



St Cross College
and
Department of Zoology
Human Evolutionary Genetics
A Thesis Submitted for the Degree of
Doctor of Philosophy
at the University of Oxford

**Integrations and Introgressions:
A Viral and Genetic History of the Human Genome**

Cindy Gilda Santander
Hilary Term 2019

Cindy Gilda Santander

Integrations and Introgressions:

A Viral and Genetic History of the Human Genome

A Thesis Submitted for the Degree of

Doctor of Philosophy

at the University of Oxford, Hilary Term 2019

Supervisor: Dr. Cristian Capelli

University of Oxford

Human Evolutionary Genetics

Department of Zoology

*And within the logic of our humanity,
we believe the lie and no one can tolerate the truth.*
— René Pérez Joglar

Acknowledgements

This thesis has been a labour of many things and it would not have been possible without the help and support from others. I would like to first thank my supervisor Dr. Cristian Capelli for all of his help, patience, and support during the DPhil. Cristian, thank you for adopting me into your group and for allowing me to complete this doctorate. I want to thank you for your raw enthusiasm in science and how it has restored my faith in research. There are no words to thank you enough for supporting me and supervising me these last few years.

I would also like to give special thanks to Dr. Francesco Montinaro. Fra, it has been a privilege to learn and work with you on this project which I know was so close to your heart. Thank you so much for your patience, for teaching me and guiding me even if it was brief. My only regret is not having had the opportunity to work with you more—it's thanks to you that this project was able to take shape. Thank you for entrusting me with it. I can only hope I've carried it out with the rigour and care it deserves.

Progress is never linear and setbacks occur in research as they do in life. I would like to thank the Department of Zoology for providing me a DPhil experience that has taught me some unexpected and harsh truths which, for better or for worse, I will always take with me.

To Tanushka, thank you for making our time together a sweet respite amongst the bitterness. Despite everything, the best thing to have come from that was our friendship.

To Andrés, Ryan, Gerardo, Joana, Nacho and Pol: You are the much needed sunshine in Oxford. Thank you for having made this process bearable, for lending an ear and sharing a pint when needed, for being the light in the dark.

To José, thanks for every peptalk and piece of advice that grounded me. In this short time, it is your genuine rawness and honesty that I've valued in our friendship. Your work ethic, moral compass, and love for biology is an inspiration. Thank you for being you and for always knowing what to say.

Momo, it's been a long time since the days we would chat outside our houses in Woodhaven. Thank you for staying true to yourself, for never giving up on that silliness which you, Ally and I have maintained for our own sanity. Thank you for always caring.

Ally, from our PS 60 days to the present we've seen each other grow. Thank you for believing in me when no one did and for reminding me that I could do this. Thanks for being a true BFF through and through.

Pippo, without you this DPhil would not have been possible. Thank you for supporting me and helping me grow. From the Death Star to the Centre of the Universe, thank you for everything.

Tefa, gran amiga—aunque no lo creas, es gracias a tí y tú ayuda con mi postulación que logro esto hoy. Gracias por tú linda amistad apesar de la distancia y el tiempo.

Rustam, it was never geriatric charity—I hope you know that our discussions have not only made me feel stronger but even optimistic about my future in research.

Jimena, amiga, clon! No existen palabras para expresar mi agradecimiento y admiración por tí. Gracias por recordarme que el mejor remedio para el dolor y la tristeza es la risa. Gracias por enseñarme a ver y valorar lo que yo no lograba ver en mi misma y por ser mi hermana cuando más lo necesité.

A mis padres, gracias por apoyarme en un sueño que, en algún momento, hasta tal vez pensaron que era un delirio. Este sueño es realidad por algo que ustedes me enseñaron: que preguntar no cuesta nada y lo peor que te pueden decir es no. Papá, apesar de todos nuestros roces, somos bastantes parecidos. Quiero agradecerte por enseñarme que la astucia no impede ser buena persona y que "está bien que me las vean pero no que me las pellizcan". La vida es más que estudios y en eso eres de los mejores profesores. Mamá, si amo la ciencia es gracias a tí. Mujer fuerte, humilde, luchadora y inteligente me haz dicho muchas veces que soy yo tú modelo a seguir pero la verdad es que tú fuiste el molde en cual me base.

Charlie, to think that I only understood 40% of what you said to me the first time we met. Thank you for putting up with me, for being a calm and comforting presence in my difficult moments even when you yourself were in pain. Out of the tragedy we've both emerged scarred, but we've also gained a best friend and companion. In the bigger scheme of things, what can be better than that? Thank you for teaching me to value those moments that I'll remember at the end of everything.

Declaration

This thesis comprises one published paper as a chapter and two chapters that contribute towards manuscripts that are in preparation. All work presented here is my own and where I have benefitted from other people's research, I have declared this in the text.

In accordance with the Regulations and with permission from the Director of Graduate Studies in the Department of Zoology at the University of Oxford, I present this thesis as a series of self-contained chapters in the style of scientific journal articles. Chapter 2 and parts of chapter 3 are part of a published article while parts of chapters 4 and 5 have been published as a Review. Additionally, a manuscript is also in preparation which will make use of results presented here in chapters 4 and 5. The publications based on the aforementioned data chapters can be found below:

Santander, C.G., et al. (2017). STEAK: A specific tool for transposable elements and retrovirus detection in high-throughput sequencing data, *Virus Evolution*, Volume 3, Issue 2, July 2017, **vex023**, <https://doi.org/10.1093/ve/vex023>

Santander, C.G., Montinaro, F., & Capelli, C. (2019). Searching for archaic contribution in Africa. *Annals of Human Biology*, May 2019, 1-39.

Oxford, Hilary Term 2019

Cindy Gilda Santander

Abstract

Integrations and Introgressions: a viral and genetic history of the human genome.

Cindy G. Santander

St. Cross College and the Department of Zoology, University of Oxford

The human genome combines archaic relics which have been obtained through two biological vehicles: viral infection and interbreeding. In this thesis I explore detection methods that search for the products of these biological processes, starting with some of the oldest fossils in our genome: viruses. I examine the most recently active endogenous retrovirus (ERV) family in humans: human endogenous retrovirus-K (HERV-K), human mouse mammary virus-like 2 (HML2), also known as HK2. I collaboratively developed and benchmarked the software STEAK for detecting mobile element and retroviral integrations in high-throughput sequencing data. Using STEAK on 70 whole genome sequences (WGS), I show that all polymorphic HK2 integrations found in Europeans are also present in African populations and estimate that HK2 activity stopped prior to the emergence of modern humans. The second half of this thesis investigates the outstanding question, did archaic and modern humans admix in Africa? Non-African populations today carry 1-5% of Neanderthal and/or Denisovan admixture. However, it remains unclear whether archaic introgression is present in African populations. I use WGS simulations to assess a current archaic reference-free method, S^* , under different demographic models varying in divergence dates, admixture dates, and admixture proportion. I show that in the absence of an archaic reference and an obscure demographic history, S^* results must be interpreted with caution given its sensitivity to demographic processes. Using S^* and Sprime on African WGS, I find that putatively introgressed regions detected by both methods are notably long and old haplotypes. I show with D -statistics that African populations such as the San share excess derived alleles with Denisovans but not with Neanderthals. These analyses collectively support a model where interbreeding took place between an ancestral human population and other archaic hominins within Africa. Overall, this thesis elucidates the relationship between the human genome and the archaic sequences it hosts.

List of Abbreviations

AMH	Anatomically modern human
ERV	Endogenous retrovirus
FDR	False discovery rate
HERV	Human endogenous retrovirus
HGDP	Human Genome Diversity Panel
HK2	Human endogenous retrovirus-K (HERV-K), human mouse mammary virus-like 2 (HML2)
HTS	High throughput sequencing
ILS	Incomplete lineage sorting
LD	Linkage disequilibrium
SGDP	The Simons Genome Diversity Project
TCGA	The Cancer Genome Atlas
TE	Transposable Elements
WGS	Whole Genome Sequencing

Contents

1	Introduction	1
1.1	Endogenous Retroviruses	1
1.1.1	Retroviruses	1
1.1.2	Biological relevance	3
1.1.3	HERV-K	5
1.1.4	Bioinformatic Challenges and Solutions	7
1.2	Human Origins	10
1.2.1	The emergence of humanity: <i>Homo</i> and anatomically modern humans	10
1.2.2	Modern Human Origins in Africa	17
1.3	Motivation and Problem Statement	20
1.4	Thesis Structure	20
2	STEAK: A specific tool for transposable elements and retrovirus detection in high-throughput sequencing data	22
2.1	Background	22
2.2	Materials and Methods	25
2.2.1	General benchmarking parameters	25
2.2.2	WGS from the Cancer Genome Atlas	25
2.2.3	WGS of CEU (Utah Residents with Northern and Western European Ancestry) pedigree from the 1000 Genomes Project	25
2.2.4	Simulated datasets and benchmarking	25
2.2.5	Target-enrichment dataset	26
2.2.6	Availability and requirements	26
2.2.7	Data availability	27
2.3	Results	27
2.3.1	STEAK algorithmic overview	27
2.3.2	Input data	28
2.3.3	Chimeric read detection and trimming	28
2.3.4	TE and retrovirus detection module	29
2.3.5	Software specifications and parallelisation	31
2.3.6	Overview of features and comparison to other TE and virus software	33
2.3.7	Evaluation of STEAK	33
2.3.8	Performance evaluation on simulated data	35

2.3.9	Performance on whole genome and target enrichment data . . .	36
2.4	Discussion	38
3	On the activity of HERV-K (HML-2)	44
3.1	Introduction	44
3.2	Materials and Methods	47
3.2.1	Estimating the HK2 integration rate for the last 5 MY	47
3.2.2	Estimating HK2 historical (30 MYA – 5 MYA) integration rates . .	47
3.2.3	Distribution for different polymorphic HK2 integration worldwide	47
3.2.4	WGS from The Cancer Genome Atlas	48
3.2.5	Marking presence/absence of polymorphic HK2 integrations . . .	48
3.3	Results	48
3.3.1	The integration rate of HK2 plunged in the last 5 million years . .	49
3.3.2	No recent HK2 activity in non-African germlines	54
3.4	Discussion	56
4	Reference-free detection of archaic introgression	58
4.1	Background	58
4.1.1	Methods to infer archaic introgression	60
4.2	Introduction	69
4.3	Materials and Methods	69
4.3.1	Coalescent whole-genome simulations	69
4.3.2	Performing S^* and Sprime	70
4.3.3	TMRCA estimates	70
4.4	Simulations	70
4.4.1	Admixture Proportion	71
4.4.2	Recombination	75
4.4.3	Demography	76
4.4.4	Divergence and Admixture Dates	80
4.4.5	TMRCA	81
4.5	Discussion	84
5	Searching for signals of archaic introgression in Africa	88
5.1	Background	88
5.1.1	The complicated demographic history of Africa	89
5.1.2	On the archaic side of things	92
5.2	Introduction	96
5.3	Materials and Methods	97
5.3.1	Whole genome sequencing data from SGDP	97
5.3.2	Performing S^* and Sprime	98
5.3.3	Papuans	98
5.3.4	African populations	98
5.3.5	TMRCA estimates	99

5.3.6	Comparison with Denisovan and Neanderthal genomes	99
5.4	Calibrating detection of archaic introgression	100
5.4.1	When Size Matters: Sample size in Papuans	100
5.5	Detection of putative archaic introgression in Africans	106
5.6	A Tale of Two Maps: HapMap YRI & African Enriched Recombination Maps	116
5.7	<i>D</i> -statistics: Archaic ancestry in Africa	120
5.8	Discussion	128
6	Conclusion	132
6.1	Future Work	136
6.1.1	The Good: Adaptive introgression and viral exaptation	137
6.1.2	The Bad: Deleterious and pathogenic burdens	137
6.1.3	The Ugly: Beyond SNPs and non-Africans	138
A	Appendices	140
A.1	Supplementary Material for Chapter 2	140
A.2	Appendix: Chapter 3	162
A.3	Appendix: Chapter 4	164
A.4	Appendix: Chapter 5	166
	Bibliography	170

Introduction

Hybridisation and viral infections have contributed a substantial portion of the human genome (Sankararaman *et al.*, 2016; Browning *et al.*, 2018; Belshaw *et al.*, 2004). In some circumstances, one has led to the other as seen in the case of modern humans and Neanderthal inbreeding (Enard & Petrov, 2018; Pimenoff *et al.*, 2017). This intimate double-edged relationship between the genus *Homo* and viral pathogens has shaped our evolutionary history as well as theirs (Enard & Petrov, 2018; Simmonds *et al.*, 2018). Both archaic admixture and ancient retroviral infections have left genetic and viral fossils, respectively, within our genomes. In this thesis, I explore forms of detecting these archaic remnants and what they may tell us about human evolutionary history. Here I review the "fossil record" left behind by ancient retroviral infections as well as the evidence that traces our modern human origins.

1.1 Endogenous Retroviruses

1.1.1 Retroviruses

The plethora of viruses that exist on earth vary extensively in their biology. Among the observed genetic diversity in viruses, retroviruses are unique as they represent the only known class of eukaryotic viruses that must integrate into the DNA of their host in order to effectively complete their life cycle (Kurth & Bannert, 2010).

The genetic material of retroviruses encodes for all the necessary enzymatic equipment to integrate and perpetuate into the genome of their host. Reverse transcriptase (RT) is the enzyme that makes it possible for their RNA genome to be reverse transcribed into a DNA copy which is essential for their successful chromosomal integration (Figure 1.2). This integrated provirus is composed of two long terminal repeats (LTRs), with a length of 300–1,200 bp, flanking each end of the coding region. Enclosed between the LTRs lies about 5–10 kb of sequence that typically encodes for three genes: *gag*, *pol*, and *env*. The *gag* gene encodes the Gag polyprotein in addition to all else that is required for the assembly of non-infectious and immature viral-particles. The transcription of *pol* produces reverse transcriptase which is crucial for proviral intergration into the host genome. Finally, the role of *env* is to encode the viral surface glycoprotein and

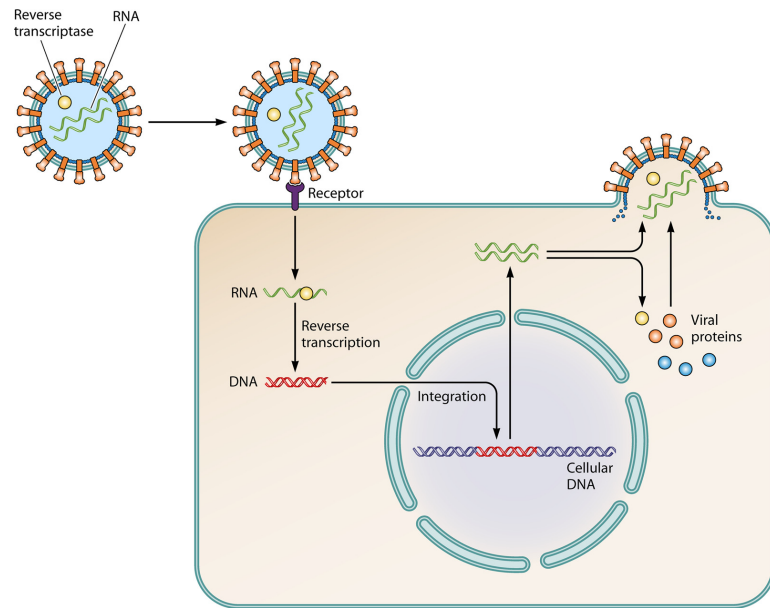


Fig. 1.1.: Retrovirus life cycle. Schematic from Denner & Tönjes (2012), *Clinical Microbiology Reviews*. The green single strands represent the RNA genomes that are converted by RT to the red DNA duplex, known as the provirus, that then integrates into the host cellular DNA. Through the normal course of host transcription, viral mRNA is produced along with host RNA. The synthesised viral proteins come together with the genomic RNA to form viral particles, which sequentially bud and cleave off.

transmembrane proteins which mediate cellular receptor binding and membrane fusion. The coding region of complex retroviruses can have additional accessory genes, allowing for evolutionary alternatives of utilising these accessory genes and their products in the optimisation of retroviral replication. The human immunodeficiency virus (HIV), for example, uses a system that employs the accessory protein Rev to increase replication efficiency. Rev essentially hijacks a transporter protein of the host that allows viral RNA transcripts, or partially spliced RNA, to be exported from the nucleus to the cytoplasm. By doing so the virus avoids the degradation of the unspliced RNA, therefore playing a key role in the replication efficiency of HIV (Graves & J, 1997).

The replication of a retrovirus depends on the permanent integration of the provirus into the host genome. Such events tend to typically occur in somatic cells. Nevertheless, there are rare occasions when retroviruses infect a germ cell. On the chance that the infected germ cell survives and the host's progeny develops from that infected germ cell, the provirus can be inherited in a Mendelian fashion (Stoye & Coffin, 1985), propagating onwards in a population while residing as an endogenous retrovirus (ERV) (Figure 1.2).

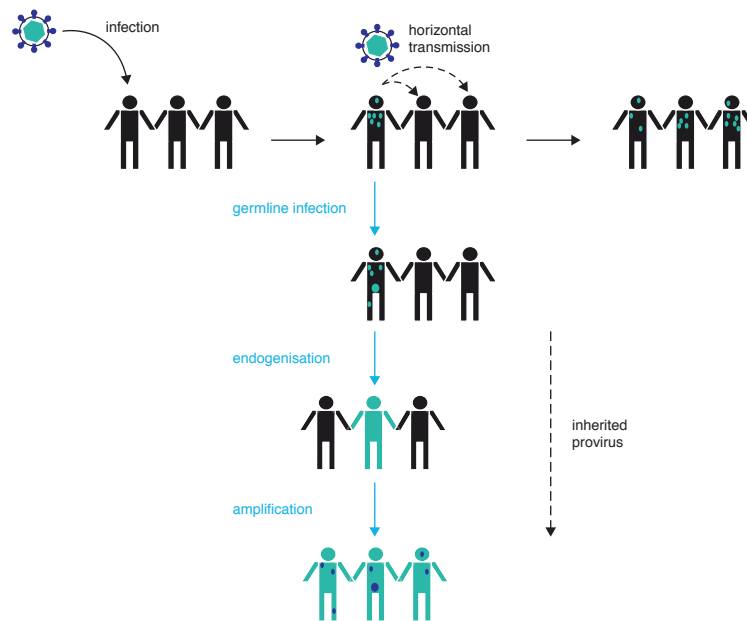


Fig. 1.2.: Mechanism of retrovirus endogenisation (Dewannieux & Heidmann, 2013). In the course of a normal infection, retroviruses can reach the germline of the host where they integrate their genome as a provirus. The latter can then be passed down through Mendelian inheritance to the offspring, and also amplify with the genome, leading to the creation of a new family of ERVs.

1.1.2 Biological relevance

Endogenous retroviruses compromise 5–8% of the human genome (Lander *et al.*, 2001; Belshaw *et al.*, 2004). Although the human genome seems to be littered with retroviral elements, only a few human endogenous retroviruses (HERVs) seem to be actively transcribed. The majority of HERVs have been in fact silenced through a number of modifications which have emerged in response to the impact on host fitness of the integration event. Some examples of how HERVs have been silenced are thought to be due to host-derived restrictions due to selective pressure (Weiss & Stoye, 2013; Stoye, 2012). These include transcriptional silencing, internal recombination events between LTRs which produce orphan integrations known as solo LTRs, and simply the accumulation of multiple mutations and/or deletions. Mutagenesis can be either a by-product of time or a protective mechanism such as seen with the restriction factors APOBEC3G and TRIM5 α (Stoye, 2012). Nevertheless, there are a few HERVs in the genome with intact opening reading frames (ORFs). Those come mainly from the HERV-K group (Villesen *et al.*, 2004), although recent reinfection has yet to be observed. Despite the silencing of ERVs, there remains an association between transcription of ERVs in conditions such as cancer, autoimmunity and some neurological conditions (Ruprecht *et al.*, 2008; Nakagawa & Harrison, 2006; Li *et al.*, 2011; Garcia-Montojo *et al.*, 2013). To be clear, this association is not necessarily causal and may be due to

the surrounding environment of the host (Voisset *et al.*, 2008; Montesion *et al.*, 2018) but further research in distinguishing between causality and effect when it comes to ERVs is still required.

