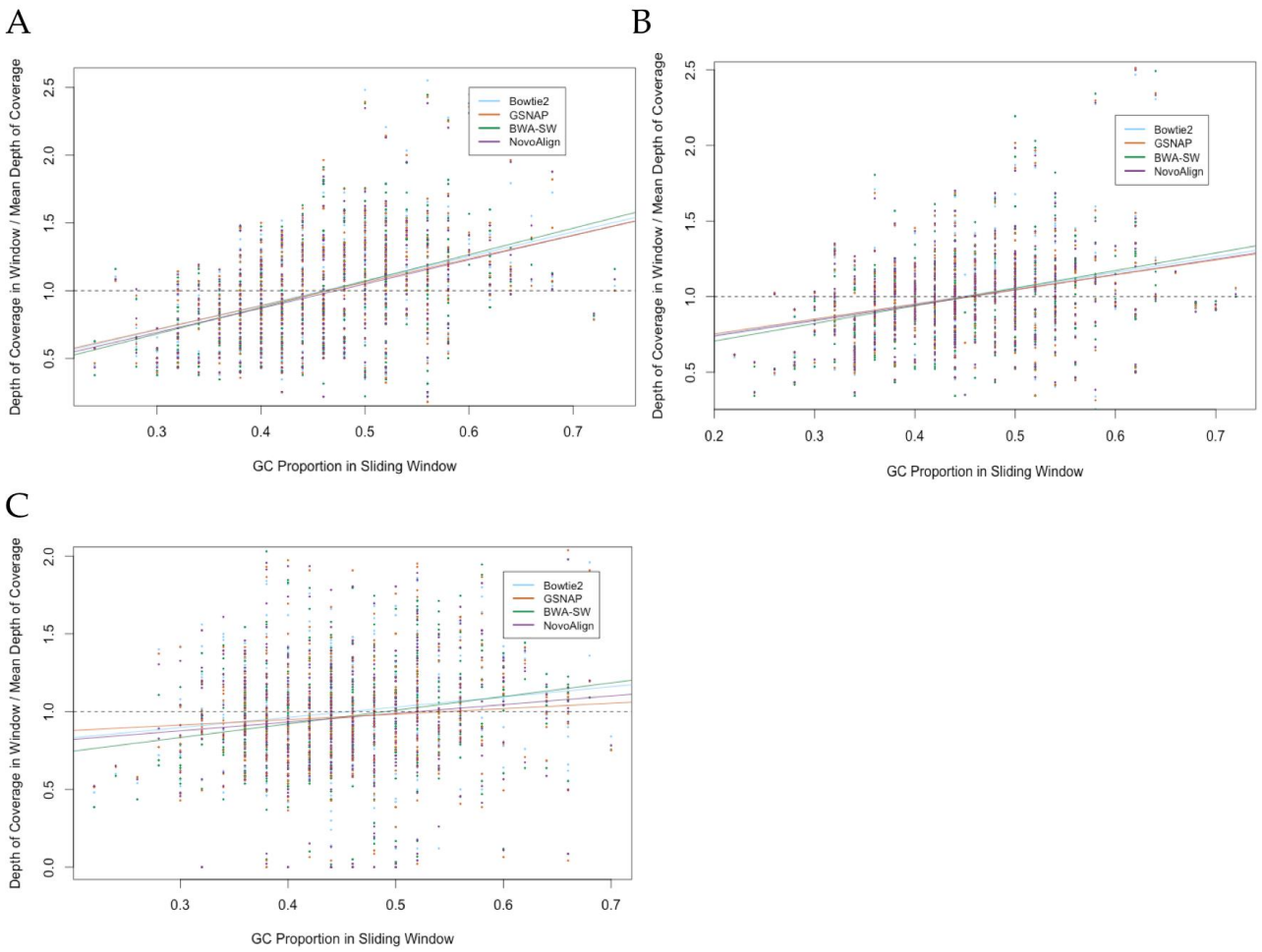



**Fig 3:** Scatter plots showing the GC bias in assembled reads. GC proportion and normalised depth of coverage in each window were plotted for each aligner individually then grouped by patient sample. Plots are shown for patient TD024 (A), TD031 (B) and TD062 (C). A linear regression was fitted to assess the magnitude and direction of the bias. Regression lines are coloured by aligner. Dashed line indicates the expected regression in the absence of any positive or negative GC bias.

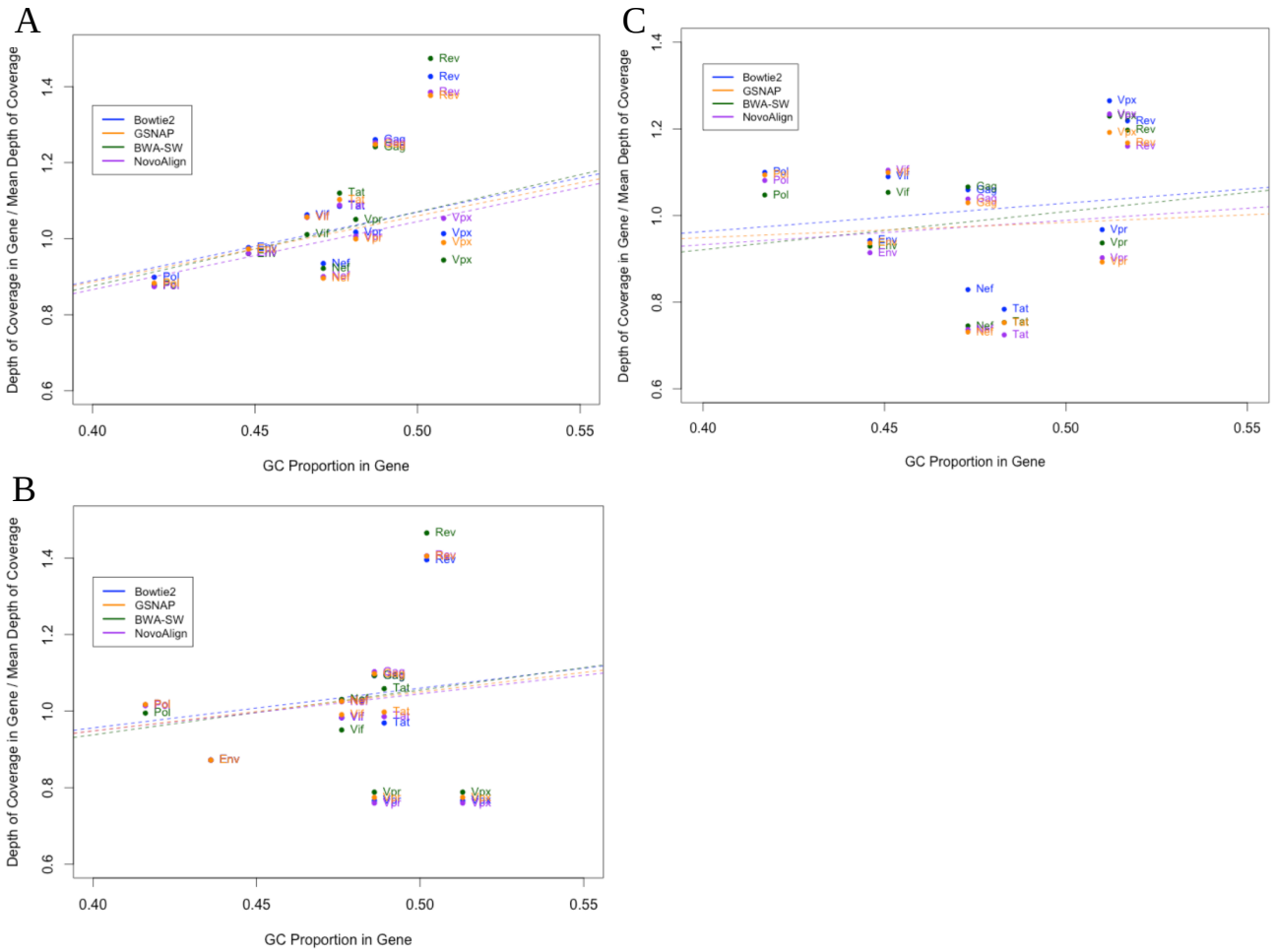



**Fig 4:** Depth of coverage as a function of genomic context. The depth of coverage was calculated individually for each gene of the HIV-2 genome for TD024 (A), TD031 (B) and TD062 (C). Depth of coverage was coloured by aligner and plotted against the mean GC content of the gene. The predicted GC bias is represented by the dashed line.

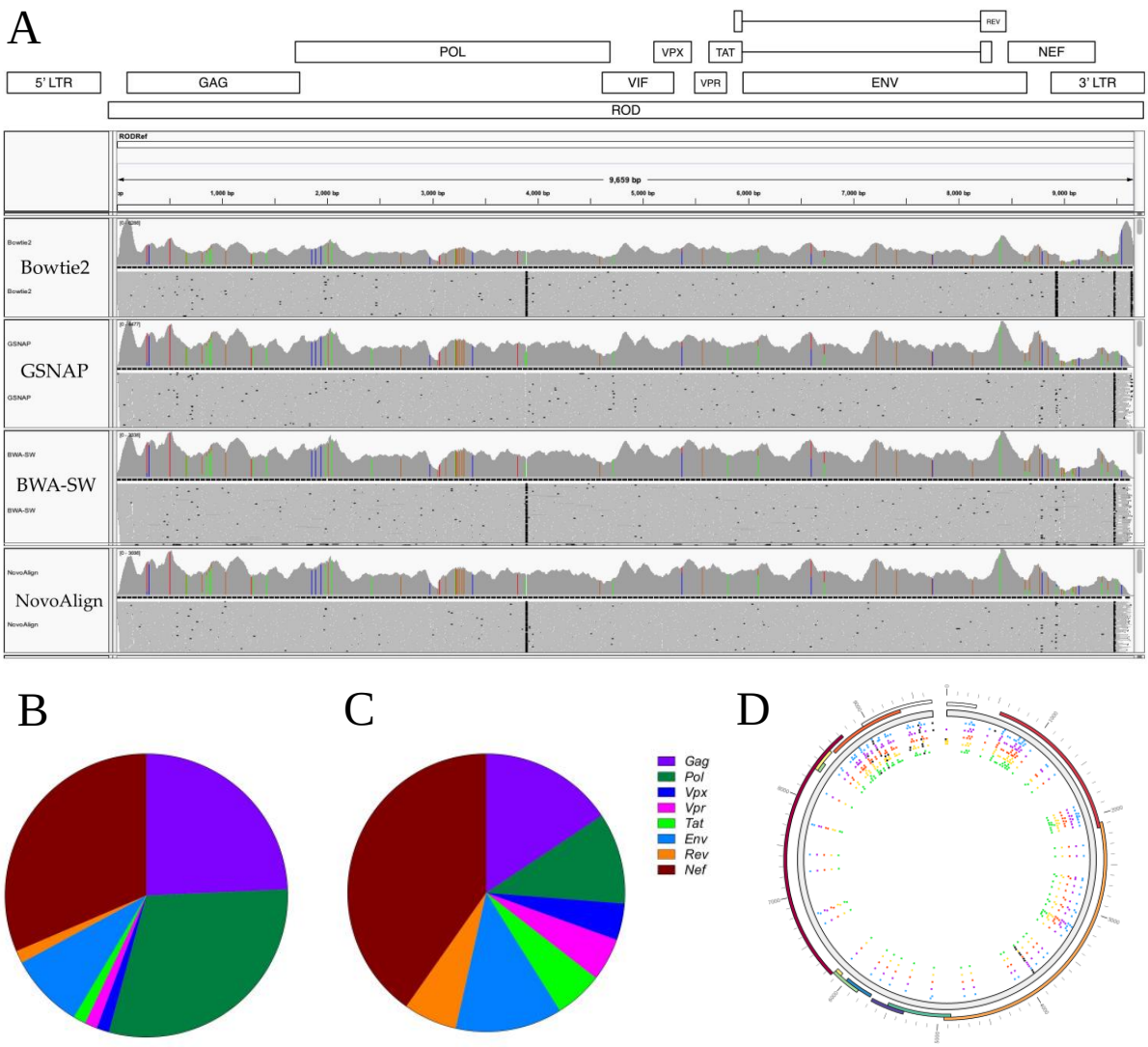



**Fig 5:** Divergence *in vitro* of the lab adapted HIV-2 isolate HIV-2 ROD. Assembled reads were visualised in IGV and mis-matched sites were coloured (A). Sites of conservation with the published reference sequence are shown in grey. SNPs were defined as fixed at a frequency of >95% and the total number of SNPs in each gene were calculated (B). In order to allow for varying gene length, the frequency of SNPs in each gene was also calculated (C). The location of fixed SNPs was visualised (D) for Bowtie2 (blue), BWA-SW (green), GSNAP (orange) and NovoAlign (purple).