Co-option and development

There are a few HERV genes that have been co-opted by the host to play important roles in regulatory function and development. Among these the well know HERV-FRD and HERV-W *env* genes that have been recruited for syncytiotrophoblast formation and placental fusion, respectively (Mi *et al.*, 2000; Rowe & Trono, 2011). LTRs of HERVs have also been found to act as promoters for genes and to consequently affect gene expression (Cohen *et al.*, 2009).

Other studies have suggested their contribution to host-immune response to exogenous viruses. For example Chuong *et al.* (2016) found that IFN-induced enhancers were introduced and amplified by ERVs in mammals, demonstrating co-option of the ERV, MER41, to regulate host genes encoding immunity factors. In sheep, endogenous retroviruses (enJSRV), which are similar to the pathogenic Jaagsiekte sheep retrovirus (JSRV), have been found to protect the host by blocking early and late steps of JSRV replication (Arnaud *et al.*, 2008). This prevents the host from developing lung cancer (Arnaud *et al.*, 2008).

In humans specifically, our ability to digest starch in the mouth is due to a proviral integration from the HERV-E family upstream of a duplicated copy of the pancreatic amylase gene *AMY2*. This proviral integration became the new promoter for this duplicated amylase gene (*AMY1*) which allowed for expression of amylase in the saliva (Meisler & Ting, 1993). Evidence of positive selection on the salivary amylase gene has been linked to the copy number variation (Perry *et al.*, 2007), where individuals from populations with high-starch diet tend to have more *AMY1* copies. This particular integration has had profound impact from a human evolutionary standpoint given the undoubted influence it has had for the development of agriculture (i.e. an increase in starch consumption). Also in humans, but from an immunological perspective, HERV-K has been found to transcribe during early human development from the point of embryonic genome activation until stem cell derivation from blastocyst outgrowths (Grow *et al.*, 2015). It may possibly be protecting embryos and germ cells from other ERVs or exogenous viral infections (Grow *et al.*, 2015).

Involvement in Cancer

Many ERV integrations in the animal kingdom have been associated with the emergence of cancer (Arnaud *et al.*, 2008; Tarlinton *et al.*, 2006; Bentvelzen *et al.*, 1970). One of the first cancer-associated viruses to be discovered was avian sarcoma leukosis virus, around the beginning of the 20th century (Boeke & Stoye, 1997). In 1936, mammary

carcinoma transmission detected in mice was thought to be a genetic disorder. It was not until 1951 that it was understood to also be infectious in addition to being inherited from one generation to the next. While cancer has been historically known to be the disease most associated with retroviruses in chickens and mice (Stoye & Coffin, 1985), such as the Rous sarcoma virus and the mouse mammary tumour virus (MMTV), the discussion of ERVs initiating cancer processes in humans still bears the controversial past of spurious results (Voisset *et al.*, 2008).

1.1.3 HERV-K

The HERV-K group, particularly of those of the HML-2 family (referred to as HK2), are the most functionally and transpositionally active group known (Medstrand *et al.*, 2001). The letter "K" is related to the fact that the viruses use a lysine tRNA to prime reverse transcription, while the "HML-2" is indicative of their resemblance to the murine retrovirus, mouse mammary tumour virus (MMTV) (van Regenmortel *et al.*, 2000). HK2 happens to be the youngest group of HERVs, having infected the germline only after the human-chimpanzee divergence (Bannert & Kurth, 2006). Being the family of HERVs with the most recent proliferations, HK2 proviruses are exceptionally well preserved and have been observed to express viral proteins and produce viral particles in tumours and cancer cell lines derived from teratocarcinomas and melanomas (Lower *et al.*, 1993; Büscher *et al.*, 2005; Schmitt *et al.*, 2013). Although ERVs have been shown to be infectious in other animals (Magiorkinis *et al.*, 2012), no infectious HERV-K has been detected thus far in nature (Brady *et al.*, 2009). However, defective human-teratocarcinoma-derived virus (HTDV), also termed HTDV/HERV-K, non-infectious virions have been observed in human teratocarcinoma cell lines (Bieda *et al.*, 2001).

Nevertheless, studies have shown the role that HK2 plays in cancer (Lee *et al.*, 2012). One mechanism of tumour induction is attributed to two HK2 alternative splicing protein products (Figure 1.3), Rec and Np9 (Armbruster *et al.*, 2002; Magin *et al.*, 1999). The proteins Rec and Np9 have shown to be expressed during tumour development and thought to function as oncogenes (Denne *et al.*, 2007). Env may also be responsible for initiating cell fusion contributing to tumour progression and even metastasis (Oricchio *et al.*, 2007).

Another way that HK2 may be able of inducing and/or promoting oncogenesis would be through *de novo* insertion or translocation of retroelements within the genome. A recombination event where HERV-K sequences are involved can lead to unstable chromosomal rearrangements and large-scale chromosomal anomalies, which are often observed in metastatic tumours (Romanish *et al.*, 2010).

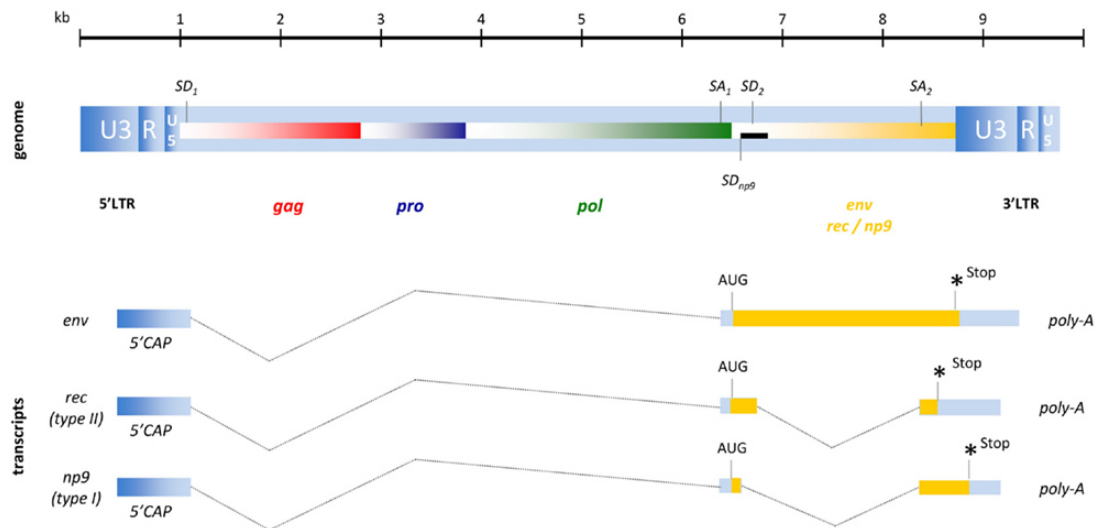


Fig. 1.3.: *Structure of HK2 provirus and RNA transcripts.* Schematic from Hohn *et al.* (2013). The first transcript is the canonical *env* ORF. The second is HK2 type 2 provirus structure with the *rec* ORF. The third structure is of HK2 type 1 provirus with the 292 bp deletion at the 5' end of *env* ORF results in *np9* splice product.

It has been shown that LTRs and other retroelements are capable of misregulating oncogenes, proto-oncogenes, and growth factors in tumours. This was observed in a study of Hodgkin's lymphoma where a hypomethylated solo LTR of an apparent mammalian LTR retrotransposon (MaLR) acted as an alternative promoter in the proto-oncogene colony-stimulating factor 1 receptor (*CSF1R*) and deregulated its normal expression (Lamprecht *et al.*, 2010).

HK2 Polymorphisms

HK2 solo LTRs have been characterised as existing in a 10:1 ratio in comparison to their full-length counterparts in the human genome (Subramanian *et al.*, 2011). These orphan loci also happen to make up the majority of HK2 polymorphisms (Hughes & Coffin, 2001). HK2 serves as a model candidate for polymorphic HERV studies because of its more recent activity and having infected the germline more than once—before and after human-ape divergence (Barbulescu *et al.*, 1999; Barbulescu *et al.*, 2001).

Turner *et al.* (2001) had previously shown that one full-length HERV-K113 with intact ORFs was present in 30% of the individuals tested and that another provirus, HERV-K115, was present in 15%. A study by Marchi *et al.* (2013) used Next Generation Sequencing (NGS) to bioinformatically explore loci that are unfixed in the human population. At the time, that work demonstrated that, on average, individuals have approximately six integrations that are not found in the human reference genome. In a more recent study, Wildschutte *et al.* (2016) searched across WGS from several diverse populations and found that about 26 HK2 integrations can be found in the human population which are not present in the human reference. Polymorphic HK2

integrations are significant for disease association studies since it is already understood that HK2 elements can affect nearby genes or even act as alternative promoters for genes in which they are integrated into (Tomlins *et al.*, 2007; Lamprecht *et al.*, 2010). Insertional polymorphisms can have different effects and signal recent transposition or reinfection. Although the frequency at which ERVs transpose or reinfect may be considered negligible, the connection between retroviral integration and its impact on human evolution can be interesting and can be explored by comparing polymorphic integrations between different populations. Mining genomic population data for unfixed integrations of recently active HK2 is expected to help in establishing associations between this HERV family and disease phenotypes, such as cancer and other autoimmune diseases. I further explore the question of HK2 replication in the human germline across different populations in Chapter 3.

1.1.4 Bioinformatic Challenges and Solutions

The size of the functional portion of our genome is still a topic of discussion even 18 years after the completion of the Human Genome Project. The ENCODE project proposed that 80% of the human genome plays some sort of functional role, on the basis that the majority of the human genome participates in at least one biochemical RNA and/or chromatin related event (Kellis *et al.*, 2014). On the other hand, others have argued that only a meagre 8.2% is responsible for any functionality in the genome if we only take into account regions of the genome which are under purifying selection (Rands *et al.*, 2014). At least 55% of the human genome is composed of repetitive elements (Criscione *et al.*, 2014) and a substantial portion of these elements are of retroviral origin. Repetitive elements have, for the most part, been left aside, condemned to their reputation of "junk DNA." Even with the advances in sequencing technology and with available bioinformatic solutions, repetitive elements remain a challenge to study—partially because of the nature of our current technology (Treangen & Salzberg, 2011) and partially because the technology caters to those working with the "functional" parts of the genome (Treangen & Salzberg, 2011).

The technical challenges that repetitive DNA poses for high throughput sequencing (HTS), essentially lies in how sequences are produced. Popular HTS sequencers tend to generate shorter reads than their Sanger ancestor—approximately 50–150 bp on average but as long as 300 bp in some cases—while producing as many as 10 billion reads per run. The Human Genome Project produced a total of about 300 million reads using Sanger sequencing. Some HTS platforms can sequence a whole human genome in little over a day, for example the HiSeq 2500 from Illumina can produce a single 30 × coverage genome in 27 hours (Reuter *et al.*, 2015). However, even with these massive outputs and 100,000-fold efficiency since the days of Sanger sequencing, the difficulties that arise with repetitive elements are due to the reliance on alignment softwares.

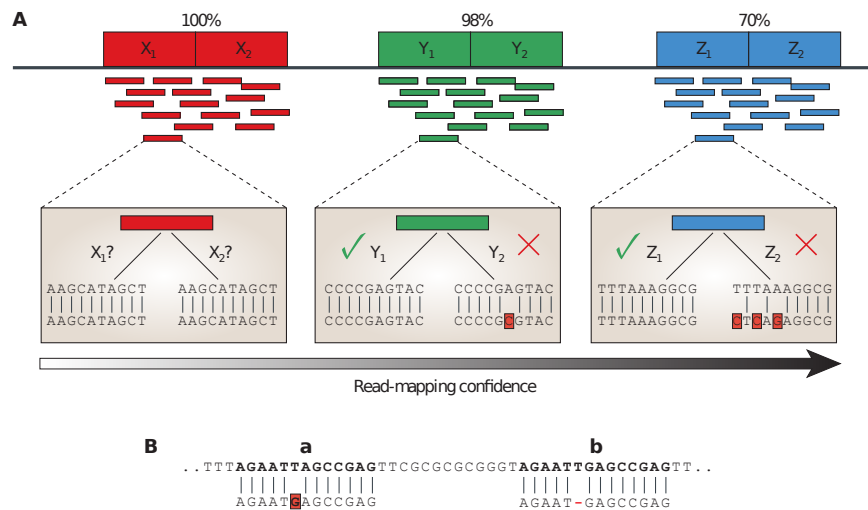


Fig. 1.4.: Ambiguities in aligning repeats. Diagram from Treangen & Salzberg (2011). (A) Above are shown three different tandem repeats with 2 copies each. Beneath these repeats are reads that are sequenced from these regions. For each of the repeats, a single read has been zoomed. In the first case, the red read has a low confidence when mapping within the tandem repeat because it aligns well to both X_1 and X_2 . In the second case, there is a higher confidence in mapping the green reads due to a single nucleotide difference— Y_1 matches better than Y_2 . In the final example, the blue read aligns perfectly to Z_1 whereas against Z_2 , there are three mismatches. (B) The 13 bp shown aligns to 2 locations: "a" and "b". Location "a" has a mismatch and location "b" has a deletion. Depending on parameters used by an aligner, mismatches are often less costly than a deletion. The alignment program may map the read in location "a" but perhaps the source of DNA may have truly belonged to location "b".

Alignment softwares are highly efficient in taking HTS reads and mapping them against an existing reference genome. The majority of the human genome, like many large genomes, is filled with repetitive sequences that appear in several places throughout the genome. From a computational point of view, repeats pose ambiguities in mapping and in de novo genome assembly, which can lead to errors in result interpretation (Figure 1.4).

Repeats that are sufficiently divergent (i.e. not ubiquitously found across the genome) or that are flanked by unique sequence are not too computationally difficult to deal with. A variety of repetitive DNA are well represented in the human genome reference and repeats that diverge enough or that continue onto unique host DNA regions can still be aligned accordingly. However, many repeats, like HERV loci, can be longer than the reads produced by HTS platforms and often are more than 97% identical to another copy of itself elsewhere in the genome (Subramanian *et al.*, 2011; Treangen & Salzberg, 2011).

The length of HTS short reads and the repetitive nature of these elements can produce many multi-mappings, or reads that can align to multiple positions in the reference. Even though the intuitive solution would be to simply assign a read to the best alignment,

the best alignment may not always be correct. By default, penalties for mismatches often tend to be lower than that of deletion or insertion amongst most aligners. This can mislead the aligner to assign a read an erroneous location where it fits with a mismatch when in reality it may have had a true deletion. More often than not, an aligner can simply leave a read unmapped if it aligns to too many positions or if it is too ambiguous. Discriminating whether a deletion, insertion or mismatch should be more tolerated amongst aligning parameters remains a difficulty we still face. In the case of the human reference, which is neither representative of the human population nor an ancestral sequence, mapping difficulties can arise due to some regions being poorly assembled, if present at all, while other challenges arise by the mere fact that repetitive elements can indeed be identical, or nearly identical, across the genome (Sherman *et al.*, 2019; Ballouz *et al.*, 2019).

Although there are now HTS platforms that produce long-reads such as Oxford Nanopore Technologies and Pacific BioSystems, these single-molecule sequencers still have elevated error rates in comparison to short-read platforms like Illumina and IonTorrent (Reuter *et al.*, 2015). A potential solution to resolve repetitive regions could be to merge the use of both short-read and long-read technologies for loci of interest (Karamitros & Magiorkinis, 2015). However, the current majority of public available data and population level studies use short-read sequencing platforms (Treangen & Salzberg, 2011). Future consortium studies may implement a combination of both to better understand the role of repetitive elements in the human genome.

With challenges such as the ones described, bioinformatic tools like RetroSeq and Tangram have been written to resolve some of the difficulties in finding repetitive elements. RetroSeq focuses on discordant pairs and pairs where a read is mapped to the genome reference and its mate is not (Keane *et al.*, 2013). The unmapped or discordant mates are aligned to a supplied reference library of mobile element sequences using and forward-and-reverse strands which are then clustered. RetroSeq henceforth uses the density of the clusters and searches soft-clipped reads, reads that have been partially trimmed at the ends to match a locus, to refine the break points of the repetitive element integrations. Tangram is similar in that it searches for read pairs where one mate is aligned to the genome reference and the other mate is unmapped or soft-clipped. Tangram then takes unmapped or soft-clipped reads and attempts to align them with a given mobile element sequence using its "split-read" module. The breakpoint is then determined by the alignment position of the first segment (Wu *et al.*, 2014). In a similar line of thinking, Marchi *et al.* (2013) produced a "trimmed-read" algorithm to mine for unfixed HK2 loci. Reads that mapped poorly against the reference or which were unmapped were trimmed if a direct 8-nucleotide match with the beginning or end (sense and antisense) of the HK2 LTR was confirmed. Trimmed reads were then realigned to the reference genome and compared to RetroSeq output positions. An

in-house script, BreakAlign, was written to establish the breakpoints of HK2 integrations amongst chimeric reads (Marchi *et al.*, 2013).

In Chapter 2 I present a tool built off the method used by Marchi *et al.* (2013) to identify novel HK2 integrations in the human population. This work is a now published computational approach, STEAK (Santander *et al.*, 2017) (generated as part of this thesis).

1.2 Human Origins

Evidence of viral infections in the human fossil record is at best difficult to recognise and at worst unrecognisable. Considering that not all viral infections lead to pathogenesis, searching for signs of ancient viral infections in prehistory, particularly ones with no morphological impact are nigh on impossible to detect with archaeology alone. Yellow fever and Spanish flu (influenza) have taken place in historical times but the causal agents are likely to be much older, (~2,400 years) (Gibbs *et al.*, 2001; Bryant *et al.*, 2007). While we may be unable to recognise viral infections on skeletal human remains, we can try to make sense of them by what *within* these remains. HERVs are a genetic fossil record of the once exogenous retrovirus that infected our primate and more recently, *Homo* ancestors (Bannert & Kurth, 2006). We can use modern human genomes today to mine for them in order to understand what we have retained and what we have lost either due to genetic drift or selection pressures.

Likewise, modern human genomes today can be used to mine other prehistoric events, such as hybridisation and interbreeding (Vernot & Akey, 2015; Vernot *et al.*, 2016; Racimo *et al.*, 2015; Lachance *et al.*, 2012; Sankararaman *et al.*, 2016; Seguin-Orlando *et al.*, 2014). We have been fortunate to have found remains of archaic hominins outside of Africa with endogenous host DNA to sequence and therefore providing us with a clearer understanding of our own genetic past. But not all remains of the fossil record are suitable to reveal their genetic history. Here I present our understanding of human origins to date beginning with the fossil record and moving into the deciphering of interbreeding events between modern humans and other species of *Homo*.

1.2.1 The emergence of humanity: *Homo* and anatomically modern humans

Africa is the cradle of life and the birthplace of our species. In particular East Africa is the most extensively excavated area where some of the earliest fossils of our genus, *Homo*, have been discovered (Kimbel *et al.*, 1997). For example, *H. habilis* dated

~2.3-1.4 MYA is the oldest known member of the genus and distinctively known for early evidence of butchered animal bones and basic stone tool use (Blumenshine *et al.*, 2003). Yet southern Africa is also home to some of the oldest fossils of the genus *Homo* with a fossil record that begins from about 2 MYA with pronounced transitional forms ranging around 200,000–600,000 ya (Dusseldorp *et al.*, 2013). A more widespread form found across Africa and Eurasia is *H. erectus* which may have existed as far back as 1.9 MYA to as late as 100,000 ya or even more recently (Antón, 2003). The beginnings of notable brain size increase starts with *H. erectus* which also displays, like anatomically modern humans (AMH), smaller molars and a lack of forelimb morphology related to climbing (Antón, 2003). The span of *H. erectus* throughout the Old World alone suggests a cognitive capacity and ecological flexibility to adapt and thrive in different environments. Notably, findings of other archaic hominins in eastern Asia such as *H. floresiensis* (Brown *et al.*, 2004) from Flores, Indonesia and *H. luzonensis* (Détroit *et al.*, 2019) from Luzon in the Philippines have shown to share anatomical similarities with *Australopithecus* and present on those respective islands since at least 700,000 years ago (Sutikna *et al.*, 2016; Van Den Bergh *et al.*, 2016; Ingicco *et al.*, 2018). This compellingly suggests an early exodus from Africa although *H. erectus* remains in the Georgian Caucasus suggest presence outside of Africa for at least 1.85 MYA (Ferring *et al.*, 2011; Wood, 2011).

From *H. erectus* emerges *H. heidelbergensis* around 700,000 ya and has been considered an archaic form of *H. sapiens* with body proportions, dental profile and cognitive abilities similar to AMH (Rightmire, 2009). Archaeological sites and behavioural evidence suggest that *H. heidelbergensis* hunted large game, produced intricate tools and had even learned to control fire (Roebroeks & Villa, 2011; Barham *et al.*, 2014). Although no remains that are clearly considered *H. heidelbergensis* have been sequenced, mainly due to no surviving aDNA to extract, it is a form that has been disputed among researchers to be potentially ancestral to Denisovans—an archaic human that had no known skeletal morphology until very recently (Chen *et al.*, 2019). Denisovan samples that have been used for aDNA extraction have been limited to a finger bone and a molar which have been used for DNA sequencing (Reich *et al.*, 2010). The hypothesis that *H. heidelbergensis* may have been ancestral to Denisova stems from a study which established that the remains at Sima de los Huesos in Spain dated to 300,000 years ago carried mitochondrial DNA (mtDNA) closer to Denisovans than to that of *H. neanderthalensis* (Neanderthals) (Meyer *et al.*, 2014). However, a follow-up study two years later showed that the nuclear DNA (nuDNA) of these remains was Neanderthal (Meyer *et al.*, 2016) and they are now considered as ancestral to European Neanderthals. These studies support earlier proposals of AMH, Neanderthal, and Denisovans putatively diverging off from *H. heidelbergensis* in Europe (Rightmire, 2009; Stringer, 2012).

Neanderthal fossils were first discovered in Flémalle, Belgium but the first fossils to be identified as "archaic" human were in Neander Valley, Germany supplying this archaic form its name (Schmitz *et al.*, 2002). With stout physiques, complex behaviours, and brain size similar to AMH (Hublin, 2009), Neanderthal populations were present across Eurasia and persisted to what may be as late as 28,000 ya, well after AMH had reached the European continent (Higham *et al.*, 2006; Devière *et al.*, 2017; Hublin, 2017; Zilhão *et al.*, 2017), although the precise radiocarbon dating on locations such as Vindija Cave (Croatia) and the Iberian peninsula are still under debate.

Since the initial studies and claims surrounding Neanderthal morphology (Hublin, 2009), DNA sequencing has allowed us to better understand their relationship to, or larger than, AMH. The first studies focusing on Neanderthal mtDNA suggested that modern humans and Neanderthals did not interbreed, as shown in the phylogenetic tree relating modern human and Neanderthal mtDNA which suggested a deep divergence of about (~500,000 years) with no contribution to present-day modern human populations (Krings *et al.*, 1997; Hodgson & Disotell, 2008). But over a decade later from the first Neanderthal mtDNA evidence, a different story was revealed. nuDNA retrieved from remains found in Croatia, Vindija cave, allowed the first draft Neanderthal genome to be assembled and released (Green *et al.*, 2010). In this new story, all non-African populations today share about 1-3% genetic similarity with Neanderthals based on the amount of shared derived alleles between modern humans and Neanderthal autosomal DNA (Green *et al.*, 2010). This was further corroborated by the remains found in the Altai mountains of Siberia, Denisova Cave (Reich *et al.*, 2010). This unique site provided fossil evidence not only to decipher interbreeding between Neanderthals and AMH (Prüfer *et al.*, 2014) but also contributed the remains that were crucial in discovering the unknown archaic hominin, Denisova (Reich *et al.*, 2010). While we may not know much about Denisova anatomy, we do know that they too interbred with AMH (Reich *et al.*, 2010; Vernot & Akey, 2015) contributing to some populations as much as 6% through admixture (Reich *et al.*, 2011; Skoglund *et al.*, 2016; Sankararaman *et al.*, 2016; Vernot & Akey, 2015). Unsurprisingly having found remains of both archaic hominins in the Denisova Cave, this site has also shed light on the interbreeding between both archaic species in a study that has described a first generation Denisovan-Neanderthal admixed individual (Slon *et al.*, 2018). Other Neanderthal samples like those from Vindija Cave (Croatia) (Prüfer *et al.*, 2017) and at a lower coverage, Goyet (Belgium), Les Cotte (France), and Mezmaiskaya (Russian Caucasus) (Hajdinjak *et al.*, 2018) have also contributed to understanding the phylogenetic relationship of Neanderthals, Denisovans, and modern human populations although at a lower resolution since low-coverage genomes are unable to provide a fine-scale understanding of the variation present in an individual.

Hybridisation between AMH and archaic hominins, also known as archaic admixture, is now widely accepted as part of human evolution. While not a novelty in other

areas of biology (Tung *et al.*, 2008; Taylor & Larson, 2019), the idea of hybridisation between different subspecies of the genus *Homo*—in particular with *H. sapiens*—has not always been well received. This refusal of accepting archaic admixture at the time had, on one hand, to do with the genetic evidence supporting one of the classical models of human evolution (Figure 1.5), the 'recent out-of-Africa' theory (RAO) (Cann *et al.*, 1987), and on the other hand, a lack of convincing evidence that the sharing of alleles between archaic hominins and non-Africans was in fact due to interbreeding (~50,000 ya) (Green *et al.*, 2010; Reich *et al.*, 2010; Plagnol & Wall, 2006) and not persistent subdivision of an ancestral population (Eriksson & Manica, 2012) which could create excess similarity between populations despite the absence of recent admixture. Although this model is now shown to be unsupported by theoretical and biological evidence, that debate continued for about 4-5 years (Martin *et al.*, 2015; Fu *et al.*, 2014).

Among the classic models of evolution, the two competing theories with the most traction in human evolutionary studies have been RAO (Howells, 1976; Stringer & Andrews, 1988) and multiregional evolution (MRE) (Wolpoff *et al.*, 1984) (Figure 1.5: left). The RAO theory proposed that *H.sapiens* evolved in Africa and migrated relatively recently from there, expanding outside of Africa and replacing archaic hominins like Neanderthals (Stringer & Andrews, 1988). To be clear, this model did not deny the possibility of interbreeding between modern humans and archaics but rather stated that any mating that could have occurred was likely to have resulted without viable and/or successful offspring thus making no impact on modern populations today (Mayr, 2000; Stringer, 2001). Contrastingly, the MRE theory put forth that the evolution of *H.sapiens* took place across the Old World emerging from discrete forms of *Homo* in Eurasia and Africa, allowing for gene flow between AMH and regional archaic hominins throughout time and with little reproductive isolation (Wolpoff *et al.*, 1984; Holliday *et al.*, 2014). Both of these model represent extremes of the human evolution hypotheses spectrum.

With what we now know from ancient DNA and paleogenomics (Green *et al.*, 2010; Reich *et al.*, 2010; Reich *et al.*, 2011; Vernot & Akey, 2015), an intermediate or "leaky replacement" model (Figure 1.5: right) seems more representative of the true human evolutionary story outside of Africa. In particular a combination of Brauer's replacement with hybridisation model (RWH) (Bräuer, 1984; Bräuer, 1992; Bräuer, 2008) and Smith's assimilation model (AM) (Smith, 1984; Smith, 1994; Smith, 2010) is more appropriate if we consider recent genetic studies that show that archaic admixture between AMH and archaic hominins indeed took place but with at most a 6% genetic contribution (Reich *et al.*, 2011; Vernot & Akey, 2015; Browning *et al.*, 2018), resonating with the RWH model view of some a few archaic genes entering a predominantly modern human gene pool (Stringer, 2001; Holliday *et al.*, 2014). Nevertheless, recent studies have shown that this small contribution has played a significant role in the selection that has shaped the phenotypic variation of present-day populations, specifically related

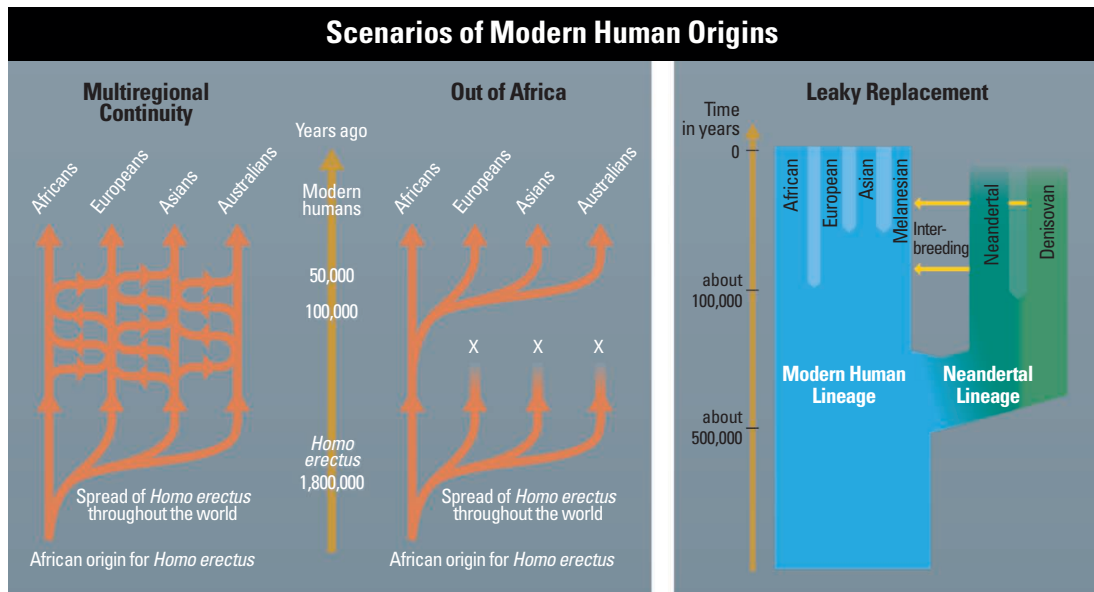


Fig. 1.5.: Proposed models of human evolution. The left side depicts the two main human evolution models that were challenged and now replaced by an intermediate model (right side) which considers interbreeding with archaic hominins outside of Africa. From Gibbons (2011) as reported in *Science* (Gibbons, 2011). Reprinted with permission from AAAS.

to immunology and disease (Racimo *et al.*, 2015; Quach *et al.*, 2016; Deschamps *et al.*, 2016; Quintana-Murci, 2019; Pimenoff *et al.*, 2017; Houldcroft & Underdown, 2016) as well as physiological adaptations like those present in Tibetans and Himalayan Sherpas who have an introgressed Denisovan haplotype in *EPAS1* (Huerta-Sanchez *et al.*, 2014; Bhandari *et al.*, 2017). This falls in line with the AM model’s emphasis on the importance of gene flow between AMH and archaics (Stringer, 2001; Holliday *et al.*, 2014). In recent years, original proponents of the RAO model have now updated the model to account for archaic admixture but with variations in the degree of gene flow (Galway-Witham & Stringer, 2018).

Evidence of archaic introgression in non-Africans

Ancient DNA has revolutionised the methods we use to understand our evolutionary past and therefore update proposed models of human evolution. Today several methods leverage genetic variation between individuals and populations, such as mutations, to make sense of our ancestral relationships.

Germline mutations represent alterations in an individual’s DNA, differentiating them from other individuals. Scally and Durbin (Scally & Durbin, 2012) present a revised human base substitution rate of approximately $0.5\text{--}1 \times 10^{-9}$ per base pair per year by combining sources from the fossil record (Benton & Donoghue, 2007), the generation time of great apes (Matsumura & Forster, 2008), as well as human pedigree studies which would suggest about 80 new mutations occur per generation (Besenbacher *et al.*,

2015; Lynch, 2010; Roach *et al.*, 2010). Typically mutations occur in new positions across the genome and thus populations collect unique mutations over time which can be used to distinguish one population from another (but also share mutations at different frequencies). However, other factors such as reproductive age can also contribute to germline mutations in an individual but these individual based differences are not expected to overwhelm population variation (Maklakov & Immler, 2016; Gao *et al.*, 2019). It is for this reason, given that Neanderthal has diverged from the common ancestor about 270,000-512,000 years ago (Green *et al.*, 2010; Hajdinjak *et al.*, 2018) (Figure 1.6), we can expect that the given DNA of any two humans will be closer to each other on average when compared to a Neanderthal segment. This expectation is under the framework of an infinite sites model where we assume that there are an infinite number of loci where mutations can occur, (2) that new mutations always occur in novel sites, and (3) that there is no recombination (Tajima, 1996). Although overly simplistic, this model allows us to approximate, underestimate even, how much variation we share between and within human populations.

In addition to mutation, recombination is a process which breaks and mixes homologous chromosomes during meiosis. Throughout this process, recombination generates a distribution of fragment lengths, some of which would be short and overlapping with possible signals from incomplete lineage sorting (ILS)—a case where two or more lineages from different populations share a common ancestor more recently than their respective most recent common ancestor within populations, producing a population tree which is at odds with a particular gene tree. For example, multiple alignments of gene for human, chimpanzee, gorilla and orangutan sequences have shown that a significant fraction of human genes are more similar to gorilla despite the human genome being over all closer to chimpanzee (Hobolth *et al.*, 2007).

Both recombination and mutation are the basis of understanding the two key properties of any introgressed tract in relation to the time since the admixture event: length and diversity. Consequently, these processes need to be accounted for when developing ways to detect tracts of archaic introgression.

The difference in the lengths of the tracts should provide a way to tell whether a shared tract with an archaic hominin is in fact introgression from a distantly related population or the genomic relic from an early shared population structure (Figure 1.7).

A crucial study that helped disentangle these two scenarios when the debate on whether Neanderthals and Denisovans had contributed to modern human populations was on-going was conducted by Fu *et al.* (2014). The femur of a ~45,000 Siberian male, Ust'-Ishim, was sequenced and found to display long Neanderthal tracts indicative of archaic introgression. Since Ust'-Ishim was closer to the admixture event between AMH and Neanderthals, if admixture did indeed take place then the tracts would be

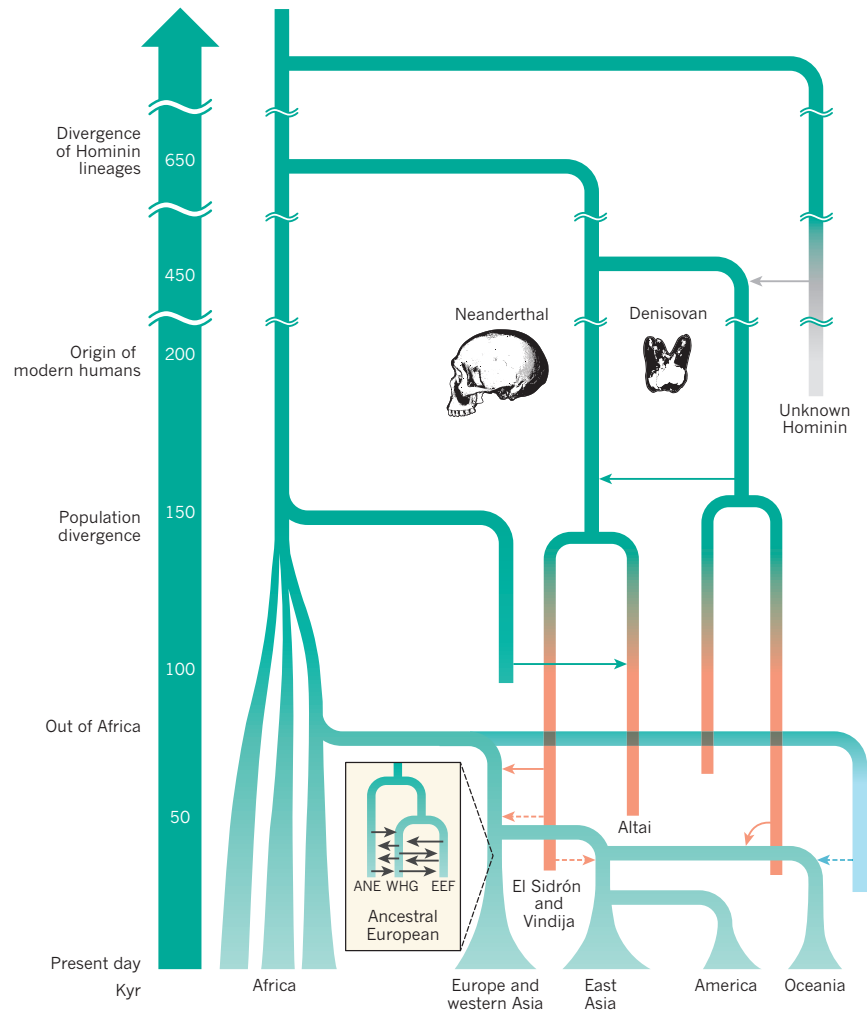


Fig. 1.6.: *Simplified model of archaic hominin and modern human history from Nielsen et al. (2017) featured in Nature.* The relationships between populations today and the estimated times where they split off from the most recent common ancestor are shown here. The solid lines represent well-established admixture events whereas dashed lines stand for tentative admixture between modern and archaic humans. The model also displays the potential a small fraction of ancestry observed in the Oceanic population, derived from an early out-of-Africa event. In the small inset we see an insight into the history of ancestral Europeans which studies have found to be an admixture between ancient north Eurasian (ANE), early European farmer (EEF), and west European hunter-gatherer (WHG).

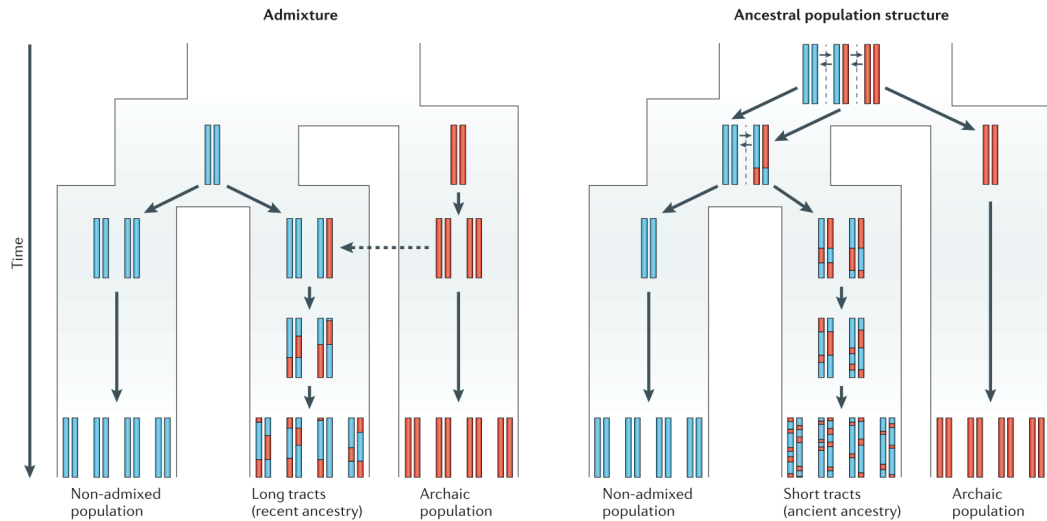


Fig. 1.7.: Admixture versus ancestral population structure. Expected lengths of archaic-like tracts for an admixture scenario (left panel) and an ancestral population structure scenario (right panel). Because there is less time for recombination to break down the migrated tracts (red) in the admixed population, the expected tract length in this case will be longer than in the case of ancestral population structure (Racimo *et al.*, 2015).

considerably longer than what is observed in non-Africans today since recombination would have less time to break them down. This would be especially true if there were multiple episodes of Neanderthal introgression as has been shown in an ancient human from Romania (Fu *et al.*, 2015; Villanea & Schraiber, 2019). On the other hand, if admixture had not taken place then the tracts would still be quite short given that the ancestral population is much further back in time than a putative admixture event. Ust'-Ishim (Fu *et al.*, 2014) was key in demonstrating concrete biological evidence to support studies that used linkage disequilibrium patterns to detected archaically introgressed segments in modern human genomes (Plagnol & Wall, 2006; Wall *et al.*, 2009; Hammer *et al.*, 2011; Vernot & Akey, 2015; Sankararaman *et al.*, 2014; Vernot & Akey, 2015; Lachance *et al.*, 2012; Hsieh *et al.*, 2016a; Browning *et al.*, 2018). How these methods work, what they have found in non-African individuals, and existing caveats when implemented on other populations are extensively discussed in Chapter 4 of this thesis.