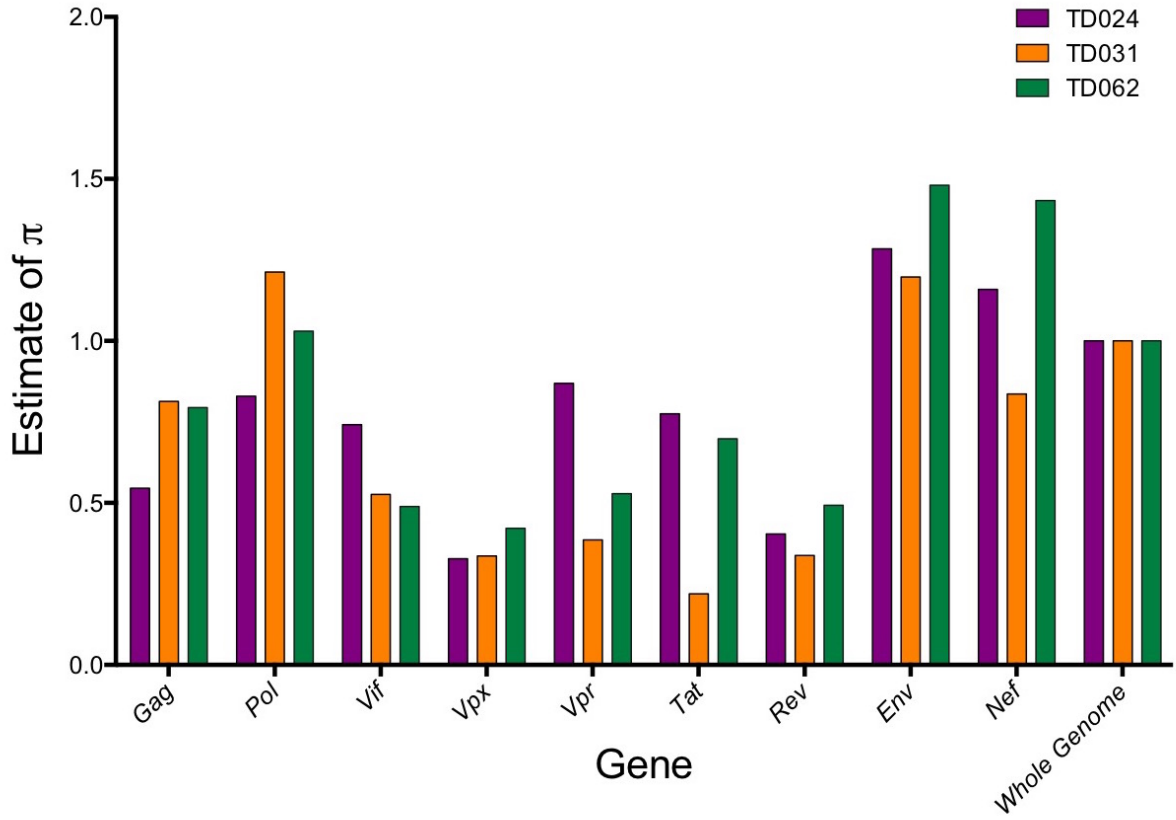



**Fig 6:** Nucleotide site diversity estimates per gene, normalised to the whole genome estimate. Diversity was estimated for samples TD024 (purple), TD031 (orange) and TD062 (green). Calculating the diversity relative to the whole genome estimate as performed to allow a comparison between patients.

**Tables:**

**TABLE 1** Samples included in this study and *de novo* assembly statistics

| Sample ID | Viral copies [a] | Total RNA (ng) [b] | Predicted HIV RNA (%) [c] | Reads aligning to viral reference | Genome covered by all contigs (%) | Genes Intact | Merged contig length (bp) |
|---|---|---|---|---|---|---|---|
| HIV-2 CBL-20 | >10 000 000 | 9.65 | >20 | 930 072 | 87 | 9 | 9885 |
| TD003 | 41 002 | 8.70 | 0.0023 | 0 | 0 | 0 | 0 |
| TD006 | <50 | 7.65 | 0.0000033 | 0 | 0 | 0 | 0 |
| TD013 | 816 | 34.00 | 0.000012 | 0 | 0 | 0 | 0 |
| TD024 | 5 280 | 2.30 | 0.0011 | 4 998 | 93 | 9 | 9531 |
| TD031 | 53 591 | 3.10 | 0.0087 | 9 065 | 90 | 9 | 9397 |
| TD062 | 69 759 | 2.85 | 0.012 | 13 304 | 87 | 9 | 9776 |

[a] Absolute viral input estimated from viral load

[b] Total RNA input used for library preparation

[c] Estimated using a viral genome length of 10 kb and absolute viral input

**TABLE 2** Summary of read mapping to sample-specific reference sequences

| | Bowtie2 | | BWA-SW | | GSNAP | | NovoAlign | |
|---|---|---|---|---|---|---|---|---|
| Sample ID | Mean depth | Reads aligning | Mean depth | Reads aligning | Mean depth | Reads aligning | Mean depth | Reads aligning |
| TD024 | 28.53x | 3709 | 31.90x | 3988 | 27.69x | 3426 | 27.79x | 3463 |
| TD031 | 62.33x | 7658 | 67.23x | 8044 | 60.33x | 7172 | 60.50x | 7267 |
| TD062 | 50.01x | 6617 | 59.61x | 7468 | 45.92x | 5658 | 46.64x | 5751 |
| HIV-2 CBL-20 | 5502x | 539906 | 4734x | 412557 | 6451x | 618464 | 5156x | 432538 |
| HIV-2 ROD | 1924x | 165506 | 1794x | 152105 | 2146x | 176885 | 1862x | 155696 |

**TABLE 3** Summary statistics for the GC% bias present in assembled reads

| | Bowtie2 | | BWA-SW | | GSNAP | | NovoAlign | |
|---|---|---|---|---|---|---|---|---|
| Sample ID | Slope [a] | Intercept [a] | Slope | Intercept | Slope | Intercept | Slope | Intercept |
| TD024 | 1.80 | 0.17 | 1.95 | 0.10 | 1.74 | 0.19 | 1.79 | 0.15 |
| TD031 | 1.04 | 0.54 | 1.17 | 0.47 | 0.97 | 0.56 | 1.02 | 0.54 |
| TD062 | 0.66 | 0.70 | 0.88 | 0.57 | 0.35 | 0.81 | 0.56 | 0.71 |

[a] Estimated by fitting a linear regression to the mean values in each sliding window

**Supplementary Information:**

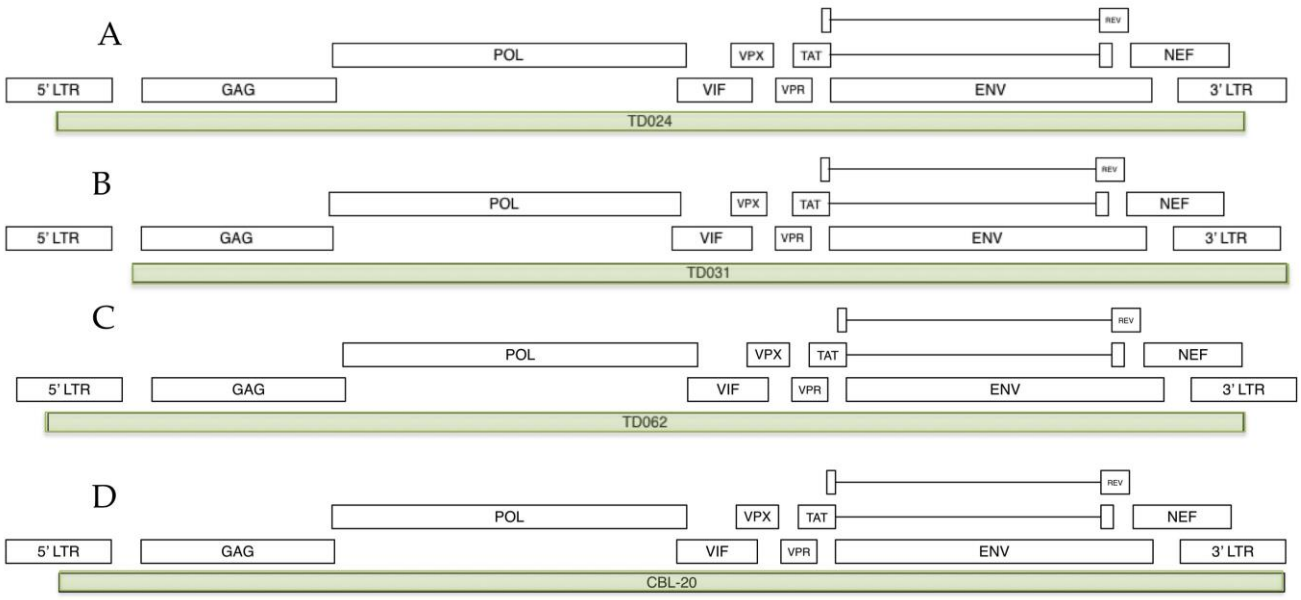
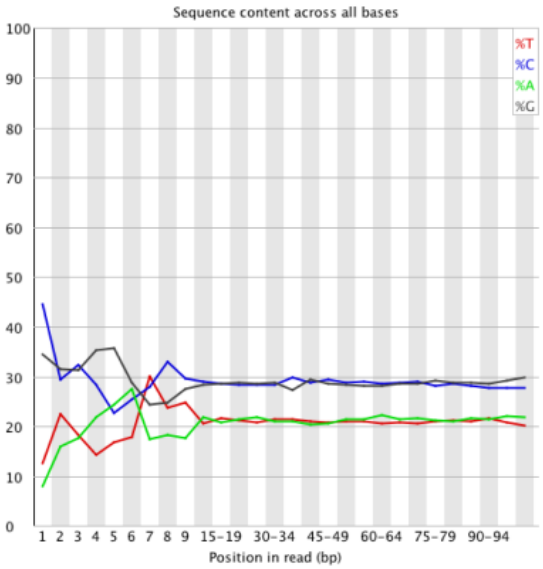
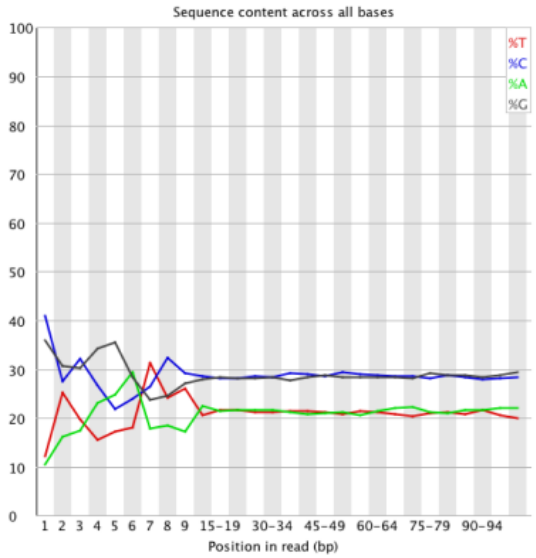



**Fig S1:** Capture of the HIV-2 genome by *de novo* assembly. The length and location of the longest merged contig is shown for TD024 (A), TD031 (B), TD062 (C) and HIV-2 CBL-20

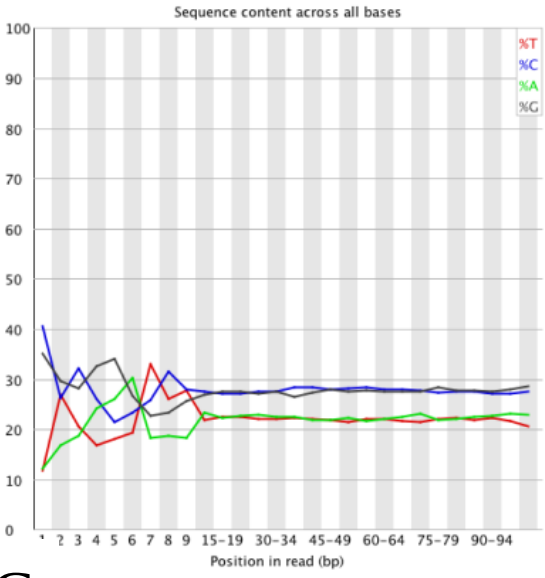
(D). Genes were annotated according to homology with the reference genome UC2. Genome capture for each sample is shown in green. The 3’ and 5’ LTRs are shown in ‘complete’ proviral form as true repeats, each consisting of U3, R and U5 domains.