1.2.2 Modern Human Origins in Africa

So far I have discussed work that primarily focused on events that affected populations which were part of the out of Africa expansion. In this section I expand on findings related to the cradle of humanity: Africa.

Forms displaying transitional AMH features have appeared in the record of southern Africa around 100,000-300,000 ya and fully AMH emerge here approximately 120,000 ya. Strikingly, while the fossil record for North Africa has been limited, it is where the oldest fossils meeting the criteria of AMH can be found—particularly the ancient remains from Jebel Irhoud, dated to 300,000 ya (Hublin, 2017). Interestingly, in a younger time period of 100,000-60,000 ya, North African remains still denote archaic morphological traits (i.e. heavy brow ridges and distinctly large teeth), despite anatomical modernity such as an anatomically modern face and cranial vault (Rightmire, 2009).

Fossils dating to 160,000-180,000 ya in East Africa, Omo Kibish and Herto, have been observed to be fully AMH and have often been used to support East Africa as the birthplace of modern humans (White *et al.*, 2003; McDougall *et al.*, 2005). Although, an East African calvaria from Lukenya Hill in Kenya, while regarded as *Homo sapiens*, demonstrates morphological features better represented in archaic hominins yet dated a recently as ~23,000 ya (Tryon *et al.*, 2015).

Of the notably few and recent assemblages from western and central Africa, some remains display features that are more in line with archaic humans but which are fairly young such as the Ishango site in the eastern part of the Democratic Republic of Congo dated to 20,000-25,000 ya and Iwo Eleru skeletal remains from Nigeria dating back to ~13,000 ya (Harvati *et al.*, 2011). Yet the human fossils from Shum Laka in Cameroon, which date 3,000-7,000 ya, are fully AMH (Lavachery, 2001).

Archaic admixture in Africa?

Some have suggested, based on both morphology and the technological replacement that took place from central to western Africa in a mosaic-like fashion, that this could putatively support a model where archaic admixture took place between surviving archaic populations and modern humans (Scerri, 2017; Schlebusch & Jakobsson, 2018). As consistent as some of the fossil record may be with supporting archaic introgression in Africa, understanding how admixture impacts morphology to better understand the mosaic-like distribution for early *Homo sapiens* across different regions has been strongly overshadowed by the efforts to prove that modern humans originated from Africa (Henn *et al.*, 2018).

Despite the several assemblages of fossils found across Africa, DNA preservation from these samples is often compromised—a great deal owed to climatic related challenges such as warmer weather accelerating the degradation of older DNA (Campana *et al.*, 2013). Alternatively, recovering demographic history from modern genomes can also be challenging with lack of large genomic datasets representing the whole of Africa—especially in the light of the continent’s complex deep past. The movements, and subsequent gene flow, from the agricultural expansion and spread of pastoralism have

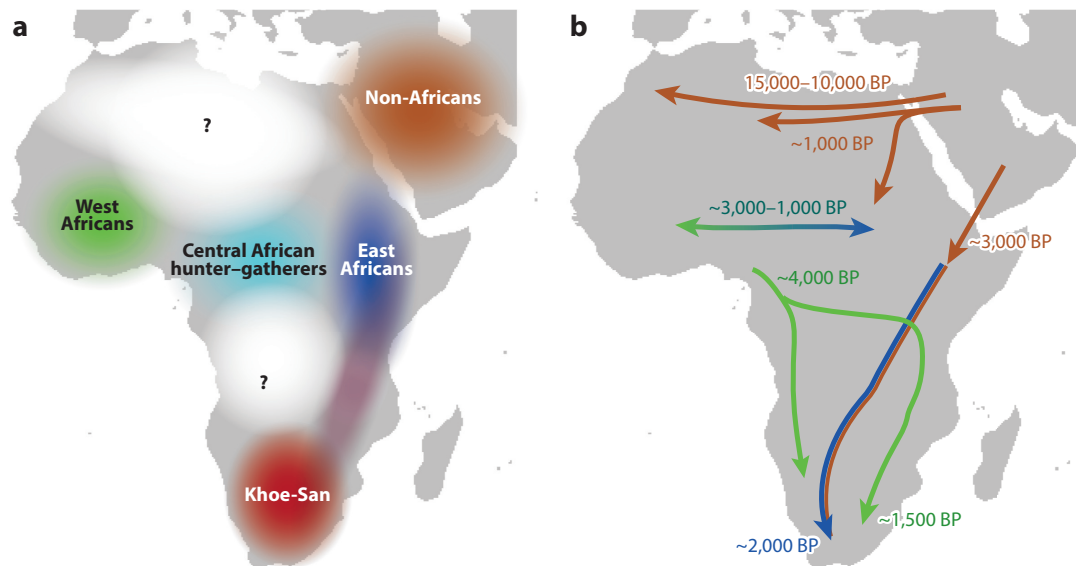


Fig. 1.8.: Stratification and migration across the African continent. (a) Map of major pre-farming population stratification across the African continent, showing non-Africans (brown), East Africans (dark blue), West Africans (green), central African hunter-gatherers (light blue), and Khoe-San (red). (b) Migration routes related to the expansion of herders and crop farmers during Holocene times. The green arrows represent the Bantu Expansion, the brown arrows represent the Eurasian back-migration, the brown/blue arrow represents the southward migration of mixed East African/Eurasian pastoralists, and the blue/green arrow represents bidirectional migration across the Sahel belt. Abbreviation: BP, years before present. (Schlebusch & Jakobsson, 2018).

cloaked the ancient variation pertaining to early modern humans (Tishkoff *et al.*, 2009). Consequently, this has complicated demographic inferences that can be recovered regarding ancient African population structure. Although recent ancient samples have provided insight into the relatedness between populations prior to monumental demographic events (Skoglund *et al.*, 2017), these sort of samples with endogenous aDNA remain rare and will most likely accumulate at a slower rate.

The veils over African prehistory: Challenges of population structure and complex demographic history

Making sense of the deep past of African populations is a challenge that still remains elusive despite the advances in sequencing modern and ancient DNA alike. While we have managed to infer some of the prehistorical stratification of African populations (Figure 1.8 a), landmark migration events in human history (Figure 1.8 b) have obscured the unique signatures of prior population structures on the continent. One of the greatest expansions where populations were displaced and also mixed with other widespread populations was the agriculturalist expansion, also referred to as the Bantu Expansion (Figure 1.8 b). It was led by Bantu-speaking people around ~4,000 ya spreading from the border of Nigeria and Cameroon in west-central Africa. Other migrations have also shaped African genetic variation including pastoralist movement

from East to South and movement across the Sahel belt (East and West). These movements and their influence on the genetics of African populations are further explored in Chapter 5.

1.3 Motivation and Problem Statement

In recent years, questions have moved from whether archaic admixture took place to how much has taken place to finally posing questions related to selection, adaptive introgression, and adaptive immunity. But as riveting as those last questions may be, the work presented here are lessons to heed. The focus of this thesis is to shed light on the challenges, potential solutions, and necessary caution required before answering those questions. By dissecting methods and algorithms which have been widely-accepted in the field, this thesis hopes to propose some insight for those working on topics that have been less catered to in retrovirology and human evolutionary genetics.

1.4 Thesis Structure

The aims of this thesis can be outlined per chapter as follows:

Chapter 2

The purpose of this project was to collaboratively develop an algorithm and bioinformatic tool that would identify novel HK2 integrations in the human population. STEAK (Santander *et al.*, 2017) characterises novel and polymorphic HK2 integrations and I show its capacity by implementing it on whole genome sequences (WGS) from different cancer patients and a high-coverage family trio. Furthermore, I adapted our algorithm to work with target enrichment sequencing data in order to identify novel HK2 integrations from LTR-bait data. I compare our method to pre-existing programs and show higher precision and a lower rate of false-positives.

Chapter 3

Most HERVs are relics from ancient events, infections that have taken place in Old World monkeys, or at the mammal basal lineage. HK2 is the exception to this with at least 150 human-specific integrations of which about a third are unfixed in the human population. By using diverse population data, what can we say regarding the distribution of HK2 polymorphisms in the human population and approximately how long ago did HK2 germline activity die out?

Chapter 4

The goal here is to study the LD-based method, S^* , and its behaviour across different

simulations representing a range of demographic models with and without admixture. This will allow us to understand what to keep in mind when implementing it on populations with very different demographic histories. Most importantly it will allow us to understand: does S^* robustly detect archaic admixture?

Chapter 5

Recent studies have been able to extract endogenous host DNA from ancient human remains in Africa (Skoglund *et al.*, 2017; Gallego Llorente *et al.*, 2015; Schlebusch & Jakobsson, 2018) but an archaic African hominin has yet to be sequenced. In absence of an archaic hominin reference for Africa, we can only turn to modern human genomes to understand our evolutionary past within Africa. Bearing in mind what I learnt in Chapter 4, I use LD-methods such as S^* and Sprime on African individuals from the Simon Genome Diversity Project (SGDP). Furthermore, I also evaluate the relationship between present-day Africans and Eurasian archaics. The aim is to cautiously consider the contentious question: did archaic admixture happen in Africa?

Chapter 6

This is a wrap up of my findings related to the activity of HK2 and the outstanding question of archaic introgression in Africa. I discuss my findings in light of recent literature and also discuss future directions related to the work presented in this thesis. The goal of this section is to highlight the importance of Africa for both viral and human host evolution.

STEAK: A specific tool for transposable elements and retrovirus detection in high-throughput sequencing data

Note: The work presented in this chapter comes from a collaborative publication in *Virus Evolution* (2017). My specific contribution to this work can be found in both the publication and in the statement at the end of this chapter. I use the term "we" instead of "I" in this chapter as it is a direct copy from the published article.

2.1 Background

High-throughput sequencing (HTS) has undoubtedly revolutionised genome sequencing with technology that has seen a 50,000-fold cost drop and an increase in capacity since the days of the Human Genome Project (Liu *et al.*, 2012; Goodwin *et al.*, 2016). Several consortia, like the Cancer Genome Atlas (TCGA) (Weinstein *et al.*, 2013), have all made use of HTS providing both researchers and clinicians with copious amounts of genomic data in the last decade. Most health- and disease-related researches have focused on exons, conventionally considered the functional portions of the genome. This has promoted the widespread use of HTS amongst disease consortia where short-reads constitute a lesser challenge for correct alignment and genome assembly. While short-read lengths remain useful and informative for unique areas of the genome (e.g. the exome), repetitive, or low-complexity regions, suffer from assembly ambiguities that arise due to HTS read length. Even with the recent advances in sequencing technology and with advanced bioinformatics solutions, repetitive regions remain a challenge - partially because of the nature of our current technology (Treangen & Salzberg, 2011) and partially because of the strong focus on working with protein-coding parts of the genome.

At least 55% of the human genome is composed of repetitive elements (Lander *et al.*, 2001), mostly transposable elements (TEs). A number of TEs, such as LINEs, SINEs, and SINE-VNTR-Alu (SVAs), have been found to actively move around the human genome with a potential pathological burden (Ostertag *et al.*, 2003; Mills *et al.*, 2007; Solyom & Kazazian, 2012; Evrony *et al.*, 2015). HERV-K HML-2 (or HK2) is a thirty-

million year-old family of endogenous retroviruses that continued integrating in the human genome even after the human-chimp divergence. Some HK2 integrations remain unfixed in the population (Marchi *et al.*, 2013; Wildschutte *et al.*, 2016), moreover, every individual carries approximately ten polymorphic HK2 integrations (Marchi *et al.*, 2013). The influence that TEs and human endogenous retroviruses (HERVs) have on altering genetic activity due to somatic rearrangements also implies a potential role in the development of disease (Hohn *et al.*, 2013) for example through insertional mutagenesis (Djebali *et al.*, 2012; Solyom & Kazazian, 2012; Shukla *et al.*, 2013; Criscione *et al.*, 2014).

Many cohorts have made use of short-read technology, however, the limitations that HTS short-reads pose for studying TEs with disease consortium data now presents an algorithmic and theoretical challenge for mapping reads with repetitive stretches in their original genomic location (Simola & Kim, 2011; Li *et al.*, 2012). Most of the available softwares make use of paired-end read information or chimeric reads (i.e. reads which are part host and part TE), to identify the genomic location of a TE (Keane *et al.*, 2013; Wu *et al.*, 2014). There are several approaches that use similar methods for discovery of TE integrations in comparison to the reference genome (Lee *et al.*, 2012; Keane *et al.*, 2013; Wu *et al.*, 2014). Other algorithms include trimming chimeric reads of the TE portion to then be remapped to the host reference genome (Marchi *et al.*, 2013). Alternatively, some softwares search for structural variation differences between the reference and HTS data such as insertions, deletions, inversions, inter-, and intra-chromosomal translocations (Chen *et al.*, 2009b).

Here we present a broadly applicable approach to annotate (i.e. mark presence or absence) known and characterise unknown insertion sites for TEs in a variety of sequencing projects. We use HK2 as a mobile element model to evaluate the identification of polymorphic integrations because of its standing as an endogenous retrovirus (Boeke & Stoye, 1997).

We have generalised the algorithm previously described by Marchi *et al.* (2013) to develop a program that will assist in marking presence or absence of any given sequence element within a reference genome as well as identify novel integrations of that sequence element compared to the reference genome. We benchmark the ability of our program Specific Transposable Element Aligner (HERV-K) (STEAK) to discover novel TE and exogenous retrovirus insertions in addition to marking the presence/absence of TEs annotated in the reference genome. We evaluate our method on simulated data as well as high-coverage HTS projects, such as those used in clinical WGS, and compare to competing systems (Table 2.1).

Tab. 2.1.: Software for detecting TEs and viruses in WGS data.

Software	Detection target	Detection Method	Detects in Reference?	Requires specific aligning?	Third -party tools	Parallellised?	Implementation
RetroSeq (Keane <i>et al.</i> , 2013)	Transposable elements	Discordant reads, then split reads	No	No, but must be in BAM	SAMtools (v0.9), bcftools, exonerate, BEDtools	No	Perl
Tangram (Wu <i>et al.</i> , 2014)	Transposable elements	Split reads and discordant reads simultaneously	Yes [†]	Yes, MOSAIK	MOSAIK (2.0), zlib, pthread lib	Yes	C, C++
VirusSeq (Chen <i>et al.</i> , 2013)	Viruses	Unmapped reads for general detection; Discordant and split-reads for integration site detection	No	Yes, MOSAIK	MOSAIK (0.9.0891)	Yes	Perl, C, C++
MELT (Gardner <i>et al.</i> , 2017)	Transposable elements	Discordant reads, then split reads	Detects deletions	No, but must be in BAM	Bowtie2	No	Java
ViralFusionSeq (VFS) (Li <i>et al.</i> , 2013)	Viruses	Unmapped reads for general detection; Discordant and split-reads for integration site detection	Yes [†]	Yes	BWA, SAMtools, BLAST, CAP3, SSAKE	Partially (BWA portion)	Pipeline (Perl)
Tlex2 (Fiston-Lavier <i>et al.</i> , 2015)	Transposable elements	Looks at host and annotated TE flanks. Searches for split reads.	Only detects in reference.	No	MAQ, SHRIMP, BLAT, RepeatMasker, Phrap	Partially (MAQ)	Pipeline (Perl)
STEAK	Transposable elements and viruses	Split-reads, retrieves mate for PE data	Yes	No	Aligner of choice, BEDtools	Yes	C, C++

[†] Can detect in reference but not designed to mark-presence absence of reference insertions.

2.2 Materials and Methods

2.2.1 General benchmarking parameters

We ran RetroSeq and Tangram under the instructions provided on their respective Github sites. For VirusSeq, we benchmarked using the instructions from the User Manual (Chen *et al.*, 2013). We ran STEAK using the parameters provided in the supplementary data. This along with the command lines for the software we benchmarked against can be found in Supplementary Note S4, p. 158.

2.2.2 WGS from the Cancer Genome Atlas

We analysed four normal blood derived samples (i.e. germlines) from patients that were available at an average range of 45× to 50× coverage. These paired-end WGS (2×100-bp) were all sequenced using Illumina Genome Analyzer platform technology (Table 2.3). All whole genomes downloaded from TCGA database were pre-aligned to the human reference, version hg19. More information on TCGA samples and cancer type abbreviations can be found in Supplementary Table S2, p. 144.

2.2.3 WGS of CEU (Utah Residents with Northern and Western European Ancestry) pedigree from the 1000 Genomes Project

We analysed the Illumina platinum genomes for the NA12878/NA12891/NA12892 pedigree. These samples were WGS sequenced at a 50× coverage with both paired-end (2×100-bp) and single-end libraries with Illumina HiSeq 2000 system. All whole genomes were downloaded from 1000 Genomes Project data portal where further information regarding fragment size and library preparation can be found.

2.2.4 Simulated datasets and benchmarking

For the TE-simulated dataset, we tested on a simulated human chromosome 20 of 5× coverage and a read length of 76-bp created by Wu *et al.* (2014) in the release of Tangram. We compared VirusSeq (Chen *et al.*, 2013), RetroSeq (Keane *et al.*, 2013), Tangram (Wu *et al.*, 2014), and STEAK on the detection sensitivity of AluY non-reference insertions in this simulated chromosome 20 data. The second simulated genome we benchmarked was one we produced from manually inserting ten HK2 LTRs into hg19 chromosome 20 across the genome (Supplementary Table S4, p. 153). Using the MASON read simulator, we created an Illumina paired-end WGS dataset with 50×

coverage and reads of 101-bp length. We mapped these reads to the hg19 reference using BWA. The resulting genome was benchmarked with RetroSeq, VirusSeq, and STEAK. Tangram runs on MOSAIK aligned genomes and did not accept any of our BAMs produced by other aligners in spite of a program that was released to add the necessary ZA tags (tangram_bam). Furthermore, Tangram is currently unmaintained and unsupported.

RetroSeq, Tangram, VirusSeq, and STEAK were all run with matching parameters to compare sensitivity and specificity on a simulated chromosome 20 with artificial HK2 insertions.

For the HIV-simulated dataset, we based it on the human genome reference (hg19) chromosome 1. Twenty full-length HIV integrations were randomly introduced into chromosome 1 using ViralFusionSeq viral insertion simulator (Li *et al.*, 2013). The full-length HIV insertion was taken from the HIV1/LAV reference (Accession number: K03455.1). We created simulated Illumina paired-end reads with the ART next generation sequencing read simulator (Huang *et al.*, 2012) with default error model and indicating 50× coverage and 100-bp reads. The reads were then aligned to hg19 using BWA MEM.

2.2.5 Target-enrichment dataset

We also tested STEAK on a sample that was prepared through targeted enrichment of the ends of HK2 LTR. Briefly, DNA was extracted from control Novagen™ human genomic DNA. Genomic regions of interest were selected using a biotin-streptavidine-based bead capture with DNA bait probes. In this case, target-specific baits used came from the beginning and end of HK2 LTR (K113): ~360-bp from each end. Five overlapping probes were used for each end; each probe was 120-bp in length. Single-stranded oligonucleotides with a common linker flanked by target-specific sequences anneal to the sequences of interest and capture them (Gnirke *et al.*, 2009; Mamanova *et al.*, 2010). After target enrichment hybridisation, the sample was sequenced using the Illumina MiSeq platform producing PE 300-bp paired-end reads.

2.2.6 Availability and requirements

STEAK relies on boost-libraries, OpenMP, gcc, python, and BEDtools. SAMtools or biobambam2 can be used to decompress BAM files. It has purposely been designed to use as little dependencies as possible for negligible installation troubleshooting. Trimmed reads can be processed with an aligner of choice we recommend a sensitive

mapper such as Novoalign (Hercus, 2009), BWA MEM (Li & Durbin, 2011) or Stampy (Lunter & Goodson, 2011).

2.2.7 Data availability

STEAK is publicly available through Github (<http://github.com/applevir/STEAK>) with an accompanying test dataset and instruction manual. Details to create the simulated genomes of 50× coverage can be found in the Methods section. The 5× coverage genome can be found on Tangram's GitHub repository. TCGA WGS data were retrieved through the Genome Data Portal (<https://gdc-portal.nci.nih.gov/legacy-archive/search/f>) after applying for access to controlled-access data through dbGAP authorisation. The CEU pedigree WGS data were retrieved through the 1000 Genomes beta Data Portal (<http://www.internationalgenome.org/data-portal/sample>) under Illumina Platinum pedigree data collection. The target enrichment library sequences data from this study have been submitted to the ENA short read archive under the accession number PRJEB21477; Sample accession numbers are ERS1796985-6 inclusive.

2.3 Results

2.3.1 STEAK algorithmic overview

Our program firstly detects reads from a HTS library that contain fragments of a minimum similarity to a given short sequence (which can be the edge of a TE or a virus, hereafter called chimeric reads) and subsequently removes this fragment to produce a library with reads that only contain host reference flank (hereafter called trimmed reads). If the HTS library contains paired-end reads, then the respective paired-end mates of the trimmed reads are retrieved (Figure 2.1). Novel integrations are often sufficiently random to produce unique chimeric reads however, in scenarios where the chimeric reads are not unique, these respective paired-end mates can be used to further support a novel integration. Users can then identify the location of an integration either from previous mapping information or by mapping trimmed reads to the host reference. Trimmed reads can be mapped either as single-end reads or with their respective mates. The latter, where trimmed reads and their mates are mapped as pairs, is referred to as guided detection and can be used to increase the possibility for uniquely mapping the trimmed read. For example, if the trimmed read comes from a highly repetitive region the mapping of the trimmed read alone will not be unique, but the combination of the mate and the trimmed read could provide a unique mapping solution. The outputs can then be used for custom downstream analyses most importantly for the reconstruction of preintegration sites, which is an excellent bioinformatic alternative

to wet-lab verification of novel integrations. In silico verification is of paramount importance as whole genomes are becoming increasingly available while their original DNA samples are either not easily accessible or too valuable to be used for multiple PCR-guided verifications.

2.3.2 Input data

STEAK requires HTS data and a reference sequence that is expected to contain the edges of the mobile element (e.g. beginning and end of an LTR in the case of HERVs). Input HTS data can be either mapped (SAM) (Li *et al.*, 2009) or raw (FASTQ). Compressed SAM, also known as BAM files, can be input for STEAK using tools such as SAMtools or biobambam2 (Tischler & Leonard, 2014). TE and retroviral reference files must be in FASTA format. STEAK performs on both single-end (SE) and paired-end (PE) libraries and on a variety of HTS sequencing strategies. All input PE data should be collated by name for respective mate retrieval.

2.3.3 Chimeric read detection and trimming

The chimeric read detection and trimming phase of STEAK is a multistep process to identify and process reads, which contain bits of the retroviral or TE reference. STEAK takes the edges of the TE or retrovirus reference and creates reverse complements, producing four baits of a given length. With these baits, it looks for similarities by first aligning each read against the TE/virus bait with the Smith-Waterman algorithm (Zhao *et al.*, 2013). Parameters can be modified to alter the length of minimum bait match sought for in a read as well as the per cent identity between reference and chimeric read. For example, current default parameters are of a 15-bp length bait and a 95% identity however, to search for HK2 integrations of five million years or younger, we used a bait length of 20-bp and a per cent identity of 90 to allow the detection of integrations which have mutated over time.

If the HTS data provided is already mapped (BAM/SAM), filtering based on the percentage of matches within a CIGAR value is also an available option for detecting chimeric reads. Filtering by CIGAR value is useful for non-reference TE discovery because reads pertaining to novel integrations are expected to have less than perfect mapping to the host. With this filtering, STEAK searches for any reads matching less than perfect with the original host reference (e.g. 94M7S for a 101-bp read or 6S280M15S for a 301-bp read), which can speed up processing significantly.

To annotate reference TE integrations, no CIGAR filtering is needed because reference integrations are expected to have reads mapping with perfect matches. In this case,

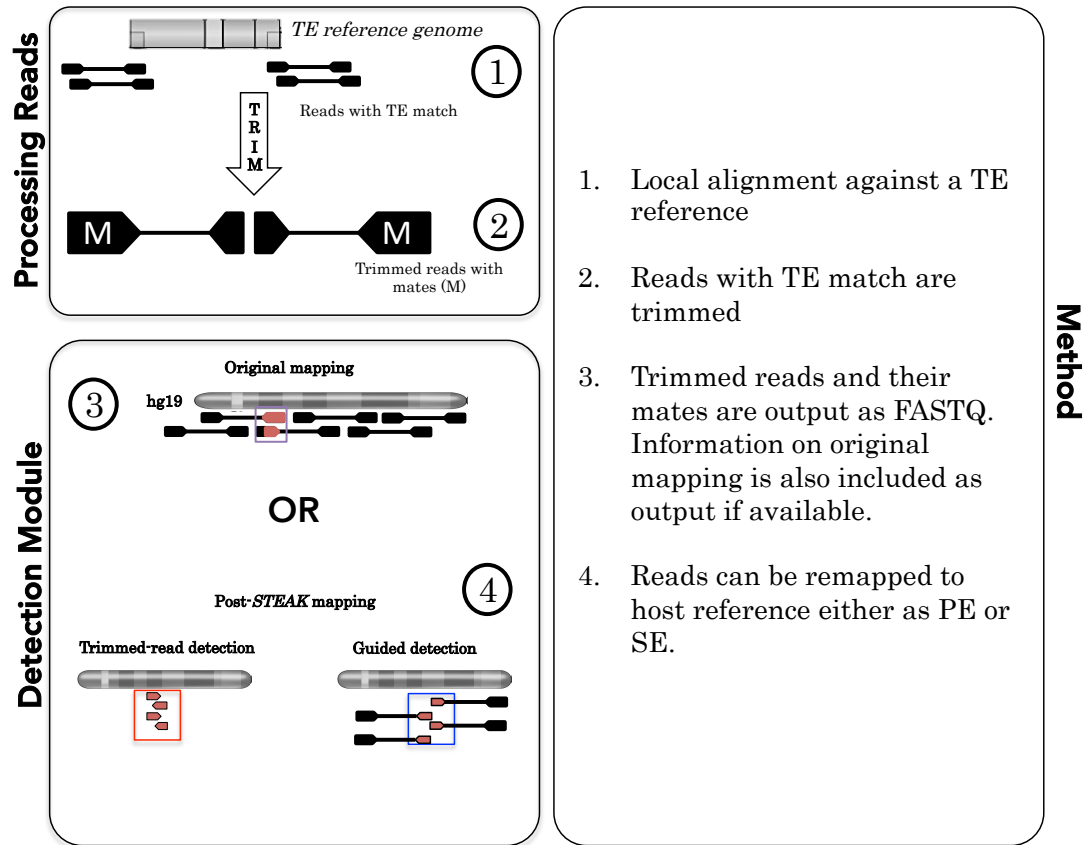


Fig. 2.1.: Workflow of STEAK. Processing data: All reads are locally aligned using the Smith Waterman algorithm and allowing mismatches when mapping reads against a TE reference (5' and 3'-ends and respective reverse complements). Reads that match with the TE are trimmed of the matching portion. Information on the trimmed reads and their mates, such as the original mapping positions, MAPQ, and sequence qualities, are kept in STEAK outputs. Detection Module: Trimmed reads can be remapped to the human reference either as single-end (trimmed read detection) or paired-end reads (guided detection).

STEAK will automatically search for chimeric reads that support integrations both present and not present in the host reference.

2.3.4 TE and retrovirus detection module

While there are a number of transposable element discovery pipelines (Ewing, 2015), STEAK is capable of finding reads supporting both reference and non-reference TE and retroviral integrations. STEAK is a parallelised software that can function as a standalone or coupled with other tools for custom downstream analyses.

If the initially provided HTS data are already mapped, STEAK is able to retrieve information of the original mapping for the chimeric reads it detects. This on its own can often provide supporting reads for integrations, which is already seen with systems like MELT (Gardner *et al.*, 2017) and RetroSeq (Keane *et al.*, 2013) (Table 2.1). But, in

addition to this, the STEAK algorithm outputs trimmed reads which can be exploited to detect both reference and non-reference integrations. The advantage of trimming chimeric reads is that it allows for re-mapping of the host flanks to the original host genome forming clusters. These clusters of host-trimmed reads can then indicate the site of an integration. Furthermore, by providing the mates of trimmed reads, it is possible to perform guided detection (Figure 2.1) where the mate can further support the proper mapping of a trimmed read.

The outputs from STEAK's operations include (1) host trimmed reads and respective mates in FASTQ format, (2) the respective TE or retrovirus match in FASTQ format, and (3) a tab-delimited file providing information on the chimeric reads detected such as length of match, per cent identity, and previous mapping coordinates.

To detect these integration sites, we aligned host trimmed reads using Novoalign (Mager & Freeman, 1995). We chose Novoalign, like Marchi *et al.* (2013), because it is an accurate aligner particularly when dealing with single-end reads as it uses the Needleman-Wunsch algorithm with affine gap penalties when scoring mapping locations. Single-end mapping was performed with default parameters. Paired-end alignment parameters were specified as end-to-end mapping with no soft clipping. Remappings were done using the host reference genomes originally used.

Our choice of downstream analyses consisted of using a combination of BEDtools (Quinlan & Hall, 2010) and command line utilities, such as AWK and grep. To detect integration sites within the host reference, we provided a TE annotation file from RepeatMasker where the coordinates and the names of the known TE integrations are supplied. For known non-reference integrations, a TE annotation file compiled from the known literature is most appropriate. To mark presence or absence of known integrations, we compiled a list (Supplementary Table S3, p. 148) made from RepeatMasker annotations of HK2, Subramanian *et al.* (2011), Marchi *et al.* (2013), Lee *et al.* (2012), and Wildschutte *et al.* (2016). For other transposable elements, users can provide a BED file of known reference TE insertions or use the RepeatMasker annotations (Tarailo-Graovac & Chen, 2009).

For novel integrations, a list of both known reference and non-reference integrations should be used to ascertain that a detected *de novo* integration is in fact a novel. Additionally, we filtered out other transposable elements that shared sequence similarity with HK2 LTR, such as Sine/VNTR/Alu (SVAs). We excluded clusters matching the non-HK2 LTR part of an SVA or which were in close proximity (within 1,000-bp) of a known HK2 or SVA locus (regions annotated in RepeatMasker). We considered a novel integration discovery when five reads or more were found clustered within a range of 10 kb. This large range is to allow for longer than expected target site duplications

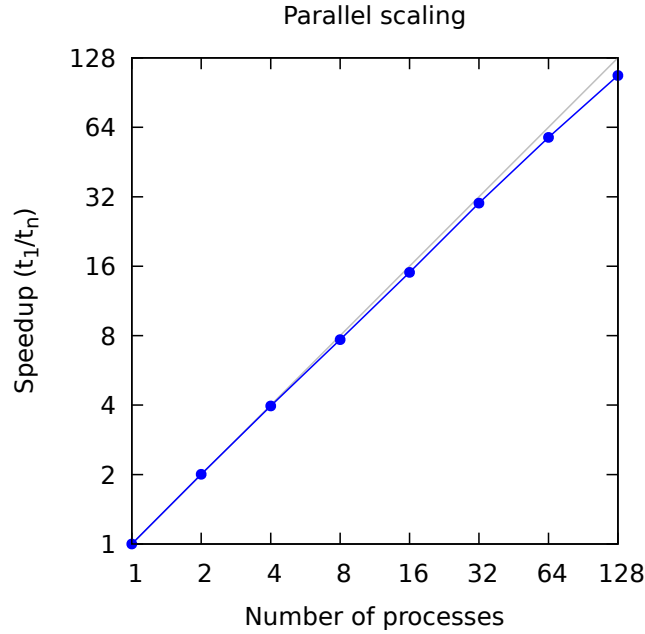


Fig. 2.2.: Parallelisation of STEAK processing. A $50\times$ coverage simulation of chromosome 1 was processed using our MPI-based software. The speedup as a function of the number of cores shows that the program scales well because few concurrency issues affect it. t_1 is the time of the problem computing at a single processor whereas t_n is the average time of the problem computing at one processor where N processors are run.

(Kahyo *et al.*, 2017). For non-endogenised elements, no annotation file is needed for integration discovery.

2.3.5 Software specifications and parallelisation

Detected candidate loci are output in two BED files: one for novel integrations and another for detection of known integrations. The BED files give coordinates of the region, the number of trimmed reads found for that locus, and if within another repetitive element. For target site duplication (TSD) identification, we used BreakAlign (Marchi *et al.*, 2013) and Geneious (Kearse *et al.*, 2012) (Supplementary Fig. S1, p. 155) was used for preintegration reconstruction which is an *in silico* alternative of verifying an integration without wet-lab validation.

STEAK has been programmed in C++ and been designed to be massively parallel, meaning that it is not limited to a single computer. This enables STEAK to run efficiently on large high-performance computing clusters and to process quickly high-coverage genomes (Figure 2.2). The input file is split in different parts that are each handled by a different process, using MPI (Figure 2.3A). The software is afflicted by very few concurrency issues and, as a consequence, should scale well until the limits arising from the file system are reached. Unfortunately, this parallelisation strategy is not

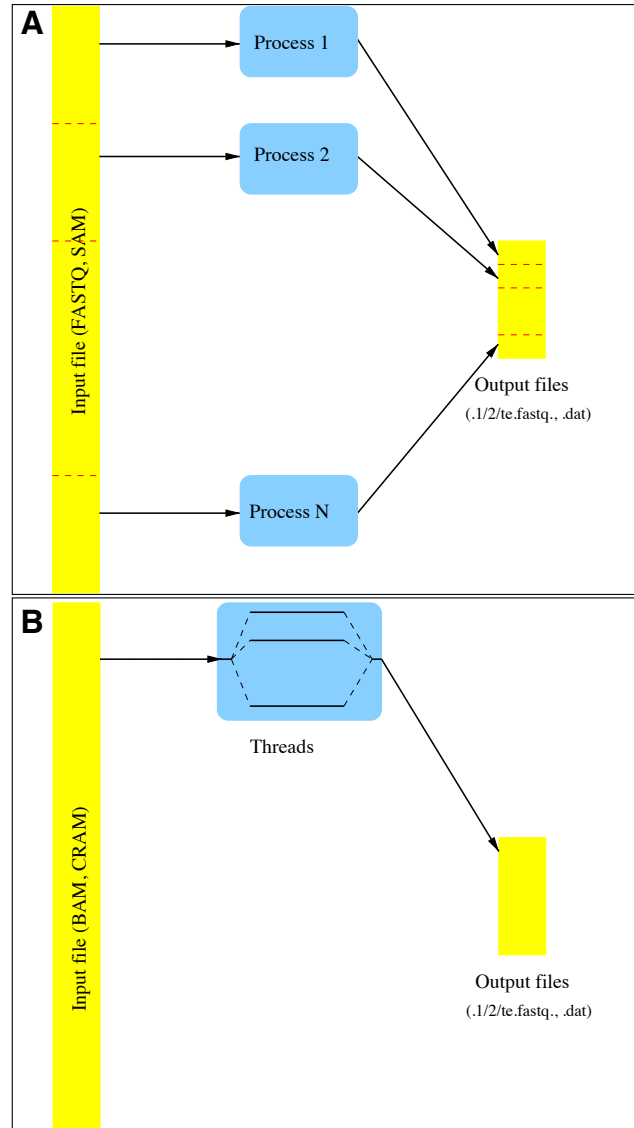


Fig. 2.3.: Processing in STEAK. (A) The software runs as several processes that read different parts of the input SAM or FASTQ file. (B) In the case of a BAM/CRAM file, because the data is read from a single point, a unique process is executed but it spawns several threads, one that accumulates the data in a buffer and the others that process them. TEs are specified in a FASTA file. The trimmed sequence is written in the first FASTQ file. The original sequence of the other read (or the trimmed sequence if there is a match as well) is placed in the second FASTQ file.

possible when we endeavour to process a BAM file because, in that case, the data are accessed sequentially, through a single point. This is because we are processing the output of the decompression fed through a pipe, i.e. the redirection of a standard output, and therefore, the whole file is not accessible to be read from several distinct locations. To cater to that possibility, we have resorted to multithreading (Figure 2.3B). A thread reads the input file and accumulates the reads in a circular buffer while several other threads process those reads in parallel. This leads to a noticeable performance improvement on fast file systems.

2.3.6 Overview of features and comparison to other TE and virus software

Our program works on re-sequencing projects and therefore additional downstream analysis requires good quality, well-annotated host reference genomes in addition to reference TE or retrovirus sequences (i.e. the TE sequence expected to be at an integration site). It is dependent on existing mapping software; for example, in order to identify potential retroviral integrations in a human genome, we would need to provide the human reference genome as well as the edges of the suspected retroviral LTRs. While the HTS data need not be mapped initially, another mapping program is required to detect integration sites (see Availability and Requirements). We pinpoint four important features of STEAK that to the best of our knowledge are combined in a package for the first time: firstly, it allows for detection of the TE (or viral) integration even if there has been deterioration of the sequence through time; secondly, it locates the absence of TE sequences that exist in the provided reference genome; thirdly, it facilitates a vast number of downstream analyses (i.e. the reconstruction of novel integrations); and fourthly, it successfully works on a variety of sequencing projects including target enrichment. Several other features of STEAK in comparison with other existing software are provided in Table 2.1. While we are aware of a number of other software that exist for transposable element (Ewing, 2015) and virus detection, we have chosen those which have comparable features, popular usage, and are currently competitive (see Supplementary Note S1, p. 142 and Supplementary Table S1, p. 143).

2.3.7 Evaluation of STEAK

For our benchmarking, we firstly tested the ability of STEAK to identify HK2 integrations in human full-genome re-sequencing HTS projects. We set bait length parameters to be 20-bp of the beginning and end of the LTR (both strands) for mining out chimeric reads (K113, Accession Number: NC_022518.1). A match between this 20-bp bait and each read was searched by means of the Smith-Waterman algorithm (Zhao *et al.*, 2013) which allows for indels and substitutions between the TE/virus reference and the

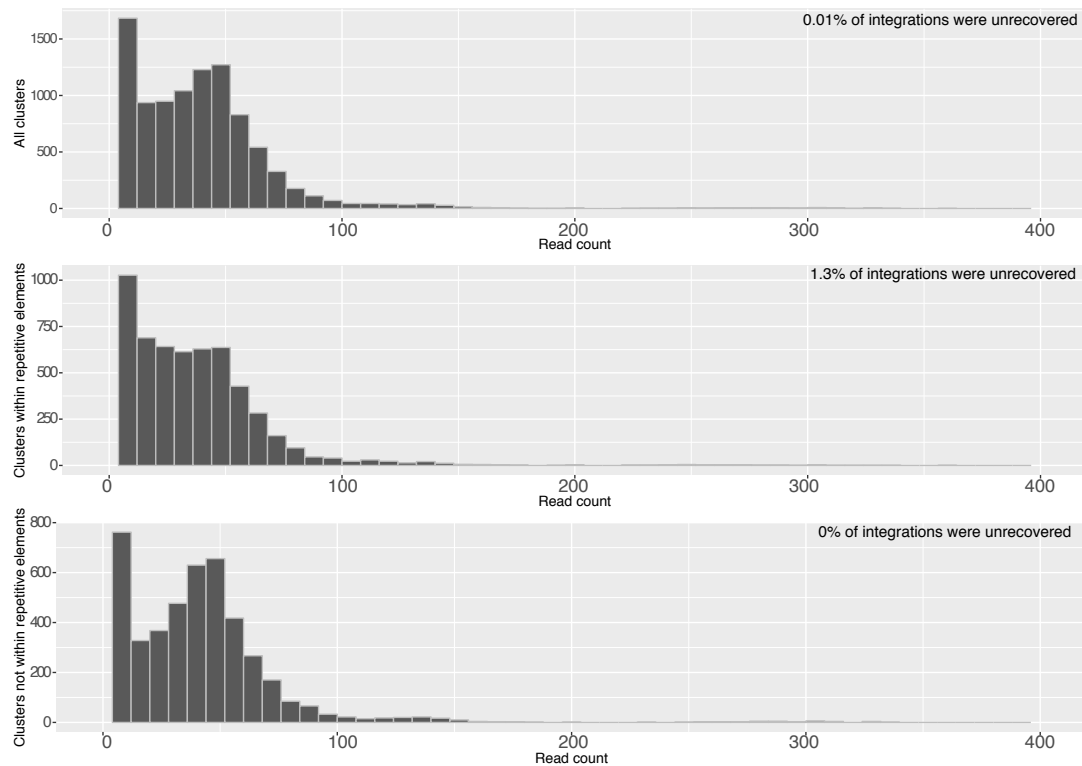


Fig. 2.4.: *Distributions for cluster read depth of known HK2 integrations.* Top: Cluster read depth for all known integrations. Middle: Cluster read depth for known integrations within repetitive elements. Bottom: Cluster read depth for known integrations that are not within repetitive elements.

read. This local alignment filtering only permitted reads through a certain threshold, allowing a limited number of mutations between the LTR reference and a read when searching for a 20-bp match (e.g. 90% similarity = up to two mismatches). Reads that passed filtering were trimmed of the LTR matching sequence (Figure 2.1). We only kept trimmed reads with a minimum length of 20-bp. Trimmed read length and TE match lengths are adjustable parameters in STEAK.