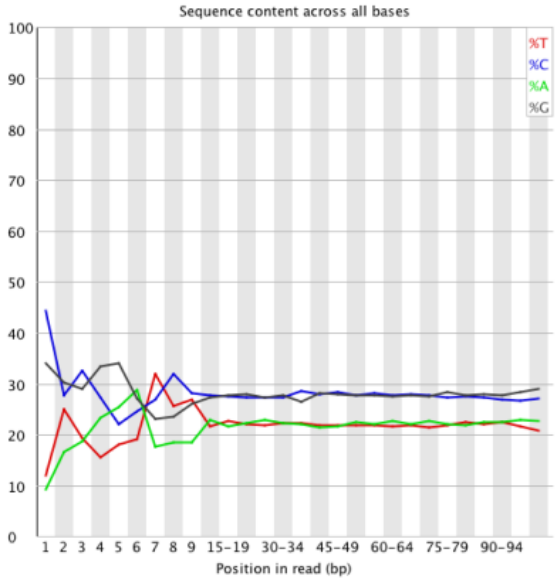
A


Sequence content across all bases
%T
%C
%A
%G
Position in read (bp)

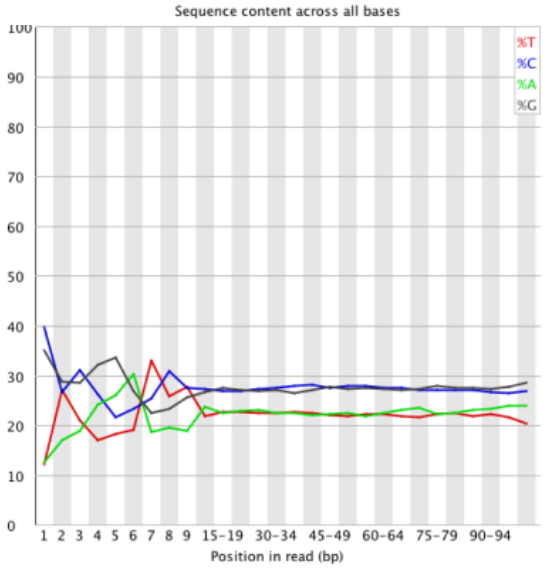

Sequence content across all bases
%T
%C
%A
%G
Position in read (bp)

B


Sequence content across all bases
%T
%C
%A
%G
Position in read (bp)


Sequence content across all bases
%T
%C
%A
%G
Position in read (bp)

C


Sequence content across all bases
%T
%C
%A
%G
Position in read (bp)

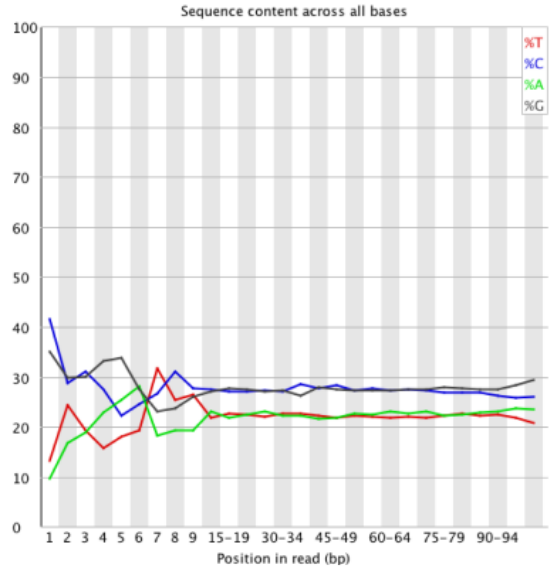

Sequence content across all bases
%T
%C
%A
%G
Position in read (bp)

**Fig S2:** Visualisation of the random hexamer bias. Nucleotide composition per read position was assessed using FastQC for TD024 (A), TD031 (B) and TD062 (C) for both the forward and reverse read mates.

**TABLE S1** Reference HIV-2 genome sequences used in phylogenetic analysis

| Name | Accession number | Country of origin |
|---|---|---|
| A.PT.ALI | AF082339 | Portugal |
| A.DE.PEI2 | U22047 | Germany |
| A.SN.ST_JSP4_27 | M31113 | Senegal |
| A.GM.MCN13 | AY509259 | Gambia |
| A.GM.MCR35 | AY509260 | Gambia |
| A.GW.NIHZ | J03654 | Guinea Bissau |
| A.SN.ROD | BD413542 | Senegal |
| A.GW.CAM2CG | D00835 | Cote D'Ivoire |
| A.IN.CRIK_147 | DQ307022 | India |
| A.IN.NNVA | EU980602 | India |
| A.GW.MDS | Z48731 | Guinea Bissau |
| A.GM.ISY_SBL | J04498 | Gambia |
| A.JP.NMC786_42 | AB731742 | Japan |
| A.JP.NMC786_41 | AB731743 | Japan |
| A.GH.GH1 | E02138 | Ghana |
| A.GM.D194 | A09995 | Gambia |
| A.CI.UC2 | U389293 | Cote D'Ivoire |
| A.DE.BEN | M30502 | Germany |

**TABLE S2** Summary of LTR capture by *de novo* assembly

| | Misplaced LTR fragments (relative to RNA genome) | |
|---|---|---|
| Sample ID | 5' LTR [a] | 3' LTR [a] |
| TD024 | +143bp U3 | -29bp U3 -90bp R |
| TD031 | -Entire LTR -177bp *gag* leader sequence | +80bp U5 |
| TD062 | +319bp U3 | -90bp R |
| CBL-20 | +117bp U3 | +80bp U5 |

[a] Relative to LTR regions expected to be present in the RNA genome.

## References:


1. Centers for Disease Control (CDC). Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men--New York City and California. *MMWR Morb. Mortal. Wkly. Rep.* **30,** 305–308 (1981).
2. Barré-Sinoussi, F. *et al.* Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* **220,** 868–871 (1983).
3. Barin, F. *et al.* Serological evidence for virus related to simian T-lymphotropic retrovirus III in residents of west Africa. *Lancet* **2,** 1387–1389 (1985).
4. Guyader, M. *et al.* Genome organization and transactivation of the human immunodeficiency virus type 2. *Nature* **326,** 662–669 (1987).
5. Clavel, F. *et al.* Isolation of a new human retrovirus from West African patients with AIDS. *Science* **233,** 343–346 (1986).
6. UNAIDS. 2013 Global Fact Sheet. at <http://www.unaids.org/en/media/unaids/contentassets/documents/epidemiology/2013/gr2013/20130923_FactSheet_Global_en.pdf>
7. WHO | Data and statistics. *WHO* at <http://www.who.int/hiv/data/en/>
8. Hahn, B. H. AIDS as a Zoonosis: Scientific and Public Health Implications. *Science* **287,** 607–614 (2000).
9. Sharp, P. M. & Hahn, B. H. Origins of HIV and the AIDS Pandemic. *Cold Spring Harb Perspect Med* **1,** a006841 (2011).
10. Keele, B. F. *et al.* Increased mortality and AIDS-like immunopathology in wild chimpanzees infected with SIVcpz. *Nature* **460,** 515–519 (2009).
11. Kestler, H. *et al.* Induction of AIDS in rhesus monkeys by molecularly cloned simian immunodeficiency virus. *Science* **248,** 1109–1112 (1990).
12. Apetrei, C. *et al.* Molecular epidemiology of simian immunodeficiency virus SIVsm in U.S. primate centers unravels the origin of SIVmac and SIVstm. *J. Virol.* **79,** 8991–9005 (2005).
13. Bailes, E. Hybrid Origin of SIV in Chimpanzees. *Science* **300,** 1713–1713 (2003).
14. Mitani, J. C. & Watts, D. P. Demographic influences on the hunting behavior of chimpanzees. *American Journal of Physical Anthropology* **109,** 439–454 (1999).
15. Locatelli, S. & Peeters, M. Non-Human Primates, Retroviruses, and Zoonotic Infection Risks in the Human Population. *Nature Education Knowledge* **3,** 62 (2012).
16. Worobey, M. *et al.* Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* **455,** 661–664 (2008).
17. Lemey, P. The Molecular Population Genetics of HIV-1 Group O. *Genetics* **167,** 1059–1068 (2004).
18. Vallari, A. *et al.* Four new HIV-1 group N isolates from Cameroon: Prevalence continues to be low. *AIDS Res. Hum. Retroviruses* **26,** 109–115 (2010).
19. Keele, B. F. *et al.* Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* **313,** 523–526 (2006).
20. Plantier, J.-C. *et al.* A new human immunodeficiency virus derived from gorillas. *Nat. Med.* **15,** 871–872 (2009).

21. Ho, D. D. *et al.* Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **373,** 123–126 (1995).
22. Preston, B. D., Poiesz, B. J. & Loeb, L. A. Fidelity of HIV-1 reverse transcriptase. *Science* **242,** 1168–1171 (1988).
23. Chen, J., Powell, D. & Hu, W.-S. High frequency of genetic recombination is a common feature of primate lentivirus replication. *J. Virol.* **80,** 9651–9658 (2006).
24. Los Alamos Laboratories. HIV Sequence Databases. at <http://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html>
25. Pepin, J. *The Origins of AIDS*. (Cambridge University Press, 2011).
26. IAVI. IAVIReports. at <www.iavireport.org>
27. Gao, F. *et al.* Human infection by genetically diverse SIVSM-related HIV-2 in west Africa. *Nature* **358,** 495–499 (1992).
28. Lemey, P. *et al.* Tracing the origin and history of the HIV-2 epidemic. *Proc. Natl. Acad. Sci. U.S.A.* **100,** 6588–6592 (2003).
29. Silvestri, G. *et al.* Nonpathogenic SIV infection of sooty mangabeys is characterized by limited bystander immunopathology despite chronic high-level viremia. *Immunity* **18,** 441–452 (2003).
30. Santiago, M. L. *et al.* Simian immunodeficiency virus infection in free-ranging sooty mangabeys (Cercocebus atys atys) from the Taï Forest, Côte d'Ivoire: implications for the origin of epidemic human immunodeficiency virus type 2. *J. Virol.* **79,** 12515–12527 (2005).
31. Damond, F. *et al.* Identification of a highly divergent HIV type 2 and proposal for a change in HIV type 2 classification. *AIDS Res. Hum. Retroviruses* **20,** 666–672 (2004).
32. Chen, Z. *et al.* Human immunodeficiency virus type 2 (HIV-2) seroprevalence and characterization of a distinct HIV-2 genetic subtype from the natural range of simian immunodeficiency virus-infected sooty mangabeys. *J. Virol.* **71,** 3953–3960 (1997).
33. Pieniazek, D. *et al.* Predominance of human immunodeficiency virus type 2 subtype B in Abidjan, Ivory Coast. *AIDS Res. Hum. Retroviruses* **15,** 603–608 (1999).
34. Peeters, M., Toure-Kane, C. & Nkengasong, J. N. Genetic diversity of HIV in Africa: impact on diagnosis, treatment, vaccine development and trials. *AIDS* **17,** 2547–2560 (2003).
35. Smith, S. M. *et al.* Isolation of a new HIV-2 group in the US. *Retrovirology* **5,** 103 (2008).
36. Ayouba, A. *et al.* Evidence for continuing cross-species transmission of SIVsmm to humans: characterization of a new HIV-2 lineage in rural Côte d'Ivoire. *AIDS* **27,** 2488–2491 (2013).
37. Ibe, S. *et al.* HIV-2 CRF01_AB: first circulating recombinant form of HIV-2. *J. Acquir. Immune Defic. Syndr.* **54,** 241–247 (2010).
38. Gao, F. *et al.* Genetic diversity of human immunodeficiency virus type 2: evidence for distinct sequence subtypes with differences in virus biology. *J. Virol.* **68,** 7433–7447 (1994).

39. Hirsch, V. M., Olmsted, R. A., Murphey-Corb, M., Purcell, R. H. & Johnson, P. R. An African primate lentivirus (SIVsm) closely related to HIV-2. *Nature* **339,** 389–392 (1989).
40. Faria, N. R. *et al.* Phylogeographical footprint of colonial history in the global dispersal of human immunodeficiency virus type 2 group A. *J Gen Virol* **93,** 889–899 (2012).
41. Collier, L., Kellam, P. & John, O. *Human Virology 4th Edition*. (Oxford University Press).
42. London School of Hygiene and Tropical Medicine. AIDS-Section 1: Virology, Immunology and Diagnosis of HIV Infection. at <http://idshowcase.lshtm.ac.uk/id501/ID501/S1S2/ID501_S1S2_050_010.html>
43. Gelderblom, H. R., Hausmann, E. H., Ozel, M., Pauli, G. & Koch, M. A. Fine structure of human immunodeficiency virus (HIV) and immunolocalization of structural proteins. *Virology* **156,** 171–176 (1987).
44. Gelderblom, H. R. Assembly and morphology of HIV: potential effect of structure on viral function. *AIDS* **5,** 617–637 (1991).
45. Zhu, P. *et al.* Distribution and three-dimensional structure of AIDS virus envelope spikes. *Nature* **441,** 847–852 (2006).
46. Wei, X. *et al.* Antibody neutralization and escape by HIV-1. *Nature* **422,** 307–312 (2003).
47. Gao, F. Amplification and cloning of near full-length HIV-2 genomes. *Methods Mol. Biol.* **304,** 399–407 (2005).
48. Gonda, M. A. Molecular genetics and structure of the human immunodeficiency virus. *J Electron Microsc Tech* **8,** 17–40 (1988).
49. Schaller, T., Bauby, H., Hué, S., Malim, M. H. & Goujon, C. New insights into an X-traordinary viral protein. *Front Microbiol* **5,** 126 (2014).
50. Pachulska-Wieczorek, K., Stefaniak, A. K. & Purzycka, K. J. Similarities and differences in the nucleic acid chaperone activity of HIV-2 and HIV-1 nucleocapsid proteins in vitro. *Retrovirology* **11,** 54 (2014).
51. Fujita, M. [Study of molecular function of proteins in human immunodeficiency virus]. *Yakugaku Zasshi* **133,** 1103–1111 (2013).
52. Bahraoui, E. *et al.* Study of the interaction of HIV-1 and HIV-2 envelope glycoproteins with the CD4 receptor and role of N-glycans. *AIDS Res. Hum. Retroviruses* **8,** 565–573 (1992).
53. Soto-Rifo, R. *et al.* Different effects of the TAR structure on HIV-1 and HIV-2 genomic RNA translation. *Nucleic Acids Res.* **40,** 2653–2667 (2012).
54. Lusvarghi, S. *et al.* The HIV-2 Rev-response element: determining secondary structure and defining folding intermediates. *Nucleic Acids Res.* **41,** 6637–6649 (2013).
55. Smith, J. L., Izumi, T., Borbet, T. C., Hagedorn, A. N. & Pathak, V. K. HIV-1 and HIV-2 Vif Interact with Human APOBEC3 Proteins Using Completely Different Determinants. *J. Virol.* **88,** 9893–9908 (2014).
56. Philippon, V., Matsuda, Z. & Essex, M. Transactivation is a conserved function among primate lentivirus Vpr proteins but is not shared by Vpx. *J. Hum. Virol.* **2,** 167–174 (1999).

57. Laguette, N. *et al.* SAMHD1 is the dendritic- and myeloid-cell-specific HIV-1 restriction factor counteracted by Vpx. *Nature* **474,** 654–657 (2011).
58. Berger, A. *et al.* Interaction of Vpx and apolipoprotein B mRNA-editing catalytic polypeptide 3 family member A (APOBEC3A) correlates with efficient lentivirus infection of monocytes. *J. Biol. Chem.* **285,** 12248–12254 (2010).
59. Cheng, X. & Ratner, L. HIV-2 Vpx protein interacts with interferon regulatory factor 5 (IRF5) and inhibits its function. *J. Biol. Chem.* **289,** 9146–9157 (2014).
60. Lindwasser, O. W., Chaudhuri, R. & Bonifacino, J. S. Mechanisms of CD4 downregulation by the Nef and Vpu proteins of primate immunodeficiency viruses. *Curr. Mol. Med.* **7,** 171–184 (2007).
61. DeGottardi, M. Q. *et al.* Selective downregulation of rhesus macaque and sooty mangabey major histocompatibility complex class I molecules by Nef alleles of simian immunodeficiency virus and human immunodeficiency virus type 2. *J. Virol.* **82,** 3139–3146 (2008).
62. Zheng, N. N. *et al.* Role of human immunodeficiency virus (HIV)-specific T-cell immunity in control of dual HIV-1 and HIV-2 infection. *J. Virol.* **81,** 9061–9071 (2007).
63. Rambaut, A., Posada, D., Crandall, K. A. & Holmes, E. C. The causes and consequences of HIV evolution. *Nature Reviews Genetics* **5,** 52–61 (2004).
64. Thomas, E. R., Shotton, C., Weiss, R. A., Clapham, P. R. & McKnight, A. CD4-dependent and CD4-independent HIV-2: consequences for neutralization. *AIDS* **17,** 291–300 (2003).
65. McKnight, A. *et al.* A broad range of chemokine receptors are used by primary isolates of human immunodeficiency virus type 2 as coreceptors with CD4. *J. Virol.* **72,** 4065–4071 (1998).
66. Blaak, H. *et al.* CCR5, GPR15, and CXCR6 Are Major Coreceptors of Human Immunodeficiency Virus Type 2 Variants Isolated from Individuals with and without Plasma Viremia. *J Virol* **79,** 1686–1700 (2005).
67. Calado, M. *et al.* Coreceptor usage by HIV-1 and HIV-2 primary isolates: the relevance of CCR8 chemokine receptor as an alternative coreceptor. *Virology* **408,** 174–182 (2010).
68. Mild, M. *et al.* High intrapatient HIV-1 evolutionary rate is associated with CCR5-to-CXCR4 coreceptor switch. *Infect. Genet. Evol.* **19,** 369–377 (2013).
69. Dean, M. *et al.* Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study. *Science* **273,** 1856–1862 (1996).
70. Li, C. *et al.* Frequency of the CCR5 delta 32 mutant allele in HIV-1-positive patients, female sex workers, and a normal population in Taiwan. *J. Formos. Med. Assoc.* **96,** 979–984 (1997).
71. Visseaux, B. *et al.* Molecular determinants of HIV-2 R5-X4 tropism in the V3 loop: development of a new genotypic tool. *J. Infect. Dis.* **205,** 111–120 (2012).

72. Shchelokovskyy, P., Tristram-Nagle, S. & Dimova, R. Effect of the HIV-1 fusion peptide on the mechanical properties and leaflet coupling of lipid bilayers. *New J Phys* **13,** 25004 (2011).
73. Greene, W. C. & Peterlin, B. M. Charting HIV's remarkable voyage through the cell: Basic science as a passport to future therapy. *Nat Med* **8,** 673–680 (2002).
74. Rasaiyaah, J. *et al.* HIV-1 evades innate immune recognition through specific cofactor recruitment. *Nature* **503,** 402–405 (2013).
75. Warrilow, D., Tachedjian, G. & Harrich, D. Maturation of the HIV reverse transcription complex: putting the jigsaw together. *Rev. Med. Virol.* **19,** 324–337 (2009).
76. Skar, H. *et al.* HIV-2 genetic evolution in patients with advanced disease is faster than that in matched HIV-1 patients. *J. Virol.* **84,** 7412–7415 (2010).
77. Bakhanashvili, M. & Hizi, A. Fidelity of the reverse transcriptase of human immunodeficiency virus type 2. *FEBS Lett.* **306,** 151–156 (1992).
78. Lanchy, J.-M. & Lodmell, J. S. Alternate usage of two dimerization initiation sites in HIV-2 viral RNA in vitro. *J. Mol. Biol.* **319,** 637–648 (2002).
79. Lanchy, J.-M., Rentz, C. A., Ivanovitch, J. D. & Lodmell, J. S. Elements located upstream and downstream of the major splice donor site influence the ability of HIV-2 leader RNA to dimerize in vitro. *Biochemistry* **42,** 2634–2642 (2003).
80. Dirac, A. M. G., Huthoff, H., Kjems, J. & Berkhout, B. Regulated HIV-2 RNA dimerization by means of alternative RNA conformations. *Nucleic Acids Res.* **30,** 2647–2655 (2002).
81. Paillart, J.-C., Shehu-Xhilaga, M., Marquet, R. & Mak, J. Dimerization of retroviral RNA genomes: an inseparable pair. *Nat. Rev. Microbiol.* **2,** 461–472 (2004).
82. Hrecka, K. *et al.* Vpx relieves inhibition of HIV-1 infection of macrophages mediated by the SAMHD1 protein. *Nature* **474,** 658–661 (2011).
83. Schauer, G., Leuba, S. & Sluis-Cremer, N. Biophysical Insights into the Inhibitory Mechanism of Non-Nucleoside HIV-1 Reverse Transcriptase Inhibitors. *Biomolecules* **3,** 889–904 (2013).
84. Nisole, S. & Saïb, A. Early steps of retrovirus replicative cycle. *Retrovirology* **1,** 9 (2004).
85. Zheng, Y.-H., Lovsin, N. & Peterlin, B. M. Newly identified host factors modulate HIV replication. *Immunol. Lett.* **97,** 225–234 (2005).
86. Phillips, A. N. *et al.* Serial CD4 lymphocyte counts and development of AIDS. *Lancet* **337,** 389–392 (1991).
87. Van der Loeff, M. F. S. *et al.* Undetectable plasma viral load predicts normal survival in HIV-2-infected people in a West African village. *Retrovirology* **7,** 46 (2010).
88. Ariyoshi, K. *et al.* Plasma RNA viral load predicts the rate of CD4 T cell decline and death in HIV-2-infected patients in West Africa. *AIDS* **14,** 339–344 (2000).
89. Clavel, F. *et al.* Human immunodeficiency virus type 2 infection associated with AIDS in West Africa. *N. Engl. J. Med.* **316,** 1180–1185 (1987).

90. Martinez-Steele, E. *et al.* Is HIV-2- induced AIDS different from HIV-1-associated AIDS? Data from a West African clinic. *AIDS* **21,** 317–324 (2007).
91. Van Tienen, C. *et al.* HTLV-1 and HIV-2 infection are associated with increased mortality in a rural West African community. *PLoS ONE* **6,** e29026 (2011).
92. Thiébaut, R. *et al.* Long-term nonprogressors and elite controllers in the ANRS CO5 HIV-2 cohort. *AIDS* **25,** 865–867 (2011).
93. Barker, E. *et al.* Virological and immunological features of long-term human immunodeficiency virus-infected individuals who have remained asymptomatic compared with those who have progressed to acquired immunodeficiency syndrome. *Blood* **92,** 3105–3114 (1998).
94. Berry, N. *et al.* Low peripheral blood viral HIV-2 RNA in individuals with high CD4 percentage differentiates HIV-2 from HIV-1 infection. *J. Hum. Virol.* **1,** 457–468 (1998).
95. Berry, N. *et al.* Low level viremia and high CD4% predict normal survival in a cohort of HIV type-2-infected villagers. *AIDS Res. Hum. Retroviruses* **18,** 1167–1173 (2002).
96. Hansmann, A. *et al.* Baseline plasma viral load and CD4 cell percentage predict survival in HIV-1- and HIV-2-infected women in a community-based cohort in The Gambia. *J. Acquir. Immune Defic. Syndr.* **38,** 335–341 (2005).
97. Gottlieb, G. S. *et al.* Equal plasma viral loads predict a similar rate of CD4+ T cell decline in human immunodeficiency virus (HIV) type 1- and HIV-2-infected individuals from Senegal, West Africa. *J. Infect. Dis.* **185,** 905–914 (2002).
98. Popper, S. J. *et al.* Low plasma human immunodeficiency virus type 2 viral load is independent of proviral load: low virus production in vivo. *J. Virol.* **74,** 1554–1557 (2000).
99. Ariyoshi, K. *et al.* A community-based study of human immunodeficiency virus type 2 provirus load in rural village in West Africa. *J. Infect. Dis.* **173,** 245–248 (1996).
100. Popper, S. J. *et al.* Lower human immunodeficiency virus (HIV) type 2 viral load reflects the difference in pathogenicity of HIV-1 and HIV-2. *J. Infect. Dis.* **180,** 1116–1121 (1999).
101. Cortes, E. *et al.* HIV-1, HIV-2, and HTLV-I infection in high-risk groups in Brazil. *N. Engl. J. Med.* **320,** 953–958 (1989).
102. Kannangai, R., Nair, S. C., Sridharan, G., Prasannakumar, S. & Daniel, D. Frequency of HIV type 2 infections among blood donor population from India: a 10-year experience. *Indian J Med Microbiol* **28,** 111–113 (2010).
103. Schim van der Loeff, M. F. & Aaby, P. Towards a better understanding of the epidemiology of HIV-2. *AIDS* **13 Suppl A,** S69–84 (1999).
104. Poulsen, A. G. *et al.* HIV-2 infection in Bissau, West Africa, 1987-1989: incidence, prevalences, and routes of transmission. *J. Acquir. Immune Defic. Syndr.* **6,** 941–948 (1993).