We initially tested STEAK on seventy WGS from The Cancer Genome Atlas (Supplementary Table S2, p. 144) to find presence and absence of already characterised HK2 integrations. We compiled a list of HK2 proviruses and solo LTRs that are five million years old or younger and can therefore be polymorphic in the human population (Supplementary Table S3, p. 148). From this list of 183 HK2 integrations, we used the 133 known fixed HK2 integrations to observe the depth of reads for present HK2 integrations amongst the 70 samples. Based on the cluster of trimmed reads found for these known integrations, we determined an appropriate threshold to consider a potential novel integration (Figure 2.4).

Our screening showed that some known fixed HK2 integrations were more difficult to recover than others. For example, integrations that are within other repetitive

Tab. 2.2.: Twenty simulated HIV integrations into human chromosome 1.

Human reference (hg19)		Original mapping	Post-trimming	
Chromosome	Position	Supporting reads	Guided detection	Trimmed read detection
1	4179520	35	100	64
1	10331435	34	113	70
1	16830086	30	96	57
1	18869777	31	112	67
1	20389049	36	76	45
1	54327146	34	93	63
1	57730318	34	101	70
1	99180019	24	92	53
1	116586277	31	85	54
1	144993094	16	18*	0*
1	149062299	36	99	63
1	165127302	43	109	71
1	170764099	15	69	44
1	188462855	43	121	85
1	191791184	25	80	53
1	197518001	34	88	55
1	213559631	26	90	52
1	219498833	37	97	63
1	223662699	38	95	61
1	231971371	26	78	44

* Where trimmed reads alone could not detect the integration.

elements were less likely to be picked up by our program, 1.3% of integrations within repetitive elements were unrecovered (Figure 2.4). Integrations with flanking repetitive regions may require lowering the tolerance below the 90% identity threshold we used or multi-location mapping can potentially resolve ambiguous trimmed read mappings. To account for these difficult regions and to increase our sensitivity, we only considered clusters with a minimum of five reads as potential novel candidates. However, for already known and characterised HK2 integrations in the literature (both reference and non-reference), we accepted single reads as evidence of presence. In spite of the difficulty that TEs within repetitive regions may pose, only one of the seventy-seven HK2 integrations within a repetitive element was unrecovered by STEAK, while all the other 132 integrations were recovered. Other regions that pose difficulties in recovering integrations include pseudogenes or genes with multiple copies in the genome (Supplementary Note S2, p. 154 and Supplementary Fig. S1, p. 155).

2.3.8 Performance evaluation on simulated data

We evaluated STEAK's ability to detect both TEs and retroviruses with a series of computational experiments. The first sets of experiments were done using Tangram's original Alu simulations into chromosome 20 and another custom simulation where HK2 integrations were inserted into chromosome 20 for benchmarking purposes (Sup-

Tab. 2.3.: The sequencing samples analysed in benchmarking.

Sample	Dataset	Coverage
TCGA-A6-2681-10A-01D-2188-10 (COAD)	TCGA: Colon adenocarcinoma	50×
TCGA-HC-7233-10A-01D-2115-08 (PRAD)	TCGA: Prostate adenocarcinoma	50×
TCGA-NJ-A4YQ-10A-01D-A46J-10 (LUAD)	TCGA: Lung adenocarcinoma	50×
TCGA-BW-A5NQ-10A-01D-A27I-10 (LIHC)	TCGA: Liver hepatocellular carcinoma	45×
NA12878	1K Genomes: CEU pedigree (Offspring)	50×
NA12891	1K Genomes: CEU pedigree (Father)	50×
NA12892	1K Genomes: CEU pedigree (Mother)	50×
HK2_Enrich01	NA	500×
HK2_Enrich02	NA	500×

plementary Table S4, p. 153). This provided insight into the positive predictive value (PPV), or the probability that a suggested integration was true, for STEAK, Tangram, and RetroSeq. When using only single-end trimmed read mapping, STEAK performed less sensitively but with higher PPV than the two other software (Supplementary Fig. S2A and B, p. 157).

For further evaluation of STEAK's ability to detect integrations, in particular retroviruses, we created a simulation based on hg19 chromosome 1 with 20 full-length HIV integrations (Figure 2.5) and with simulated reads that match realistic Illumina sequencing errors (~ 0.1) (Rands *et al.*, 2014). With this simulation (Methods 2.2), we evaluated the different forms of detection that STEAK has to offer: using original alignment information, using trimmed read detection, and using guided detection. Guided detection performs the most sensitively in its ability to recover all simulated integrations with the most amount of supporting reads per integration within a 100-bp window from the original simulated insertion site (Table 2.2). These results suggest that while trimmed reads can provide specific integration detection, if mapped alone they lose out on the sensitivity that being paired offers in mapping. Similarly, VFS (Table 2.1) was also able to detect all simulated integrations by making use of both chimeric reads and discordant pairs although with a substantially slower pipeline (Supplementary Note S1, p. 142).

2.3.9 Performance on whole genome and target enrichment data

We screened samples of whole genome sequence datasets provided by The Cancer Genomes Atlas Project (TCGA) (Weinstein *et al.*, 2013), the 1000 Genomes Project, and a target-enrichment project (Methods 2.2). The CEU pedigree, four patient genomes, and the target enrichment sample are listed in Table 2.3. These samples were chosen for their high coverage and representativeness of different projects that are publicly available (e.g. 1000 Genomes Project) as well as clinical (e.g. TCGA). The CEU pedigree

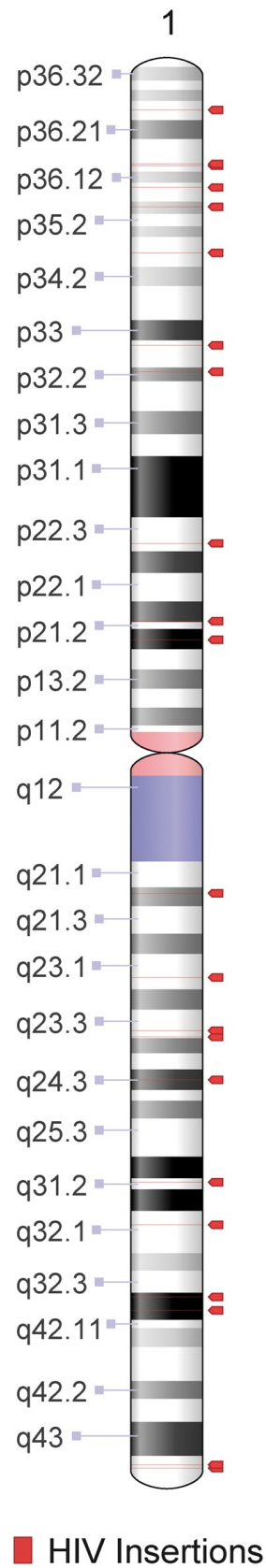


Fig. 2.5.: *Distribution of twenty simulated HIV integrations within human reference (hg19) in chromosome 1. Respective genomic coordinates can be found in Table 2.2.*

was chosen because it is one of the best-sequenced pedigrees, which is also publicly available making it ideal for benchmarking purposes.

With STEAK, we are able to mark the presence and absence of both reference and non-reference integrations. For comparison purposes, we benchmarked RetroSeq and MELT using non-reference integrations. For benchmarking presence and absence of reference integrations, we compared our results to MELT.

We observed that RetroSeq and STEAK often exceed the sensitivity of MELT for non-reference integrations (Figure 2.6A) and that STEAK performs more sensitively than MELT in detecting known polymorphic HK2 integrations regardless of whether they are reference or non-reference (Figure 2.6B). We also intended to compare our reference integration results to tlex2 results, but tlex2 is unable to handle high-coverage genomes (Supplementary Note S1, p. 142).

2.4 Discussion

Crucially, STEAK clearly demonstrates its ability to handle long fragment-sized libraries, such as target-enrichment data, and significantly outperforms MELT and RetroSeq in detecting polymorphic integrations both in and not in the reference genome (Figure 2.7). In the case of MELT's deletion genotyping module, it marks 137 reference integrations, of which a great majority tend to be fixed within the population, as absent in the target enrichment data.

The driving motive behind developing STEAK was to detect polymorphic endogenous retrovirus integrations in HTS data with high coverage. The overarching difference of STEAK, compared to previous algorithms that were published (Marchi *et al.*, 2013), was that it allowed for the potential evolution of the transposable element or virus by tolerating for a controllable number of mutations to be present in the reference sequence. This latter part is computationally intensive, thus we accelerated the process by massively parallelising our software.

STEAK has been designed to include functionality features that are largely missing from other similar available software. RetroSeq, MELT, and TEA rely on paired-end data and work on discordant or unmapped mates (Li *et al.*, 2009). Those algorithms cannot work with single-end reads, longer-read libraries where pairs partially overlap, and do not provide outputs for users to continue their own downstream analysis. While MELT, RetroSeq, and TEA provide the coordinates for TSD breakpoint intervals, they do not provide outputs for reconstructing pre-integration sites or to further characterise the novel integrations found, all of which can be particularly useful when processing

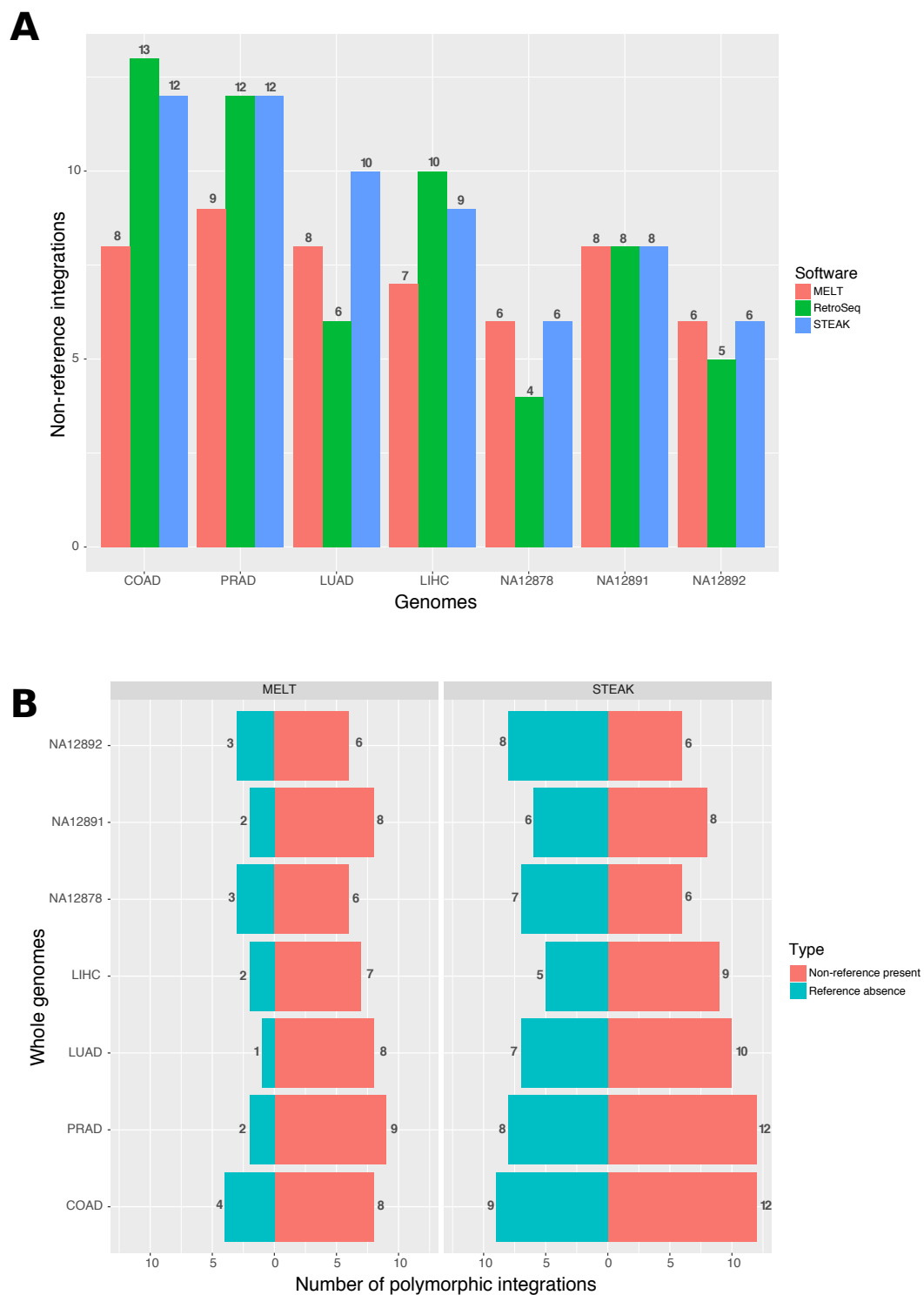


Fig. 2.6.: *Comparative performance of HK2 detection in whole genome sequencing.* (A) Bar graph displays known non-reference integrations detected in WGS projects by each respective system. (B) Pyramid plot depicts number of polymorphic integrations detected per genome—number of present non-reference and number of absent reference integrations. Abbreviations correspond to samples in Table 2.3.

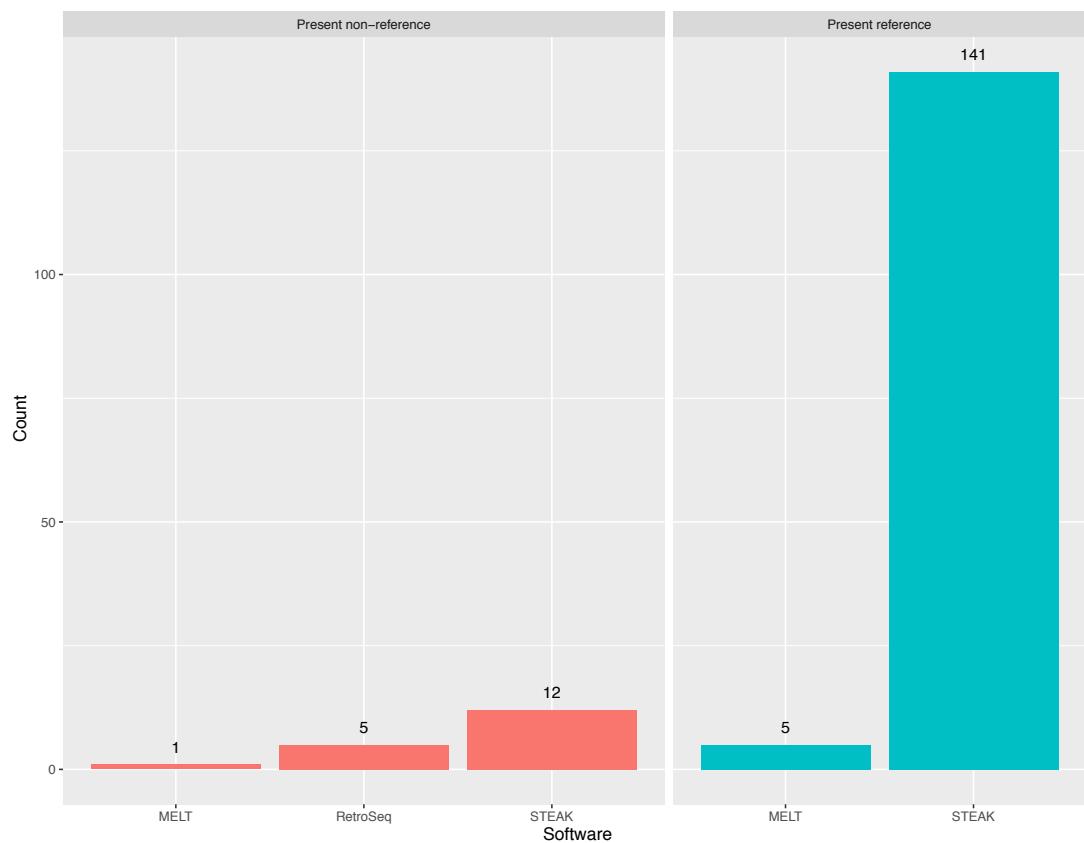


Fig. 2.7.: *Detection performance in target enrichment data.* The left facet depicts presence of non-reference integrations detected by each respective system. The right facet depicts the marking of presence of reference integrations. MELT results detect 5 reference presences and 137 absences.

whole genome sequences without access to original DNA samples. This is highlighted in clinical datasets, like TCGA, which often include data from individuals that have been sequenced with 30× coverage but have restricted access to samples. In such cases, the ability to recover supporting chimeric reads for TEs or retroviral integrations in high-coverage genomes is valuable. One example would be using WGS to observe the changes in the TE profile of an individual, such as an abundance of a TE in tumour tissue when compared to the germline (Criscione *et al.*, 2014). On the other hand, screening deep sequencing target enrichment experiments can be equally as crucial for clinical research: for example, target enrichment is a specialised alternative to WGS, particularly if there is a need to sequence multiple samples to observe a specific active retrotransposon or retrovirus. In this regard, STEAK was the only software capable of sensitively retrieving integrations from both WGS and target enrichment sequencing projects (Figures 2.6 and 2.7).

Our program will efficiently process both mapped (SAM/BAM) and unmapped (FASTQ) HTS data, paired and single-end sequencing projects, and whole genome sequencing as well as target enrichment strategies. STEAK is also the only program that has an integrated approach for users to detect reads that support TE and retroviral integrations: using original mapping information, trimmed read mapping and guided detection. Moreover, it is the only software flexible enough to provide users with outputs to move forward with custom downstream analysis.

How does STEAK compare with other similar software with respect to performance? We compared STEAK with MELT, RetroSeq, and Tangram on paired-end datasets (which in principle can be handled by all of them). RetroSeq and Tangram fared well in sensitively detecting TEs within low-coverage genomes (Rishishwar *et al.*, 2017a), but we observed an increase of false-positives when used for discovery in deep-coverage WGS (Supplementary Fig. S2A and B, p. 157). Such programs can be used with low-coverage WGS but the verification of the breakpoint would need to go through a wet-lab approach. STEAK performs better than other TE discovery software when recovering specific integrations with high confidence, which is the case with individual deep-coverage genomes. It is a program that can both specifically and sensitively mark presence or absence of reference and non-reference TEs with its adjustable parameters. We observed that certain HK2 proviruses were more difficult to recover by STEAK when integrated within other mobile elements. When such proviruses within repetitive elements are already described (i.e. no verification is required) we can increase the sensitivity of our downstream analyses by accepting the presence of the provirus even with a single chimeric read spanning the known junction. As demonstrated in the simulation (Table 2.2), the guided detection that STEAK facilitates allows rescuing of integrations where host trimmed reads do not map uniquely. However, when these proviruses are not catalogued an approach like RetroSeq would be more likely to recover them as potential candidates (Figure 2.6A); although these candidate integrations are

likely to need wet-lab verification, as the PPV of RetroSeq remains low (Supplementary Fig. S2B, p. 157).