105. Da Silva, Z. J. *et al.* Changes in prevalence and incidence of HIV-1, HIV-2 and dual infections in urban areas of Bissau, Guinea-Bissau: is HIV-2 disappearing? *AIDS* **22,** 1195–1202 (2008).
106. Hamel, D. J. *et al.* Twenty years of prospective molecular epidemiology in Senegal: changes in HIV diversity. *AIDS Res. Hum. Retroviruses* **23,** 1189–1196 (2007).
107. Tienen, C. van *et al.* Two distinct epidemics: the rise of HIV-1 and decline of HIV-2 infection between 1990 and 2007 in rural Guinea-Bissau. *J. Acquir. Immune Defic. Syndr.* **53,** 640–647 (2010).
108. Fryer, H. *et al.* Predicting the Extinction of HIV-2 in Guinea-Bissau. *21st Conference on Retroviruses and Opportunistic Infections, Boston, USA 2014* **Abstract 575,**
109. O'Donovan, D. *et al.* Maternal plasma viral RNA levels determine marked differences in mother-to-child transmission rates of HIV-1 and HIV-2 in The Gambia. MRC/Gambia Government/University College London Medical School working group on mother-child transmission of HIV. *AIDS* **14,** 441–448 (2000).
110. Andersson, S. *et al.* Plasma viral load in HIV-1 and HIV-2 singly and dually infected individuals in Guinea-Bissau, West Africa: significantly lower plasma virus set point in HIV-2 infection than in HIV-1 infection. *Arch. Intern. Med.* **160,** 3286–3293 (2000).
111. Hawes, S. E. *et al.* Lower levels of HIV-2 than HIV-1 in the female genital tract: correlates and longitudinal assessment of viral shedding. *AIDS* **22,** 2517–2525 (2008).
112. Gottlieb, G. S. *et al.* Lower levels of HIV RNA in semen in HIV-2 compared with HIV-1 infection: implications for differences in transmission. *AIDS* **20,** 895–900 (2006).
113. Poulsen, A. G., Aaby, P., Jensen, H. & Dias, F. Risk factors for HIV-2 seropositivity among older people in Guinea-Bissau. A search for the early history of HIV-2 infection. *Scand. J. Infect. Dis.* **32,** 169–175 (2000).
114. Pépin, J., Plamondon, M., Alves, A. C., Beaudet, M. & Labbé, A.-C. Parenteral transmission during excision and treatment of tuberculosis and trypanosomiasis may be responsible for the HIV-2 epidemic in Guinea-Bissau. *AIDS* **20,** 1303–1311 (2006).
115. Schim van der Loeff, M. F. *et al.* Mortality of HIV-1, HIV-2 and HIV-1/HIV-2 dually infected patients in a clinic-based cohort in The Gambia. *AIDS* **16,** 1775–1783 (2002).
116. Aaby, P. *et al.* Age of wife as a major determinant of male-to-female transmission of HIV-2 infection: a community study from rural West Africa. *AIDS* **10,** 1585–1590 (1996).
117. Schmidt, W. P. *et al.* Behaviour change and competitive exclusion can explain the diverging HIV-1 and HIV-2 prevalence trends in Guinea-Bissau. *Epidemiol. Infect.* **136,** 551–561 (2008).

118. Norrgren, H. *et al.* Trends and interaction of HIV-1 and HIV-2 in Guinea-Bissau, west Africa: no protection of HIV-2 against HIV-1 infection. *AIDS* **13,** 701–707 (1999).
119. Greenberg, A. E. Possible protective effect of HIV-2 against incident HIV-1 infection: review of available epidemiological and in vitro data. *AIDS* **15,** 2319–2321 (2001).
120. Schim van der Loeff, M. F. *et al.* HIV-2 does not protect against HIV-1 infection in a rural community in Guinea-Bissau. *AIDS* **15,** 2303–2310 (2001).
121. Esbjörnsson, J. *et al.* Inhibition of HIV-1 disease progression by contemporaneous HIV-2 infection. *N. Engl. J. Med.* **367,** 224–232 (2012).
122. Esbjörnsson, J. *et al.* Increased survival among HIV-1 and HIV-2 dual-infected individuals compared to HIV-1 single-infected individuals. *AIDS* **28,** 949–957 (2014).
123. Al-Harthi, L., Owais, M. & Arya, S. K. Molecular inhibition of HIV type 1 by HIV type 2: effectiveness in peripheral blood mononuclear cells. *AIDS Res. Hum. Retroviruses* **14,** 59–64 (1998).
124. Kokkotou, E. G. *et al.* In vitro correlates of HIV-2-mediated HIV-1 protection. *Proc. Natl. Acad. Sci. U.S.A.* **97,** 6797–6802 (2000).
125. Prince, P. D., Matser, A., van Tienen, C., Whittle, H. C. & Schim van der Loeff, M. F. Mortality rates in people dually infected with HIV-1/2 and those infected with either HIV-1 or HIV-2: a systematic review and meta-analysis. *AIDS* **28,** 549–558 (2014).
126. Dunham, R. *et al.* The AIDS resistance of naturally SIV-infected sooty mangabeys is independent of cellular immunity to the virus. *Blood* **108,** 209–217 (2006).
127. Giorgi, J. V. *et al.* Shorter survival in advanced human immunodeficiency virus type 1 infection is more closely associated with T lymphocyte activation than with plasma virus burden or virus chemokine coreceptor usage. *J. Infect. Dis.* **179,** 859–870 (1999).
128. Jaffar, S. *et al.* Immunological predictors of survival in HIV type 2-infected rural villagers in Guinea-Bissau. *AIDS Res. Hum. Retroviruses* **21,** 560–564 (2005).
129. Leligdowicz, A. *et al.* Direct relationship between virus load and systemic immune activation in HIV-2 infection. *J. Infect. Dis.* **201,** 114–122 (2010).
130. Nyamweya, S. *et al.* Are plasma biomarkers of immune activation predictive of HIV progression: a longitudinal comparison and analyses in HIV-1 and HIV-2 infections? *PLoS ONE* **7,** e44411 (2012).
131. Schindler, M. *et al.* Nef-mediated suppression of T cell activation was lost in a lentiviral lineage that gave rise to HIV-1. *Cell* **125,** 1055–1067 (2006).
132. Feldmann, J. *et al.* Downregulation of the T-cell receptor by human immunodeficiency virus type 2 Nef does not protect against disease progression. *J. Virol.* **83,** 12968–12972 (2009).
133. Khalid, M. *et al.* Efficient Nef-mediated downmodulation of TCR-CD3 and CD28 is associated with high CD4+ T cell counts in viremic HIV-2 infection. *J. Virol.* **86,** 4906–4920 (2012).

134. Michel, P. *et al.* Reduced immune activation and T cell apoptosis in human immunodeficiency virus type 2 compared with type 1: correlation of T cell apoptosis with beta2 microglobulin concentration and disease evolution. *J. Infect. Dis.* **181,** 64–75 (2000).
135. Hegedus, A. *et al.* Protection Versus Pathology in Aviremic and High Viral Load HIV-2 Infection--The Pivotal Role of Immune Activation and T-cell Kinetics. *J. Infect. Dis.* **210**, 752-761 (2014).
136. Björling, E. *et al.* Autologous neutralizing antibodies prevail in HIV-2 but not in HIV-1 infection. *Virology* **193,** 528–530 (1993).
137. Shi, Y. *et al.* Evolution of human immunodeficiency virus type 2 coreceptor usage, autologous neutralization, envelope sequence and glycosylation. *J. Gen. Virol.* **86,** 3385–3396 (2005).
138. Blaak, H., van der Ende, M. E., Boers, P. H. M., Schuitemaker, H. & Osterhaus, A. D. M. E. In vitro replication capacity of HIV-2 variants from long-term aviremic individuals. *Virology* **353,** 144–154 (2006).
139. MacNeil, A. *et al.* Long-term intrapatient viral evolution during HIV-2 infection. *J. Infect. Dis.* **195,** 726–733 (2007).
140. Rodriguez, S. K. *et al.* Comparison of heterologous neutralizing antibody responses of human immunodeficiency virus type 1 (HIV-1)- and HIV-2-infected Senegalese patients: distinct patterns of breadth and magnitude distinguish HIV-1 and HIV-2 infections. *J. Virol.* **81,** 5331–5338 (2007).
141. Tendeiro, R. *et al.* Memory B-cell depletion is a feature of HIV-2 infection even in the absence of detectable viremia. *AIDS* **26,** 1607–1617 (2012).
142. Borrow, P., Lewicki, H., Hahn, B. H., Shaw, G. M. & Oldstone, M. B. Virus-specific CD8+ cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. *J. Virol.* **68,** 6103–6110 (1994).
143. Ogg, G. S. *et al.* Quantitation of HIV-1-specific cytotoxic T lymphocytes and plasma load of viral RNA. *Science* **279,** 2103–2106 (1998).
144. Schmitz, J. E. *et al.* Control of viremia in simian immunodeficiency virus infection by CD8+ lymphocytes. *Science* **283,** 857–860 (1999).
145. Migueles, S. A. *et al.* HLA B*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors. *Proc. Natl. Acad. Sci. U.S.A.* **97,** 2709–2714 (2000).
146. International HIV Controllers Study *et al.* The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* **330,** 1551–1557 (2010).
147. Goulder, P. J. *et al.* Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS. *Nat. Med.* **3,** 212–217 (1997).
148. Moore, C. B. *et al.* Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science* **296,** 1439–1443 (2002).
149. Duvall, M. G. *et al.* Maintenance of HIV-specific CD4+ T cell help distinguishes HIV-2 from HIV-1 infection. *J. Immunol.* **176,** 6973–6981 (2006).

150. Duvall, M. G. *et al.* Polyfunctional T cell responses are a hallmark of HIV-2 infection. *Eur. J. Immunol.* **38,** 350–363 (2008).
151. Leligdowicz, A. *et al.* Robust Gag-specific T cell responses characterize viremia control in HIV-2 infection. *J. Clin. Invest.* **117,** 3067–3074 (2007).
152. Leligdowicz, A. *et al.* Highly avid, oligoclonal, early-differentiated antigen-specific CD8+ T cells in chronic HIV-2 infection. *Eur. J. Immunol.* **40,** 1963–1972 (2010).
153. De Silva, T. I. *et al.* Correlates of T-cell-mediated viral control and phenotype of CD8+ T cells in HIV-2, a naturally contained human retroviral infection. *Blood* **121,** 4330–4339 (2013).
154. Appay, V. *et al.* Memory CD8+ T cells vary in differentiation phenotype in different persistent virus infections. *Nat. Med.* **8,** 379–385 (2002).
155. Hu, J. *et al.* Characterization and comparison of recombinant simian immunodeficiency virus from drill (Mandrillus leucophaeus) and mandrill (Mandrillus sphinx) isolates. *J. Virol.* **77,** 4867–4880 (2003).
156. Henderson, L. E., Sowder, R. C., Copeland, T. D., Benveniste, R. E. & Oroszlan, S. Isolation and characterization of a novel protein (X-ORF product) from SIV and HIV-2. *Science* **241,** 199–201 (1988).
157. Yu, X. F., Ito, S., Essex, M. & Lee, T. H. A naturally immunogenic virion-associated protein specific for HIV-2 and SIV. *Nature* **335,** 262–265 (1988).
158. Tristem, M., Marshall, C., Karpas, A. & Hill, F. Evolution of the primate lentiviruses: evidence from vpx and vpr. *EMBO J.* **11,** 3405–3412 (1992).
159. Wong-Staal, F., Chanda, P. K. & Ghrayeb, J. Human immunodeficiency virus: the eighth gene. *AIDS Res. Hum. Retroviruses* **3,** 33–39 (1987).
160. Hu, W., Vander Heyden, N. & Ratner, L. Analysis of the function of viral protein X (VPX) of HIV-2. *Virology* **173,** 624–630 (1989).
161. Shibata, R. *et al.* Mutational analysis of the human immunodeficiency virus type 2 (HIV-2) genome in relation to HIV-1 and simian immunodeficiency virus SIV (AGM). *J. Virol.* **64,** 742–747 (1990).
162. Park, I. W. & Sodroski, J. Functional analysis of the vpx, vpr, and nef genes of simian immunodeficiency virus. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.* **8,** 335–344 (1995).
163. Ueno, F. *et al.* Vpx and Vpr proteins of HIV-2 up-regulate the viral infectivity by a distinct mechanism in lymphocytic cells. *Microbes Infect.* **5,** 387–395 (2003).
164. Gibbs, J. S., Regier, D. A. & Desrosiers, R. C. Construction and in vitro properties of SIVmac mutants with deletions in 'nonessential' genes. *AIDS Res. Hum. Retroviruses* **10,** 607–616 (1994).
165. Akari, H. *et al.* Biological characterization of human immunodeficiency virus type 1 and type 2 mutants in human peripheral blood mononuclear cells. *Arch. Virol.* **123,** 157–167 (1992).
166. Guyader, M., Emerman, M., Montagnier, L. & Peden, K. VPX mutants of HIV-2 are infectious in established cell lines but display a severe defect in peripheral blood lymphocytes. *EMBO J.* **8,** 1169–1175 (1989).

167. Jin, L., Zhou, Y. & Ratner, L. HIV type 2 Vpx interaction with Gag and incorporation into virus-like particles. *AIDS Res. Hum. Retroviruses* **17,** 105–111 (2001).
168. Singh, S. P. *et al.* Epitope-tagging approach to determine the stoichiometry of the structural and nonstructural proteins in the virus particles: amount of Vpr in relation to Gag in HIV-1. *Virology* **268,** 364–371 (2000).
169. Belshan, M. & Ratner, L. Identification of the nuclear localization signal of human immunodeficiency virus type 2 Vpx. *Virology* **311,** 7–15 (2003).
170. Rajendra Kumar, P., Singhal, P. K., Vinod, S. S. & Mahalingam, S. A non-canonical transferable signal mediates nuclear import of simian immunodeficiency virus Vpx protein. *J. Mol. Biol.* **331,** 1141–1156 (2003).
171. Rajendra Kumar, P., Singhal, P. K., Subba Rao, M. R. K. & Mahalingam, S. Phosphorylation by MAPK regulates simian immunodeficiency virus Vpx protein nuclear import and virus infectivity. *J. Biol. Chem.* **280,** 8553–8563 (2005).
172. Powell, R. D., Holland, P. J., Hollis, T. & Perrino, F. W. Aicardi-Goutieres syndrome gene and HIV-1 restriction factor SAMHD1 is a dGTP-regulated deoxynucleotide triphosphohydrolase. *J. Biol. Chem.* **286,** 43596–43600 (2011).
173. Li, N., Zhang, W. & Cao, X. Identification of human homologue of mouse IFN-gamma induced protein from human dendritic cells. *Immunol. Lett.* **74,** 221–224 (2000).
174. Amie, S. M., Bambara, R. A. & Kim, B. GTP is the primary activator of the anti-HIV restriction factor SAMHD1. *J. Biol. Chem.* **288,** 25001–25006 (2013).
175. Goldstone, D. C. *et al.* HIV-1 restriction factor SAMHD1 is a deoxynucleoside triphosphate triphosphohydrolase. *Nature* **480,** 379–382 (2011).
176. Kennedy, E. M. *et al.* Ribonucleoside triphosphates as substrate of human immunodeficiency virus type 1 reverse transcriptase in human macrophages. *J. Biol. Chem.* **285,** 39380–39391 (2010).
177. Lahouassa, H. *et al.* SAMHD1 restricts the replication of human immunodeficiency virus type 1 by depleting the intracellular pool of deoxynucleoside triphosphates. *Nat. Immunol.* **13,** 223–228 (2012).
178. St Gelais, C. *et al.* SAMHD1 restricts HIV-1 infection in dendritic cells (DCs) by dNTP depletion, but its expression in DCs and primary CD4+ T-lymphocytes cannot be upregulated by interferons. *Retrovirology* **9,** 105 (2012).
179. Gramberg, T. *et al.* Restriction of diverse retroviruses by SAMHD1. *Retrovirology* **10,** 26 (2013).
180. Hollenbaugh, J. A. *et al.* Host factor SAMHD1 restricts DNA viruses in non-dividing myeloid cells. *PLoS Pathog.* **9,** e1003481 (2013).
181. Kim, E. T., White, T. E., Brandariz-Núñez, A., Diaz-Griffero, F. & Weitzman, M. D. SAMHD1 restricts herpes simplex virus 1 in macrophages by limiting DNA replication. *J. Virol.* **87,** 12949–12956 (2013).
182. Zhao, K. *et al.* Modulation of LINE-1 and Alu/SVA retrotransposition by Aicardi-Goutières syndrome-related SAMHD1. *Cell Rep* **4,** 1108–1115 (2013).

183. Ahn, J. *et al.* HIV/simian immunodeficiency virus (SIV) accessory virulence factor Vpx loads the host cell restriction factor SAMHD1 onto the E3 ubiquitin ligase complex CRL4DCAF1. *J. Biol. Chem.* **287,** 12550–12558 (2012).
184. Brandariz-Nuñez, A. *et al.* Role of SAMHD1 nuclear localization in restriction of HIV-1 and SIVmac. *Retrovirology* **9,** 49 (2012).
185. Schwefel, D. *et al.* Structural basis of lentiviral subversion of a cellular protein degradation pathway. *Nature* **505,** 234–238 (2014).
186. St Gelais, C. & Wu, L. SAMHD1: a new insight into HIV-1 restriction in myeloid cells. *Retrovirology* **8,** 55 (2011).
187. Rice, G. I. *et al.* Mutations involved in Aicardi-Goutières syndrome implicate SAMHD1 as regulator of the innate immune response. *Nat. Genet.* **41,** 829–832 (2009).
188. Chahwan, C. & Chahwan, R. Aicardi-Goutieres syndrome: from patients to genes and beyond. *Clin. Genet.* **81,** 413–420 (2012).
189. Berger, A. *et al.* SAMHD1-deficient CD14+ cells from individuals with Aicardi-Goutières syndrome are highly susceptible to HIV-1 infection. *PLoS Pathog.* **7,** e1002425 (2011).
190. Pan, X., Baldauf, H.-M., Keppler, O. T. & Fackler, O. T. Restrictions to HIV-1 replication in resting CD4+ T lymphocytes. *Cell Res.* **23,** 876–885 (2013).
191. Baldauf, H.-M. *et al.* SAMHD1 restricts HIV-1 infection in resting CD4(+) T cells. *Nat. Med.* **18,** 1682–1687 (2012).
192. Cribier, A., Descours, B., Valadão, A. L. C., Laguette, N. & Benkirane, M. Phosphorylation of SAMHD1 by cyclin A2/CDK1 regulates its restriction activity toward HIV-1. *Cell Rep* **3,** 1036–1043 (2013).
193. White, T. E. *et al.* The retroviral restriction ability of SAMHD1, but not its deoxynucleotide triphosphohydrolase activity, is regulated by phosphorylation. *Cell Host Microbe* **13,** 441–451 (2013).
194. Welbourn, S., Dutta, S. M., Semmes, O. J. & Strebel, K. Restriction of virus infection but not catalytic dNTPase activity is regulated by phosphorylation of SAMHD1. *J. Virol.* **87,** 11516–11524 (2013).
195. Diamond, T. L. *et al.* Macrophage tropism of HIV-1 depends on efficient cellular dNTP utilization by reverse transcriptase. *J. Biol. Chem.* **279,** 51545–51553 (2004).
196. Ryoo, J. *et al.* The ribonuclease activity of SAMHD1 is required for HIV-1 restriction. *Nat. Med.* **20,** 936–941 (2014).
197. Yang, Z. & Greene, W. C. A new activity for SAMHD1 in HIV restriction. *Nat. Med.* **20,** 808–809 (2014).
198. Berger, G. *et al.* APOBEC3A is a specific inhibitor of the early phases of HIV-1 infection in myeloid cells. *PLoS Pathog.* **7,** e1002221 (2011).
199. Manel, N. *et al.* A cryptic sensor for HIV-1 activates antiviral innate immunity in dendritic cells. *Nature* **467,** 214–217 (2010).
200. Lahaye, X. *et al.* The capsids of HIV-1 and HIV-2 determine immune detection of the viral cDNA by the innate sensor cGAS in dendritic cells. *Immunity* **39,** 1132–1142 (2013).

201. Weiss, K. K. *et al.* A role for dNTP binding of human immunodeficiency virus type 1 reverse transcriptase in viral mutagenesis. *Biochemistry* **43,** 4490–4500 (2004).
202. Lim, E. S. *et al.* The ability of primate lentiviruses to degrade the monocyte restriction factor SAMHD1 preceded the birth of the viral accessory protein Vpx. *Cell Host Microbe* **11,** 194–204 (2012).
203. Laguette, N. *et al.* Evolutionary and functional analyses of the interaction between the myeloid restriction factor SAMHD1 and the lentiviral Vpx protein. *Cell Host Microbe* **11,** 205–217 (2012).
204. Zhang, C., de Silva, S., Wang, J.-H. & Wu, L. Co-evolution of primate SAMHD1 and lentivirus Vpx leads to the loss of the vpx gene in HIV-1 ancestor. *PLoS ONE* **7,** e37477 (2012).
205. Allan, J. S. *et al.* Species-specific diversity among simian immunodeficiency viruses from African green monkeys. *J. Virol.* **65,** 2816–2828 (1991).
206. Wertheim, J. O. & Worobey, M. A challenge to the ancient origin of SIVagm based on African green monkey mitochondrial genomes. *PLoS Pathog.* **3,** e95 (2007).
207. Spragg, C. J. & Emerman, M. Antagonism of SAMHD1 is actively maintained in natural infections of simian immunodeficiency virus. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 21136–21141 (2013).
208. Etienne, L., Hahn, B. H., Sharp, P. M., Matsen, F. A. & Emerman, M. Gene loss and adaptation to hominids underlie the ancient origin of HIV-1. *Cell Host Microbe* **14,** 85–92 (2013).
209. Bishop, K. N., Verma, M., Kim, E.-Y., Wolinsky, S. M. & Malim, M. H. APOBEC3G inhibits elongation of HIV-1 reverse transcripts. *PLoS Pathog.* **4,** e1000231 (2008).
210. Compton, A. A. & Emerman, M. Convergence and divergence in the evolution of the APOBEC3G-Vif interaction reveal ancient origins of simian immunodeficiency viruses. *PLoS Pathog.* **9,** e1003135 (2013).
211. Keese, P. K. & Gibbs, A. Origins of genes: 'big bang' or continuous creation? *Proc. Natl. Acad. Sci. U.S.A.* **89,** 9489–9493 (1992).
212. Gaur, R. & Strebel, K. Insights into the dual activity of SIVmac239 Vif against human and African green monkey APOBEC3G. *PLoS ONE* **7,** e48850 (2012).
213. Gramberg, T., Sunseri, N. & Landau, N. R. Evidence for an activation domain at the amino terminus of simian immunodeficiency virus Vpx. *J. Virol.* **84,** 1387–1396 (2010).
214. Goujon, C. *et al.* Characterization of simian immunodeficiency virus SIVSM/human immunodeficiency virus type 2 Vpx function in human myeloid cells. *J. Virol.* **82,** 12335–12345 (2008).
215. Berger, G. *et al.* Functional analysis of the relationship between Vpx and the restriction factor SAMHD1. *J. Biol. Chem.* **287,** 41210–41217 (2012).
216. Wei, W. *et al.* A novel DCAF1-binding motif required for Vpx-mediated degradation of nuclear SAMHD1 and Vpr-induced G2 arrest. *Cell. Microbiol.* **14,** 1745–1756 (2012).