STEAK stands out in its ability to identify more integrations within other forms of sequencing strategies, such as target-enrichment data, in comparison to already existing systems (Figure 2.7). Numerous sequencing platforms are making efforts to output longer reads which will inevitably provide longer reads flanking the integration junctions. STEAK was designed with this in mind and already proves to handle such sequencing strategies without any difficulties.

The limitations of STEAK in comparison to other software, like MELT and RetroSeq, lie in that it does not mark zygosity of an integration and that it does not calculate the exact breakpoint interval as this can be done separately in the downstream processing. In our analyses of non-reference HK2 integrations, STEAK provides flexible outputs to easily recur to software such as BreakAlign (Marchi *et al.*, 2013) and Geneious (Kearse *et al.*, 2012) for TSD detection and pre-integration site reconstruction (Supplementary Fig. S1, p. 155) in order to calculate breakpoint intervals.

We show that STEAK performs as well as or better than competitive systems available when detecting non-reference integrations (Figure 2.6B). As an algorithm and software, it is remarkable in its flexibility to accept a variety of HTS data and process high-coverage genomes in a parallelised manner where other systems struggle or even fail (Supplementary Note S1, p. 142). The purpose of STEAK is to fill a gap that exists in regards to detecting mobile elements (either virus or transposable elements) while also providing adjustability in its detection strategies and assisting users in their custom downstream analyses.

Statement of Authorship for joint/multi-authored papers for PGR thesis
To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (**only required where there isn't already a statement of contribution within the paper itself**).

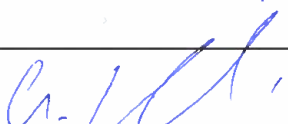
Title of Paper	STEAK: A specific tool for transposable elements and retrovirus detection in high-throughput sequencing data
Publication Status	☒ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Cindy G. Santander, Philippe Gambron, Emanuele Marchi, Timokratis Karamitros, Aris Katzourakis, Gkikas Magiorkinis, STEAK: A specific tool for transposable elements and retrovirus detection in high-throughput sequencing data, <i>Virus Evolution</i> , Volume 3, Issue 2, July 2017, vex023, https://doi.org/10.1093/ve/vex023

Student Confirmation

Student Name:	Cindy Gilda Santander		
Contribution to the Paper	I helped develop the concept for STEAK. I wrote the code the in collaboration with PG. I implemented the method on the data, benchmarked the other software, and performed all the analyses with guidance of GM. I wrote the original draft of the manuscript and revised accordingly using co-author's feedback. I created all the figures with guidance of GM.		
Signature		Date	25/04/2019

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Dr Cristian Capelli			
Supervisor comments <i>The statement by the candidate matches the declaration of contribution in the paper</i>			
Signature		Date	25/4/19

This completed form should be included in the thesis, at the end of the relevant chapter.

On the activity of HERV-K (HML-2)

3.1 Introduction

Capable of integrating themselves into a host's genome, retroviruses can have profound influences on the evolution of their hosts. The genus *Homo* and retroviruses have an intimate and reticulate evolutionary relationship dating back to millions of years ago. This aspect of retroviral biology is often eclipsed by their role as devastating disease vectors, such as HIV, responsible for the AIDS epidemic, and also as the oncogenic Human T cell lymphotropic virus type 1 (HTLV-1). However not all viruses are pathogenic and in some situations they have been advantageous for human evolution. Such is the case of the human endogenous retrovirus W family (HERV-W) which infected the germline of Old World monkeys 25 million years ago (MYA). This retroviral infection left behind an *env* gene, canonically known as *syncytin-1*, which has since been co-opted to play a crucial role for trophoblast fusion in placental formation (Mi *et al.*, 2000). Another example is a provirus from the ERV9 family present in only humans and great apes which transcribes *TP63* for the testis-specific expression of germ cell-associated protein, p63 which protects male germ cells from cancer (Beyer *et al.*, 2011).

The human genome has accumulated an abundance of similar retroviral elements from ancient infections where the invading retrovirus has reintegrated continuously as far as the germline (Bannert & Kurth, 2006) sometimes resulting in endogenisation (Chapter 1.2). ERVs can replicate both as transposable elements (TE) and viruses. As transposable elements, ERVs can copy into the host entirely through intercellular mechanisms as LTR retrotransposons (Havecker *et al.*, 2004). On the other hand, as viruses they can reinfect the germline as an exogenous retrovirus would and proliferate in somatic cells (Jaenisch, 1976; Tarlinton *et al.*, 2006). An ERV's capability to retrotranspose or reinfect relies on the integrity of the *env* gene which encodes the protein on the surface of the viral particle (viral envelope), responsible for cell entry (Kurth & Bannert, 2010). As explained in Chapter 1, host restriction factors can inactivate a retrovirus by damaging or deleting *env* which is crucial for extracellular mobility. However, as an LTR retrotransposon, an ERV requires only a minimum of *gag* and *pol* to be replication competent. On rare occasions (Ribet *et al.*, 2008), a retrovirus with a damaged *env* gene can regain its ability to reinfect by "hitch-hiking" the *env* of another infecting retrovirus, a mechanism referred to as "complementation"

(Hanafusa *et al.*, 1963). Of the well-characterised HK2 integrations, a major subset display a 292-bp deletion at the junction of *pol* and *env* which most likely renders them incapable of infection and extracellular transmission. This subset, also known as type I viruses in the HK2 family, are an example of HK2s which replicated as retrotransposons. This deletion which encodes the alternative splice product, *np9*, has been repeatedly exchanged across different HK2 subgroups through gene conversion, as well as through recombination (Macfarlane & Simmonds, 2004). More than 44% of human-specific HK2 integrations happen to be type I viruses yet the importance of the alternative protein (Bannert & Kurth, 2006; Subramanian *et al.*, 2011), Np9, remains unclear with no known physiological function in HK2 replication (Hohn *et al.*, 2013).

The relative abundance of ERVs within different hosts depends on the interplay of several factors which include, but are not limited to, viral life cycle (Magiorkinis *et al.*, 2015), host mode of fertilisation (Hayward *et al.*, 2014), and body size (Katzourakis *et al.*, 2014). Like cancer cells (Caulin & Maley, 2011), ERV activity is thought to be limited by the pathogenic costs that scales with host body size (Katzourakis *et al.*, 2014). While humans are less efficient in limiting the abundance of cancer cells (Rifkin *et al.*, 2017; Caulin & Maley, 2011), there is to date neither evidence of oncogenic ERVs in humans nor convincing studies presenting HERVs as pathogenic despite their established oncogenicity elsewhere in the animal kingdom (Tarlinton *et al.*, 2006; Arnaud *et al.*, 2008; Bentvelzen *et al.*, 1970).

ERV acquisition can have, mechanistically, three different effects on the host genome: 1) expression of the integrated virus or viral proteins, 2) host RNA transcription controlled by viral sequences, and 3) insertional mutagenesis as bulk DNA (Stoye, 2012; Johnson, 2019). For the recently integrated HERV-K (HML-2) family, referred here as HK2, the first mechanism has been shown across several studies where expression of viral genes and proteins has been observed in healthy and malignant tissues alike (Schmitt *et al.*, 2015; Lin *et al.*, 2018; Lower *et al.*, 1984; Büscher *et al.*, 2005). Some studies have even noted viral particles (Lower *et al.*, 1993; Hirschl *et al.*, 2007; Boller *et al.*, 1993; Grow *et al.*, 2015; Dewannieux *et al.*, 2006). The over-expression of HK2 viral genes is strongly correlated with germ cell tumours and melanomas and can therefore be used as a biomarker for these types of cancers (Lin *et al.*, 2018; Muster *et al.*, 2003). Furthermore, some studies suggest that HK2 in tumourigenic cells may regulate nearby genes through antisense transcription, specifically through LTR promoter activity (Patzke *et al.*, 2002; Montesio *et al.*, 2018). Nevertheless, it is important to note that this sort of ERV transcription appears to be dependent on the altered transcriptional environment induced by a cancer and not causal (Montesio *et al.*, 2018). HK2 integrations, however, as bulk DNA insertions, can also interrupt normal transcription of nearby genes without necessarily regulating transcription (Karamitros *et al.*, 2018) but whether they are deleterious in humans remains, at best, unproven.

To date, there is no convincing evidence of HK2 presently reinfecting or retrotransposing the human germline despite there existing HK2 integrations with intact open reading frames for *gag*, *pol*, and *env*. Among the youngest dated HK2 integrations, two are polymorphic in the human population dated $\sim 0.67 - 1.3$ MYA and $\sim 1.8 - 3.3$ MYA (Wildschutte *et al.*, 2016). They are both proviruses with full Open Reading Frames (ORFs), one of which is located in chromosome X (Xq21.33) and under investigation for potential infectivity (Wildschutte *et al.*, 2016). The Xq21.33 provirus, which is the youngest HK2 integration dated, also happens to be found in the Denisovan genome (Wildschutte *et al.*, 2016; Agoni *et al.*, 2012) suggesting that the integration of this provirus is most consistent with integrating before Denisovans and anatomically modern humans split from their most recent common ancestor (Reich *et al.*, 2010) (See Figure 1.6).

Previous studies using neutral population genetic models have suggested that HK2 was active—reintegrating in the germline—until 150,000-250,000 years ago (Marchi *et al.*, 2013; Jha *et al.*, 2011), still leaving open the possibility that HK2 could have died-out within the last $\sim 200,000$ years of human history. However, approaches based on population genetic modelling alone cannot definitely prove or reject very recent HK2 activity. In a study that estimated frequencies of polymorphic HK2 integrations in more than 2,400 individuals from the 1000 Genomes Project (Wildschutte *et al.*, 2016), no non-African specific HK2 integrations could be found. This particular finding suggests that HK2 activity must have ceased before modern humans dispersed outside of Africa, 50,000-60,000 years ago (Henn *et al.*, 2012).

In this Chapter I show that the frequencies of polymorphic HK2 integrations worldwide (Wildschutte *et al.*, 2016) (Figure 3.2) fall within the expected levels of diversity (Tishkoff *et al.*, 2009; Sudmant *et al.*, 2015) with no evidence of recent germline activity in non-Africans. Furthermore, using high coverage genomes from The Cancer Genome Atlas (TCGA), I find only one novel non-African-specific HK2 integration, which remains to be validated through wet-lab work. My results suggest that HK2 activity, whether through reinfection, retrotransposition or complementation, likely ceased more than 250,000 years ago. Together these results propose that if HK2 were to be a pathogenic burden and/or still active, the most appropriate individuals to search would be African populations which carry the youngest HK2 integrations.

3.2 Materials and Methods

3.2.1 Estimating the HK2 integration rate for the last 5 MY

To calculate the most recent HK2 integration rate, I used Wildschutte *et al.* (2016) data from the 1000 Genomes Project and the HGDP genomes: allele count of polymorphic HK2 integrations for a total of 2,453 individuals. They observed 42 polymorphic integrations. These segregating sites (polymorphic/unfixed integrations) are treated as point mutations in the rate calculation by means of the Watterson estimator (Watterson, 1975).

3.2.2 Estimating HK2 historical (30 MYA – 5 MYA) integration rates

Calculating the integration rate of HK2 from 5 MYA to 30 MYA differed from the calculation for the present-day integration in that we calculated the rate by dividing the number of insertions that were dated accordingly (Subramanian *et al.*, 2011) by the number of generations in 2 million year windows and assuming an average generation time of 20 years, a generation time relevant for both early modern humans as well as chimpanzees and gorillas (Teleki *et al.*, 1976; Plagnol & Wall, 2006; Matsumura & Forster, 2008). These insertions encompass loci that have integrated before the divergence of human and chimpanzee therefore including integrations that went into fixation in humans. The most recent common ancestor lived approximately 4.5 MYA, which has been used to calculate a neutral mutation rate of 0.24 to 0.5 substitutions per 1000 bp per million years for the common ancestor (Jha *et al.*, 2009; Johnson & Coffin, 1999). The age of HK2 insertions allows us to confidently calculate the rate of integrations older than 5 MYA without the use of the Watterson estimator. This is because LTRs at the time of integration are identical. Integrations with base differences between the LTRs that give age estimates lower than the upper bound are difficult to date because neutral substitutions could have occurred anytime between today and 5 MYA.

3.2.3 Distribution for different polymorphic HK2 integration worldwide

Using Wildschutte *et al.* (2016) compiled data from the 1000 Genomes Project and the HGDP genomes, I calculated allele frequencies per global ancestry and mapped them to their corresponding geographic (ancestry-derived) region with the exception of CEU and ASW which were mapped to the US. For example, Gujarati Indians from Houston,

Texas (GIH) HK2 frequencies were mapped to India. Frequencies were calculated for soloLTRs, full-proviruses in the hg19 reference detected using only the LTR, and full-proviruses confirmed using a unique k-mer method that reconstructed part of the viral genes (*gag*, *pol*, *env*).

3.2.4 WGS from The Cancer Genome Atlas

I analysed 70 normal blood-derived samples from patients that were available at an average range of 45-50× coverage. These paired-end WGS (2 x 100 bp) were all sequenced using Illumina Genome Analyzer platform technology (Table A.1). All whole genomes downloaded from TCGA database were pre-aligned to the human reference, version hg19. Germline samples belonged to 36 females and 34 males. Ethnicity/race were self-reported (i.e. TCGA is an American cancer database). Only cancer types where more than one individual was downloaded were included in the analyses. However, all individual genomes were used to search for inR and nonR integrations. ASW, CEU, and CHS frequencies were used for comparisons based on self-reported ethnicities and race in the TCGA.

3.2.5 Marking presence/absence of polymorphic HK2 integrations

I ran STEAK using the parameters found in Supplementary Note S4, p. 158 on WGS from TCGA. I searched for the presence and absence of 147 HK2 integrations that are younger than 5 MY compiled from several different studies (Barbulescu *et al.*, 1999; Marchi *et al.*, 2013; Subramanian *et al.*, 2011; Wildschutte *et al.*, 2016; Hughes & Coffin, 2001; Agoni *et al.*, 2012; Turner *et al.*, 2001; Lee *et al.*, 2012; Belshaw *et al.*, 2005) using the LTR to mine out reads from the WGS. Reads which matched with 95% similarity to the LTR sequence provided were then used trimmed of the LTR leaving only unique host-flank. The trimmed reads were then mapped to the hg19 reference. These trimmed-reads were used to identify regions where an integration was present or absent.

3.3 Results

HK2 first integrated into the germline of Old World primates around 35 MYA (Bannert & Kurth, 2006) with some integrations holding an upper bound date of over 40 MY (Subramanian *et al.*, 2011) (Figure 3.1). The most recent reinfection of the germline took place after the divergence of human and chimpanzee lineages about 6 MYA (Bannert & Kurth, 2006). Today the human population shares some fixed HK2 integrations with chimpanzees and other primates, however a great deal of

integrations in the population are human-specific ($\sim 43\%$) (The Chimpanzee Sequencing and Analysis Consortium *et al.*, 2005). About 30% of human-specific HK2 integrations are insertionally polymorphic (Wildschutte *et al.*, 2016) with more than 80% of them in the form of solo LTRs rather than full-proviruses or simply absent from certain populations (Subramanian *et al.*, 2011; Wildschutte *et al.*, 2016; Bannert & Kurth, 2006). Polymorphic integrations are represented by the youngest HK2 insertions, primarily those younger than 5 MY (Subramanian *et al.*, 2011; Bannert & Kurth, 2006). Considering that previous studies have established that a neutral HERV would require about 40,000 generations (0.8-1 MY) to reach fixation in the human population ($N_e = 10,000$) (Hughes & Coffin, 2004), polymorphic HK2 integrations in the human population today are representative of putatively recent activity in the past one million years.

It still remains unclear when the HK2 family ceased germline activity and if this occurred suddenly or gradually. Based on the simulations of Marchi *et al.* (2013), if HK2 had been in a gradual "die-out" trajectory then we should observe a drop of the integration activity within the last 5 MY compared to the historical levels of activity before 5 MYA. In line with this, Wildschutte *et al.* (2016) showed that the highest frequency of HK2 integrations can be found in the San, an African population with the deepest divergence time among humans, beyond 260,000 years ago (Schlebusch *et al.*, 2012) (See Chapter 5). This study helped established three important key findings: 1) all HK2 integrations in non-African populations are also present in African populations and therefore precede the out of Africa bottleneck, 2) the San display the highest frequencies of HK2 integrations suggesting that most recent HK2 integrations took place prior to the divergence of the San from other human populations (260,000-350,000 years ago) (Schlebusch *et al.*, 2012), and 3) the youngest HK2 provirus found to date is 0.67-1.3 MY old and also shared with Denisovans, which is consistent with this insertion having occurred in the most recent common ancestor of humans and Denisovans. All together this would suggest that cessation of HK2 activity may have taken place somewhere between 0.26 and 1.3 MYA.

3.3.1 The integration rate of HK2 plunged in the last 5 million years

To identify when a HERV integrated into the germline one can compare the two LTRs of a specific integration. At the moment of integration, the LTRs are identical but accumulate mutations over time at the same rate as the host genome although ectopic recombination events, such as gene conversion, can compromise the accuracy of integration time estimates (Boeke & Stoye, 1997; Hughes & Coffin, 2005). But even if one did not account for recombination, given that HK2 LTRs are ~ 1000 -bp in length and the substitution rate of the human genome is ≈ 0.5 -bp per 1000-bp per million years,

this dating approach has a very low resolution for recent events; it cannot discriminate between integrations that happened today or within the last million years (Stringer & Andrews, 1988). While proviruses with identical LTRs have been identified, this finding is consistent with present day activity but also with a cessation of activity within the last one million years. Other approaches to dating HERV integrations have been attempted, including population genetics of known polymorphic integrations through PCR screening (Belshaw *et al.*, 2005), search for integrations absent from the human reference genome (Wildschutte *et al.*, 2016; Marchi *et al.*, 2013) or mutations observed in the LTRs (Jha *et al.*, 2011).

In order to understand the direction of HK2 activity in the last one million years, the number of polymorphic HK2 integrations in the population can be used to calculate an integration rate the same way segregating sites can be used to calculate a mutation rate under an infinite site model (Hartl & Clark, 1997; Hudson, 2002). Because HERV integrations within the last million years are difficult to date (Wildschutte *et al.*, 2016; Subramanian *et al.*, 2011; Jha *et al.*, 2011), using integrations that are estimated to date back to 5 MYA or less can be used to calculate a rate for the last 5 MY. This rate in comparison to the historical trajectory of HK2 integration rate can help us understand if HK2 was already in decline in the last 5 MY. Although other studies have done similar calculations (Marchi *et al.*, 2013; Belshaw *et al.*, 2005), they used limited cohorts which were not as ethnically diverse as the 1000 Genomes Project and the Human Genome Diversity Panel (HGDP) (Cann *et al.*, 2002; The 1000 Genomes Project Consortium, 2015). Their study yielded an integration rate that in the past five million years equated to 3.32×10^{-4} integrations per genome per generation which in a simulation model gave a cessation of HK2 activity around 150,000-250,000 years ago when compared to observed data (Marchi *et al.*, 2013).

I used data from Wildschutte *et al.* (2016), which include diverse populations in the 1000 Genomes Project and the HGDP, to calculate the most recent rate of integration based on the HK2 integrations that were shown to be polymorphic in that study ($K=42$), 26 integrations not present in the hg19 reference and 16 already present in the hg19 reference among 2,453 individuals. Here I use the Watterson estimator to obtain an integration rate similar to Marchi *et al.* (2013), according to the infinite sites model (Hartl & Clark, 1997; Hudson, 2002),

$$K = 4N_e\mu \sum_{i=1}^{n-1} \frac{1}{i} \quad (3.1)$$

where K is the number of polymorphic integrations, N_e is the effective population size, μ is the integration rate and i is the number of analysed haploid genomes $n=4,906$.