217. Belshan, M. *et al.* Vpx is critical for SIVmne infection of pigtail macaques. *Retrovirology* **9,** 32 (2012).
218. Srivastava, S. *et al.* Lentiviral Vpx accessory factor targets VprBP/DCAF1 substrate adaptor for cullin 4 E3 ubiquitin ligase to enable macrophage infection. *PLoS Pathog.* **4,** e1000059 (2008).
219. Bergamaschi, A. *et al.* The human immunodeficiency virus type 2 Vpx protein usurps the CUL4A-DDB1 DCAF1 ubiquitin ligase to overcome a postentry block in macrophage infection. *J. Virol.* **83,** 4854–4860 (2009).
220. Pertel, T., Reinhard, C. & Luban, J. Vpx rescues HIV-1 transduction of dendritic cells from the antiviral state established by type 1 interferon. *Retrovirology* **8,** 49 (2011).
221. Miyake, A. *et al.* Poly-proline motif in HIV-2 Vpx is critical for its efficient translation. *J. Gen. Virol.* **95,** 179–189 (2014).
222. Miyake, A., Miyazaki, Y., Fujita, M., Nomaguchi, M. & Adachi, A. Role of poly-proline motif in HIV-2 Vpx expression. *Front Microbiol* **5,** 24 (2014).
223. Yu, H. *et al.* The efficiency of Vpx-mediated SAMHD1 antagonism does not correlate with the potency of viral control in HIV-2-infected individuals. *Retrovirology* **10,** 27 (2013).
224. Onyango, C. O. *et al.* HIV-2 capsids distinguish high and low virus load patients in a West African community cohort. *Vaccine* **28 Suppl 2,** B60–67 (2010).
225. Gaschen, B. *et al.* Diversity considerations in HIV-1 vaccine selection. *Science* **296,** 2354–2360 (2002).
226. Barouch, D. H. Challenges in the development of an HIV-1 vaccine. *Nature* **455,** 613–619 (2008).
227. Lemey, P. *et al.* Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. *PLoS Comput. Biol.* **3,** e29 (2007).
228. Keele, B. F. *et al.* Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 7552–7557 (2008).
229. Hightower, G. K. *et al.* Higher HIV-1 Genetic Diversity is Associated with AIDS and Neuropsychological Impairment. *Virology* **433,** 498–505 (2012).
230. Barroso, H. *et al.* Evolutionary and structural features of the C2, V3 and C3 envelope regions underlying the differences in HIV-1 and HIV-2 biology and infection. *PLoS ONE* **6,** e14548 (2011).
231. Barroso, H. & Taveira, N. Evidence for negative selective pressure in HIV-2 evolution in vivo. *Infect. Genet. Evol.* **5,** 239–246 (2005).
232. Soares, R. S. *et al.* Cell-associated viral burden provides evidence of ongoing viral replication in aviremic HIV-2-infected patients. *J. Virol.* **85,** 2429–2438 (2011).
233. Charpentier, C. *et al.* Genotypic resistance profiles of HIV-2-treated patients in West Africa. *AIDS* **28,** 1161–1169 (2014).
234. Domingo, E. *et al.* Basic concepts in RNA virus evolution. *FASEB J.* **10,** 859–864 (1996).

235. Shiina, T., Hosomichi, K., Inoko, H. & Kulski, J. K. The HLA genomic loci map: expression, interaction, diversity and disease. *J. Hum. Genet.* **54,** 15–39 (2009).
236. Bashirova, A. A., Thomas, R. & Carrington, M. HLA/KIR Restraint of HIV: Surviving the Fittest. *Annual Review of Immunology* **29,** 295–317 (2011).
237. Carrington, M. & O'Brien, S. J. The Influence of HLA Genotype on AIDS*. *Annual Review of Medicine* **54,** 535–551 (2003).
238. Kawashima, Y. *et al.* Adaptation of HIV-1 to human leukocyte antigen class I. *Nature* **458,** 641–645 (2009).
239. McMichael, A. J., Borrow, P., Tomaras, G. D., Goonetilleke, N. & Haynes, B. F. The immune response during acute HIV-1 infection: clues for vaccine development. *Nat Rev Immunol* **10,** 11–23 (2010).
240. Edwards, B. H. *et al.* Magnitude of functional CD8+ T-cell responses to the gag protein of human immunodeficiency virus type 1 correlates inversely with viral load in plasma. *J. Virol.* **76,** 2298–2305 (2002).
241. Sacha, J. B. *et al.* Gag-specific CD8+ T lymphocytes recognize infected cells before AIDS-virus integration and viral protein expression. *J. Immunol.* **178,** 2746–2754 (2007).
242. Gao, X. *et al.* Effect of a single amino acid change in MHC class I molecules on the rate of progression to AIDS. *N. Engl. J. Med.* **344,** 1668–1675 (2001).
243. Diouf, K. *et al.* Associations between MHC class I and susceptibility to HIV-2 disease progression. *J. Hum. Virol.* **5,** 1–7 (2002).
244. Yindom, L.-M. *et al.* Influence of HLA class I and HLA-KIR compound genotypes on HIV-2 infection and markers of disease progression in a Manjako community in West Africa. *J. Virol.* **84,** 8202–8208 (2010).
245. Vivier, E., Tomasello, E., Baratin, M., Walzer, T. & Ugolini, S. Functions of natural killer cells. *Nat. Immunol.* **9,** 503–510 (2008).
246. Wilson, M. J. *et al.* Plasticity in the organization and sequences of human KIR/ILT gene families. *Proc. Natl. Acad. Sci. U.S.A.* **97,** 4778–4783 (2000).
247. Middleton, D. & Gonzelez, F. The extensive polymorphism of KIR genes. *Immunology* **129,** 8–19 (2010).
248. Valiante, N. M. *et al.* Functionally and structurally distinct NK cell receptor repertoires in the peripheral blood of two human donors. *Immunity* **7,** 739–751 (1997).
249. Gardiner, C. M. *et al.* Different NK cell surface phenotypes defined by the DX9 antibody are due to KIR3DL1 gene polymorphism. *J. Immunol.* **166,** 2992–3001 (2001).
250. Carrington, M., Martin, M. P. & van Bergen, J. KIR-HLA intercourse in HIV disease. *Trends Microbiol.* **16,** 620–627 (2008).
251. Alter, G. *et al.* HLA class I subtype-dependent expansion of KIR3DS1+ and KIR3DL1+ NK cells during acute human immunodeficiency virus type 1 infection. *J. Virol.* **83,** 6798–6805 (2009).
252. Vivian, J. P. *et al.* Killer cell immunoglobulin-like receptor 3DL1-mediated recognition of human leukocyte antigen B. *Nature* **479,** 401–405 (2011).

253. Carr, W. H., Pando, M. J. & Parham, P. KIR3DL1 polymorphisms that affect NK cell inhibition by HLA-Bw4 ligand. *J. Immunol.* **175,** 5222–5229 (2005).
254. Harris, R. S., Hultquist, J. F. & Evans, D. T. The Restriction Factors of Human Immunodeficiency Virus. *J. Biol. Chem.* **287,** 40875–40883 (2012).
255. Compton, A. A., Malik, H. S. & Emerman, M. Host gene evolution traces the evolutionary history of ancient primate lentiviruses. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **368,** 20120496 (2013).
256. Harris, R. S. *et al.* DNA deamination mediates innate immunity to retroviral infection. *Cell* **113,** 803–809 (2003).
257. Mangeat, B. *et al.* Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. *Nature* **424,** 99–103 (2003).
258. Lecossier, D., Bouchonnet, F., Clavel, F. & Hance, A. J. Hypermutation of HIV-1 DNA in the absence of the Vif protein. *Science* **300,** 1112 (2003).
259. Letko, M. *et al.* Vif proteins from diverse primate lentiviral lineages use the same binding site in APOBEC3G. *J. Virol.* **87,** 11861–11871 (2013).
260. Yu, X. *et al.* Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex. *Science* **302,** 1056–1060 (2003).
261. He, Z., Zhang, W., Chen, G., Xu, R. & Yu, X.-F. Characterization of conserved motifs in HIV-1 Vif required for APOBEC3G and APOBEC3F interaction. *J. Mol. Biol.* **381,** 1000–1011 (2008).
262. Kupzig, S. *et al.* Bst-2/HM1.24 is a raft-associated apical membrane protein with an unusual topology. *Traffic* **4,** 694–709 (2003).
263. Perez-Caballero, D. *et al.* Tetherin inhibits HIV-1 release by directly tethering virions to cells. *Cell* **139,** 499–511 (2009).
264. Neil, S. J. D., Zang, T. & Bieniasz, P. D. Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. *Nature* **451,** 425–430 (2008).
265. Lau, D., Kwan, W. & Guatelli, J. Role of the endocytic pathway in the counteraction of BST-2 by human lentiviral pathogens. *J. Virol.* **85,** 9834–9846 (2011).
266. Le Tortorec, A. & Neil, S. J. D. Antagonism to and intracellular sequestration of human tetherin by the human immunodeficiency virus type 2 envelope glycoprotein. *J. Virol.* **83,** 11966–11978 (2009).
267. Gupta, R. K. *et al.* Simian immunodeficiency virus envelope glycoprotein counteracts tetherin/BST-2/CD317 by intracellular sequestration. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 20889–20894 (2009).
268. Zhang, F. *et al.* Nef proteins from simian immunodeficiency viruses are tetherin antagonists. *Cell Host Microbe* **6,** 54–67 (2009).
269. Zhang, F. *et al.* SIV Nef proteins recruit the AP-2 complex to antagonize Tetherin and facilitate virion release. *PLoS Pathog.* **7,** e1002039 (2011).
270. Gupta, R. K. *et al.* Mutation of a single residue renders human tetherin resistant to HIV-1 Vpu-mediated depletion. *PLoS Pathog.* **5,** e1000443 (2009).
271. Sauter, D. *et al.* Tetherin-driven adaptation of Vpu and Nef function and the evolution of pandemic and nonpandemic HIV-1 strains. *Cell Host Microbe* **6,** 409–421 (2009).

272. Keckesova, Z., Ylinen, L. M. J. & Towers, G. J. The human and African green monkey TRIM5alpha genes encode Ref1 and Lv1 retroviral restriction factor activities. *Proc. Natl. Acad. Sci. U.S.A.* **101,** 10780–10785 (2004).
273. Towers, G. J. The control of viral infection by tripartite motif proteins and cyclophilin A. *Retrovirology* **4,** 40 (2007).
274. Yap, M. W., Nisole, S. & Stoye, J. P. A single amino acid change in the SPRY domain of human Trim5alpha leads to HIV-1 restriction. *Curr. Biol.* **15,** 73–78 (2005).
275. Takeuchi, J. S. *et al.* High level of susceptibility to human TRIM5$\alpha$ conferred by HIV-2 capsid sequences. *Retrovirology* **10,** 50 (2013).
276. Holley, R. W. *et al.* Structure of a Ribonucleic Acid. *Science* **147,** 1462–1465 (1965).
277. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* **94,** 441–448 (1975).
278. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **74,** 5463–5467 (1977).
279. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
280. Robertson, D. L., Hahn, B. H. & Sharp, P. M. Recombination in AIDS viruses. *J. Mol. Evol.* **40,** 249–259 (1995).
281. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).
282. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
283. Gall, A. *et al.* Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. *J. Clin. Microbiol.* **50,** 3838–3844 (2012).
284. Malboeuf, C. M. *et al.* Complete viral RNA genome sequencing of ultra-low copy samples by sequence-independent amplification. *Nucleic Acids Res.* **41,** e13 (2013).
285. Batty, E. M. *et al.* A Modified RNA-Seq Approach for Whole Genome Sequencing of RNA Viruses from Faecal and Blood Samples. *PLoS ONE* **8,** e66129 (2013).
286. Laszlo, A. H. *et al.* Decoding long nanopore sequencing reads of natural DNA. *Nat Biotech* **32,** 829–833 (2014).
287. Ledford, H. HIV rebound dashes hope of cure. *Nature* (2014). doi:10.1038/nature.2014.15535
288. Purcell, D. F., Elliott, J. H., Ross, A.-L. & Frater, J. Towards an HIV cure: science and debate from the International AIDS Society 2013 symposium. *Retrovirology* **10,** 134 (2013).
289. Buckner, M. 22. Village women as town prostitutes: cultural factors relevant to prostitution and HIV epidemiology in Guinea-Bissau. at <https://www.codesria.org/IMG/pdf/22LBUCKNER_.pdf>
290. Wilkins, A. *et al.* The epidemiology of HIV infection in a rural area of Guinea-Bissau. *AIDS* **7,** 1119–1122 (1993).

291. Cooper, M. *et al.* Cause of death among people with retroviral infection in rural Guinea-Bissau. *Trop Doct* **40,** 181–183 (2010).
292. Edwards, J. K. *et al.* Loss to clinic and five-year mortality among HIV-infected antiretroviral therapy initiators. *PLoS ONE* **9,** e102305 (2014).
293. Graham, S. M. *et al.* Loss to follow-up as a competing risk in an observational study of HIV-1 incidence. *PLoS ONE* **8,** e59480 (2013).
294. R Core Team. R: A Language and Environment for Statistical Computing. (2013). at <http://www.R-project.org/>
295. Koressaar, T. & Remm, M. Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23,** 1289–1291 (2007).
296. Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28,** 1647–1649 (2012).
297. Geospiza Inc. FinchTV 1.4.0. at <http://www.geospiza.com>
298. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5,** 113 (2004).
299. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Meth* **9,** 772–772 (2012).
300. Matsubara, H., Jukes, T. H. & Cantor, C. R. Structural and evolutionary relationships of ferredoxins. *Brookhaven Symp. Biol.* **21,** 201–216 (1968).
301. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17,** 368–376 (1981).
302. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22,** 160–174 (1985).
303. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16,** 111–120 (1980).
304. Turner, F. S. Assessment of insert sizes and adapter content in fastq data from NexteraXT libraries. *Front Genet* **5,** 5 (2014).
305. Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data. *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/* at <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
306. Dear, S. & Staden, R. A standard file format for data from DNA sequencing instruments. *Mitochondrial DNA* **3,** 107–110 (1992).
307. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36,** e105 (2008).
308. JN. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. at <https://github.com/najoshi/sickle>
309. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9,** 357–359 (2012).
310. Picard. at <http://picard.sourceforge.net.>
311. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

312. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14,** 178–192 (2013).
313. Dobrovolny, P. L. & Bess, D. Optimized PCR-based detection of mycoplasma. *J Vis Exp* **52,** 3057 (2011).
314. Edo-Matas, D. *et al.* Genetic composition of replication competent clonal HIV-1 variants isolated from peripheral blood mononuclear cells (PBMC), HIV-1 proviral DNA from PBMC and HIV-1 RNA in serum in the course of HIV-1 infection. *Virology* **405,** 492–504 (2010).
315. Banks, L., Gholamin, S., White, E., Zijenah, L. & Katzenstein, D. A. Comparing Peripheral Blood Mononuclear Cell DNA and Circulating Plasma viral RNA pol Genotypes of Subtype C HIV-1. *J AIDS Clin Res* **3,** 141–147 (2012).
316. Butler, D. M., Pacold, M. E., Jordan, P. S., Richman, D. D. & Smith, D. M. The Efficiency of Single Genome Amplification and Sequencing is Improved by Quantitation and Use of a Bioinformatics Tool. *J Virol Methods* **162,** 280–283 (2009).
317. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *PNAS* **76,** 5269–5273 (1979).
318. Schweitzer, F. *et al.* Longitudinal comparison of HIV-1 plasma viral load and cellular proviral load. *J Int AIDS Soc* **17,** 19669 (2014).
319. Luzuriaga, K. *et al.* HIV type 1 (HIV-1) proviral reservoirs decay continuously under sustained virologic control in HIV-1-infected children who received early treatment. *J. Infect. Dis.* **210,** 1529–1538 (2014).
320. Buzon, M. J. *et al.* HIV-1 persistence in CD4+ T cells with stem cell-like properties. *Nat. Med.* **20,** 139–142 (2014).
321. Williams, J. P. *et al.* HIV-1 DNA predicts disease progression and post-treatment virological control. *Elife* **3,** e03821 (2014).
322. Damond, F. *et al.* Quantification of proviral load of human immunodeficiency virus type 2 subtypes A and B using real-time PCR. *J. Clin. Microbiol.* **39,** 4264–4268 (2001).
323. Damond, F. *et al.* Plasma RNA Viral Load in Human Immunodeficiency Virus Type 2 Subtype A and Subtype B Infections. *J Clin Microbiol* **40,** 3654–3659 (2002).
324. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Mol Biol Evol* **26,** 1641–1650 (2009).
325. Shimodaira, H. & Hasegawa, M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution* **16,** 1114–1116 (1999).
326. De Silva, T. I. *et al.* Population dynamics of HIV-2 in rural West Africa: comparison with HIV-1 and ongoing transmission at the heart of the epidemic. *AIDS* **27,** 125–134 (2013).
327. De Silva, T. I., Cotten, M. & Rowland-Jones, S. L. HIV-2: the forgotten AIDS virus. *Trends Microbiol.* **16,** 588–595 (2008).

328. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
329. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 (2007).
330. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29,** 1969–1973 (2012).
331. Rambaut, A., Suchard, M. A., Xie, D. & Drummond, A. J. Tracer v1.6. (2014). at <http://beast.bio.ed.ac.uk/tracer>
332. FigTree. at <http://beast.bio.ed.ac.uk/figtree>
333. Slatkin, M. & Hudson, R. R. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129,** 555–562 (1991).
334. Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol. Biol. Evol.* **24,** 1792–1800 (2007).
335. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *PNAS* **89,** 10915–10919 (1992).
336. Toni, T. & Stumpf, M. P. H. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics* **26,** 104–110 (2010).
337. Lemey, P., Rambaut, A. & Pybus, O. G. HIV evolutionary dynamics within and among hosts. *AIDS Rev* **8,** 125–140 (2006).
338. Lemey, P., Minin, V. N., Bielejec, F., Kosakovsky Pond, S. L. & Suchard, M. A. A counting renaissance: combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. *Bioinformatics* **28,** 3248–3256 (2012).
339. Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25,** 1605–1612 (2004).
340. Henn, M. R. *et al.* Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants Upon Immune Recognition During Acute Infection. *PLoS Pathog* **8,** e1002529 (2012).
341. Smith, E. N. *et al.* Biased estimates of clonal evolution and subclonal heterogeneity can arise from PCR duplicates in deep sequencing experiments. *Genome Biol.* **15,** 420 (2014).
342. Edwards, C. T. T. *et al.* Evolution of the Human Immunodeficiency Virus Envelope Gene Is Dominated by Purifying Selection. *Genetics* **174,** 1441–1453 (2006).
343. Moutsianas, L. & Morris, A. P. Methodology for the analysis of rare genetic variation in genome-wide association and re-sequencing studies of complex human traits. *Brief Funct Genomics* **13,** 362–370 (2014).
344. McElroy, K., Thomas, T. & Luciani, F. Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. *Microb Inform Exp* **4,** 1 (2014).
345. Pan, W. *et al.* DNA polymerase preference determines PCR priming efficiency. *BMC Biotechnol.* **14,** 10 (2014).

346. Brodin, J. *et al.* PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data. *PLoS ONE* **8,** e70388 (2013).
347. Pinto, A. J. & Raskin, L. PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLoS ONE* **7,** e43093 (2012).
348. Jabara, C. B., Jones, C. D., Roach, J., Anderson, J. A. & Swanstrom, R. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 20166–20171 (2011).
349. McInerney, P., Adams, P. & Hadi, M. Z. Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. *Mol Biol Int* **2014,** 287430 (2014).
350. Zagordi, O., Klein, R., Däumer, M. & Beerenwinkel, N. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res.* **38,** 7400–7409 (2010).
351. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth* **5,** 621–628 (2008).
352. Qi, Y.-X., Liu, Y.-B. & Rong, W.-H. [RNA-Seq and its applications: a new technology for transcriptomics]. *Yi Chuan* **33,** 1191–1202 (2011).
353. Yang, X. *et al.* De novo assembly of highly diverse viral populations. *BMC Genomics* **13,** 1–13 (2012).
354. Borozan, I., Watt, S. N. & Ferretti, V. Evaluation of Alignment Algorithms for Discovery and Identification of Pathogens Using RNA-Seq. *PLoS ONE* **8,** e76935 (2013).
355. Chen, Y.-C., Liu, T., Yu, C.-H., Chiang, T.-Y. & Hwang, C.-C. Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. *PLoS ONE* **8,** e62856 (2013).
356. Hansen, K. D., Brenner, S. E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* **38,** e131 (2010).
357. Van Dijk, E. L., Jaszczyszyn, Y. & Thermes, C. Library preparation methods for next-generation sequencing: tone down the bias. *Exp. Cell Res.* **322,** 12–20 (2014).
358. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
359. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21,** 1859–1875 (2005).
360. Novocraft.com: Novoalign short read mapper (http://www.novocraft.com/main/downloadpage.php). at <http://www.novocraft.com/main/downloadpage.php>
361. Quail, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13,** 341 (2012).
362. Cohen, N., Dagan, T., Stone, L. & Graur, D. GC composition of the human genome: in search of isochores. *Mol. Biol. Evol.* **22,** 1260–1272 (2005).

363. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* **12,** 87–98 (2011).
364. Glenn, T. C. Field guide to next-generation DNA sequencers. *Mol Ecol Resour* **11,** 759–769 (2011).
365. Display Sequence. at <http://www.hiv.lanl.gov/components/sequence/HIV/asearch/query_one.comp?se_id=BD413542>
366. Landi, A., Iannucci, V., Nuffel, A. V., Meuwissen, P. & Verhasselt, B. One protein to rule them all: modulation of cell surface receptors and molecules by HIV Nef. *Curr. HIV Res.* **9,** 496–504 (2011).
367. Sebastiani, G., Gkouvatsos, K. & Pantopoulos, K. Chronic hepatitis C and liver fibrosis. *World J. Gastroenterol.* **20,** 11033–11053 (2014).
368. Ninomiya, M. *et al.* Use of illumina deep sequencing technology to differentiate hepatitis C virus variants. *J. Clin. Microbiol.* **50,** 857–866 (2012).
369. Goodgame, R. Norovirus gastroenteritis. *Curr Gastroenterol Rep* **8,** 401–408 (2006).
370. Van Alphen, L. B. *et al.* The application of new molecular methods in the investigation of a waterborne outbreak of norovirus in denmark, 2012. *PLoS ONE* **9,** e105053 (2014).
371. Faria, N. R. *et al.* HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science* **346,** 56–61 (2014).
372. Lauck, M. *et al.* Discovery and full genome characterization of two highly divergent simian immunodeficiency viruses infecting black-and-white colobus monkeys (Colobus guereza) in Kibale National Park, Uganda. *Retrovirology* **10,** 107 (2013).
373. Kurn, N. *et al.* Novel isothermal, linear nucleic acid amplification systems for highly multiplexed applications. *Clin. Chem.* **51,** 1973–1981 (2005).
374. Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345** 1369-1372 (2014).
375. Todd, J. *et al.* Performance characteristics for the quantitation of plasma HIV-1 RNA using branched DNA signal amplification technology. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.* **10 Suppl 2,** S35–44 (1995).
376. Baker, M. De novo genome assembly: what every biologist should know. *Nat Meth* **9,** 333–337 (2012).
377. Li, Z. *et al.* Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Brief Funct Genomics* **11,** 25–37 (2012).
378. Miller, J. R., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* **95,** 315–327 (2010).
379. Schliesky, S., Gowik, U., Weber, A. P. M. & Bräutigam, A. RNA-Seq Assembly - Are We There Yet? *Front Plant Sci* **3,** 220 (2012).
380. Knipe, D. M. & Howley, P. M. *Fields' Virology*. (Lippincott Williams & Wilkins, 2007).

381. Gorski, J. L., Gonzalez, I. L. & Schmickel, R. D. The secondary structure of human 28S rRNA: the structure and evolution of a mosaic rRNA gene. *J. Mol. Evol.* **24,** 236–251 (1987).
382. Rao, M. *et al.* HIV-1 variable loop 2 and its importance in HIV-1 infection and vaccine development. *Curr. HIV Res.* **11,** 427–438 (2013).
383. Brandin, E. *et al.* pol Gene Sequence Variation in Swedish HIV-2 Patients Failing Antiretroviral Therapy. *AIDS Research and Human Retroviruses* **19,** 543–550 (2003).
384. Lipscomb, J. T. *et al.* HIV Reverse-Transcriptase Drug Resistance Mutations During Early Infection Reveal Greater Transmission Diversity Than in Envelope Sequences. *J. Infect. Dis.* **210,** 1827-1837 (2014).
385. Feldhahn, M. *et al.* No evidence of viral genomes in whole-transcriptome sequencing of three melanoma metastases. *Exp. Dermatol.* **20,** 766–768 (2011).
386. So-Armah, K. A. *et al.* HIV infection, antiretroviral therapy initiation and longitudinal changes in biomarkers of organ function. *Curr. HIV Res.* **12,** 50–59 (2014).