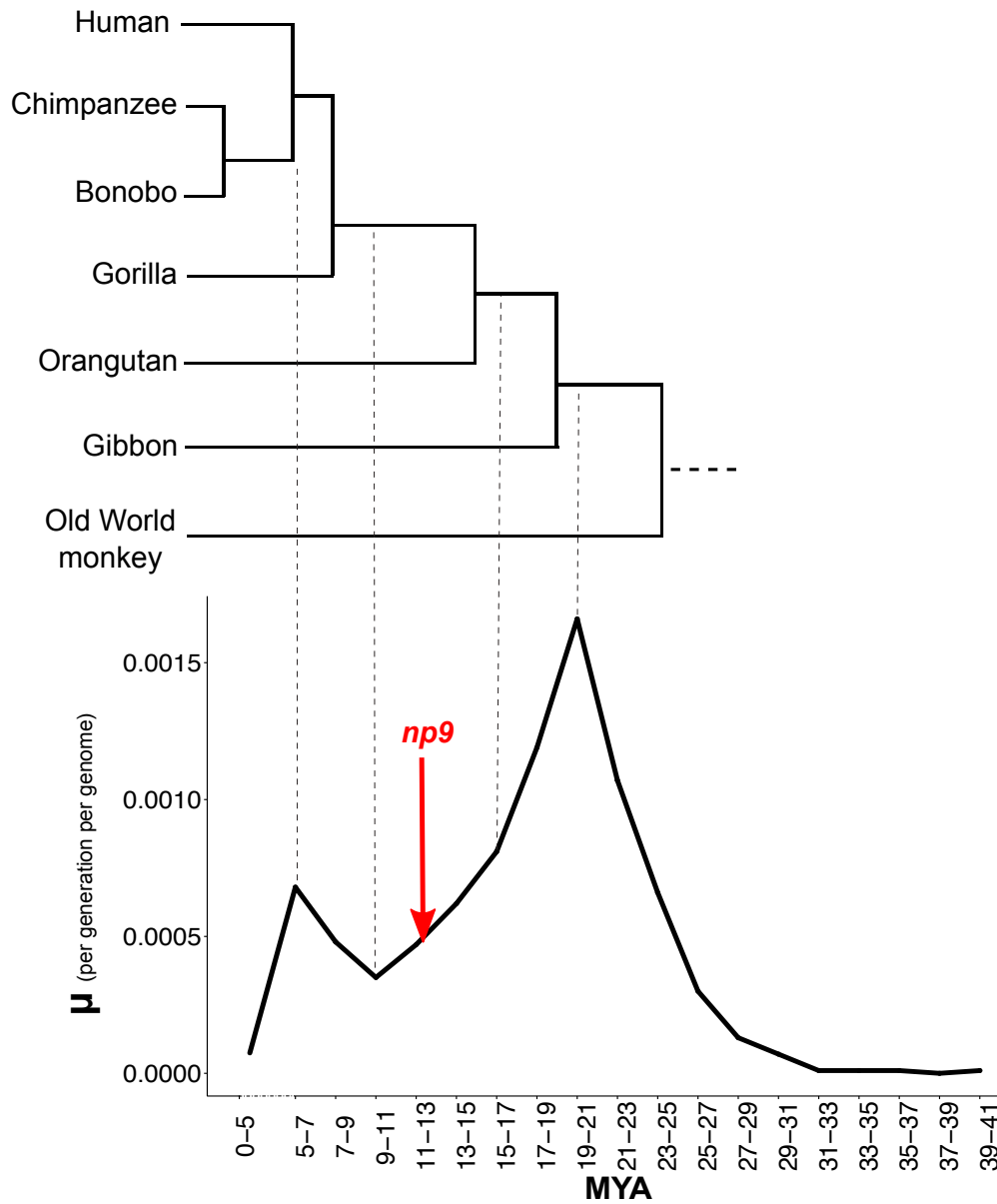


Fig. 3.1.: Trajectory of HERV-K (HML-2) integration rate (μ) across primate evolution. Upper panel depicts phylogeny of primate species where HK2 colonised at different time points in the last 41 million years. The dotted grey lines connect phylogeny to times of HK2 amplification. The appearance of alternative splicing product *np9* is pointed to in red. Dates and integration rates older than 5 MYA were calculated from Subramanian *et al.* (2011).

Based on the Watterson estimator (Eq.1) and the observed number of polymorphic integrations ($K=42$) in the analysed genomes from Wildschutte *et al.* (2016) and an N_e of 10,000, I estimate the most recent integration rate (μ) of HK2 to have been 1.16×10^{-4} per genome per generation before the cessation of germline activity.

Looking into the historical record of HK2 germline activity using the catalogue of dated HK2 integrations by Subramanian *et al.* (2011), we can observe that the integration rate peaked at 17-19 MYA and then declined until 9-11 MYA, but afterwards recovered until 5 MYA (Figure 3.1). That rate recovery coincides with the emergence of type-I proviruses with *np9* around 13 MYA when the copy number increased, on average, by 1.65×10^{-4} integrations per generation within 2 million years. However, it is clear that there was a plunge in activity from 5 MYA onward (Figure 3.1) suggesting a reduction in HK2's replication efficacy. Thus, the observation of 1.16×10^{-4} integrations per generation (Marchi *et al.*, 2013) is consistent with a decrease in accumulated HK2 integrations in the genus *Homo*.

The distribution of polymorphic HK2 across the globe varies on the type of integration present. Solo LTRs (See Chapter 1) which are thought to have a negligible impact on nearby genes (van de Lagemaat *et al.*, 2006; Medstrand *et al.*, 2002), are in higher frequency in Africa whereas HK2 frequencies begin to decrease moving up and outside of Africa (Figure 3.2A). This is consistent with the expected levels of diversity (Tishkoff *et al.*, 2009) and what has been found for other structural variants worldwide (Sudmant *et al.*, 2015; Rishishwar *et al.*, 2017b).

Wildschutte *et al.* (2016) use two methods of identifying HK2 proviral integrations: first using the LTR and second, reconstructing the provirus using unique k-mers. While this has revealed that in addition to there being polymorphisms in HK2 in terms of presence/absence, there are also polymorphic integrations that in some populations appear to be solo LTRs and in others the full provirus has been retained. Such is the case for the African-specific provirus 1q22, a type I provirus which appears to be less than 2 million years old but where the proviral genes seem absent in non-Africans (Wildschutte *et al.*, 2016). In particular, we can observe ancestry-based differences when detecting proviral integrations using LTRs (Figure 3.2B) versus detecting presence by constructing part of the viral genes (Figure 3.2C). Variation in integration type for the same loci can most likely be explained by drift, however polymorphic HK2 integrations which are present as solo LTRs in some populations while others retain the full-length provirus may be useful in understanding the dynamics and pressures for the inactivation of HK2 integrations.

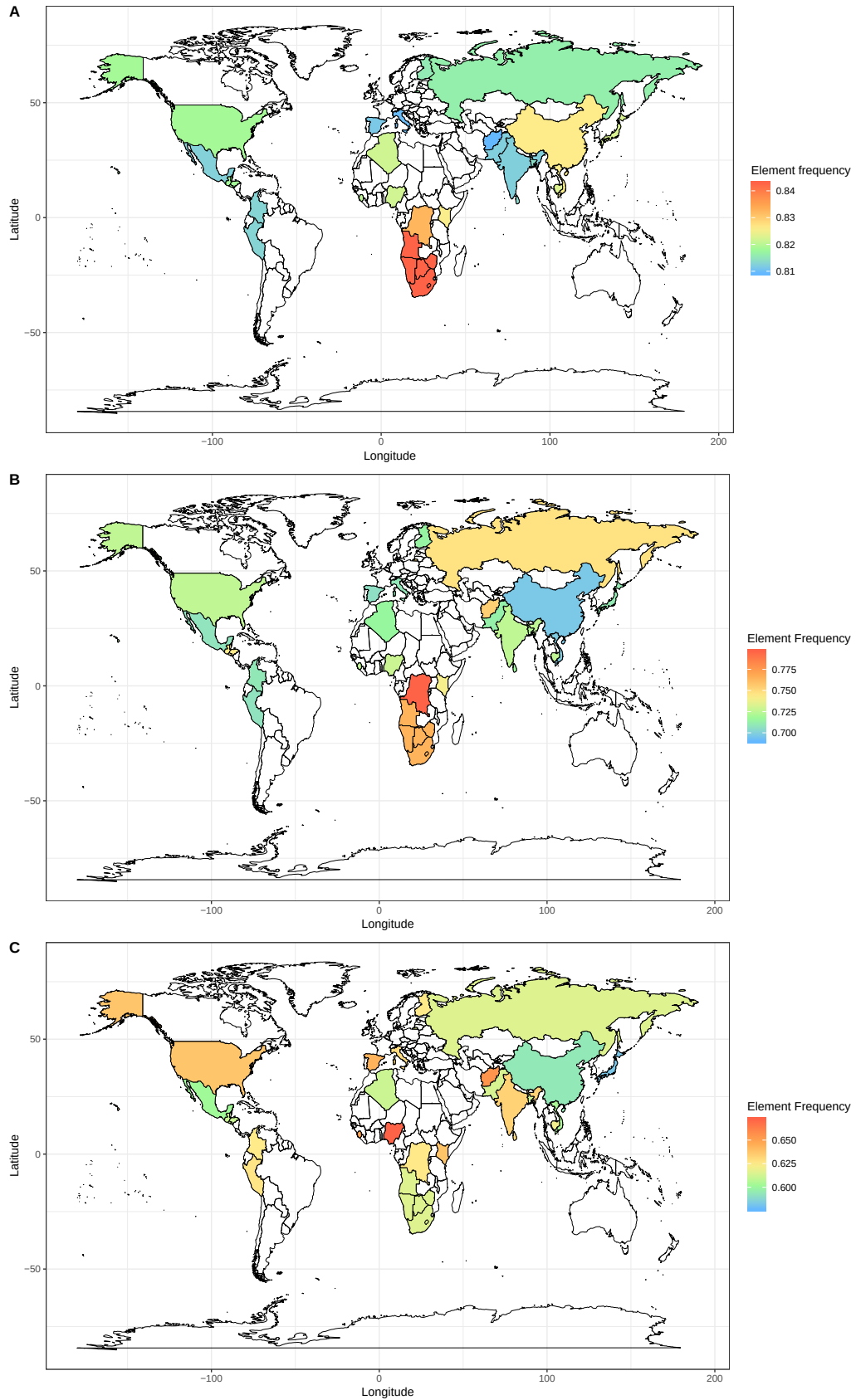


Fig. 3.2.: Estimated allele frequencies of polymorphic HK2 integrations distributed world-wide. Each map represents the frequency of an integration type: (A) Solo-LTRs, (B) 2-LTRs, (C) Presence of the full-length provirus. The difference between (B) and (C) is that (B) reflects the presence of the integration through LTR detection while (C) reflects the presence of the full-provirus using unique k-mers to reconstruct the viral genes.

3.3.2 No recent HK2 activity in non-African germlines

Before the release of the Wildschutte *et al.* (2016) study, I analysed 70 WGS derived from normal blood samples ("germlines") at an average range of 39× to 50× coverage from TCGA. I processed these paired-end WGS (2×100 bp) (Table A.1) with the algorithm STEAK (Specific Transposable Element Aligner (HERV-K)) (See Chapter 2) to identify integrations of HK2 that are present or absent from the reference human genome. A key aspect of this approach is the extensive validation that takes into account the sequence homology of the viral integration. This minimises false positive and false negative results by accurately detecting integrations not found in the reference (nonR) as well as absence of integrations found in the reference (inR) genome.

I compiled a list of 147 HK2 full-length proviruses and solo-LTRs in the reference genome (hg19) that were calculated to have integrated less than 5 MYA (Stringer & Andrews, 1988). As previously mentioned, older integrations are highly unlikely to be polymorphic, and, crucially, are not relevant to characterising on-going activity in the human population. I marked the presence/absence of these 147 HK2 integrations with high confidence in 70 patients from TCGA: 56 self-reported as "White", 12 as "Asian", 1 as "African-American", and 1 with no reported ethnicity or race (See 3.2). Overall I detected 33 canonical polymorphic integrations, 14 of which are inR and 19 of which are nonR among the 70 patients. These 33 integrations have been shown to already be present in the human population (Wildschutte *et al.*, 2016; Marchi *et al.*, 2013; Jha *et al.*, 2011). In addition to these known integrations, I found 1 nonR integration in chromosome 6 that is noted here for the first time (Table A.2). The results I obtained from TCGA samples support what has been shown by Wildschutte *et al.* (2016)—HK2 integrations found in non-Africans are already present in African populations. While I did detect a novel integration in chromosome 6 *in silico* among non-Africans, proper wet-lab validation would be necessary to confirm it as a true positive. However, this would still not rule out that it may exist in African populations.

Although I have focused on marking polymorphic and human-specific HK2 integrations, a great deal of these have already reached fixation (Figure 3.3). Notably, Xq21.31 while clearly frequent in the human population (Figure 3.3A) was not present in the TCGA cohort of cancer patients analysed.

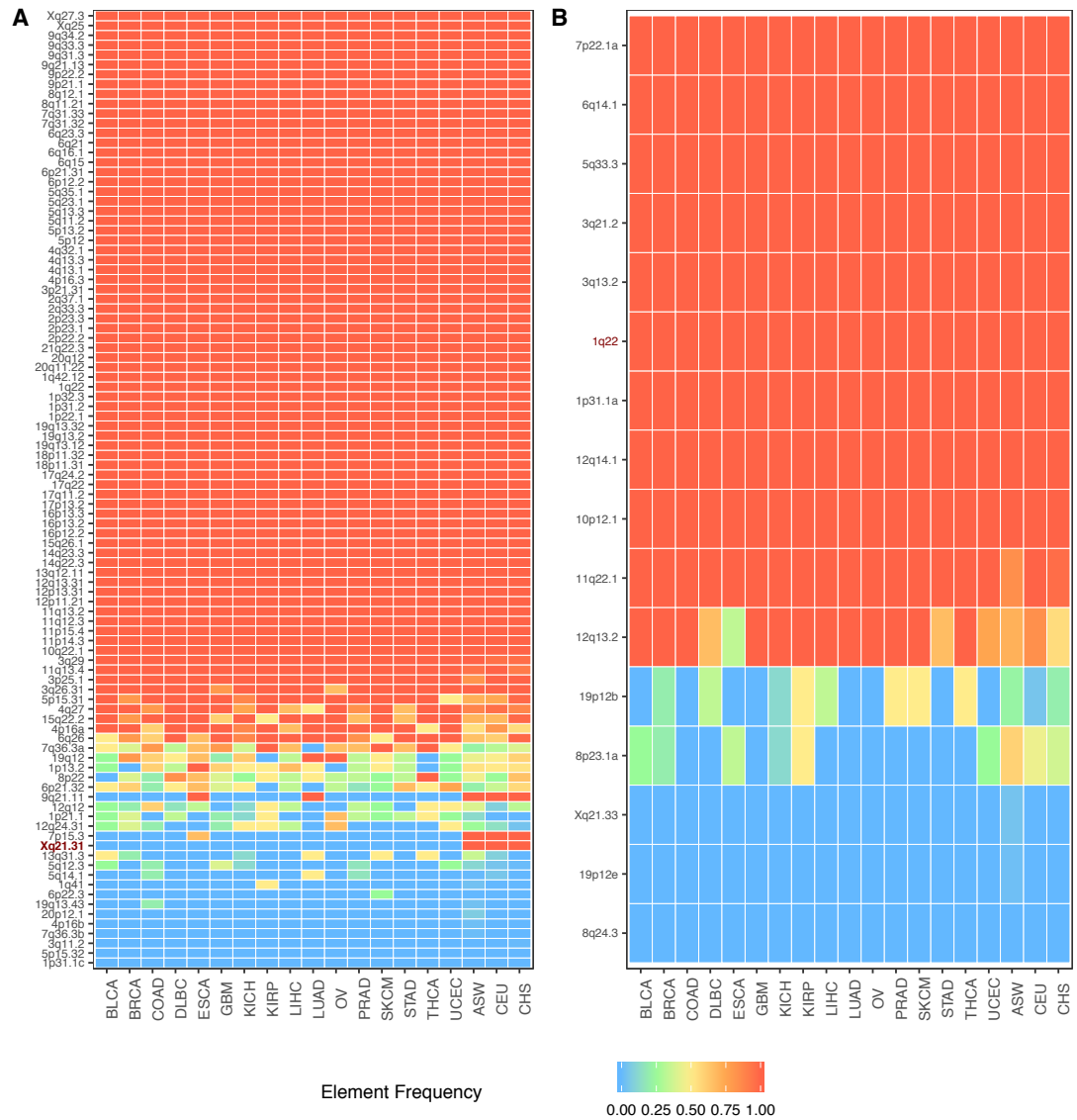


Fig. 3.3.: Inferred frequencies of unfixed HK2 integrations in germline across individuals from 1000 KGP and patients with different cancers. Figure (A) shows the frequency of solo-LTRs and (B) the frequencies of 2-LTR integrations. Frequency estimates for Americans of African ancestry in Southwest of the US (ASW), Utah residents with Northern and Western European ancestries (CEU), and Southern Han Chinese ancestries (CHS) (Wildschutte *et al.*, 2016) are displayed at the bottom-end for comparison purposes (i.e. TCGA samples were from self-reported "White", "Asian", and "African-American" individuals). Only integrations searched for both the Wildschutte *et al.* (2016) study and this study are displayed. TCGA abbreviations: Bladder Urothelial Carcinoma (BLCA), Breast invasive carcinoma (BRCA), Colon adenocarcinoma (COAD), Esophageal carcinoma (ESCA), Lymphoid Neoplasm Diffuse Large B-cell Lymphoma (DLBC), Glioblastoma multiforme (GBM), Kidney Chromophobe (KICH), Kidney renal papillary cell carcinoma (KIRP), Liver hepatocellular carcinoma (LIHC), Lung adenocarcinoma (LUAD), Ovarian serous cystadenocarcinoma (OV), Prostate adenocarcinoma (PRAD), Skin Cutaneous Melanoma (SKCM), Stomach adenocarcinoma (STAD), Thyroid carcinoma (THCA), Uterine Corpus Endometrial Carcinoma (UCEC).

3.4 Discussion

Within the human-specific clade, an integration rate of 1.16×10^{-4} integrations per generation suggests that HK2 ceased germline proliferation before 250,000 years ago, pushing back previous estimates (Marchi *et al.*, 2013; Jha *et al.*, 2011) of HK2 activity to possibly before the emergence of anatomically modern humans (Schlebusch *et al.*, 2012; Tattersall & Schwartz, 2009). A deceleration of the germline integration rate also coincides with the general decline of the HERV integration rate observed in humans (Magiorkinis *et al.*, 2013).

The activity of type I proviruses (those that do not have a functional *env* gene) during the last 10 MY suggests that infectivity may not be required to explain the observed amount of polymorphic integrations. An intriguing hypothesis arising from the observed pattern of HK2 accelerated activity is that the "invention" of *np9*, being the result of a partial deletion in *env*, may have enhanced the germline integration efficiency of HK2 as an LTR-retrotransposon before its eventual decline. In particular, type I integrations, like 1q22, which are solo LTRs in some individuals and full proviruses in others might be useful in understanding the putative pathogenicity of HK2 as a retrotransposon. Additionally, the very recent HK2 proviral integrations found in gorillas (Holloway *et al.*, 2019) may also provide a study model for us to understand how HK2 activity may have declined in our early human ancestors.

Interestingly, both of the youngest and rarest HK2 proviruses have been detected in the central African rainforest hunter gatherers, Mbuti Pygmies (Wildschutte *et al.*, 2016). Like the San, Mbuti Pygmies not only have an older date of divergence in comparison to other human populations, but also a larger past effective population size (Lopez *et al.*, 2018). This larger effective population size in the distant past would suggest an ancestral population that was demographically successful, possibly more interconnected by gene flow across several populations providing ample opportunity for HK2 integrations to proliferate the germlines of the ancestors of modern humans. The presence of young intact HK2 proviruses in the Mbuti fits with the very recent bottleneck that central African rainforest hunter-gatherers have suffered when we consider that small effective population sizes tend to struggle in purifying deleterious mutations (Hartl & Clark, 1997; Patin *et al.*, 2014; Lopez *et al.*, 2018). Yet despite the shorter amount of time to recover this recent bottleneck, Mbuti Pygmies have been shown to have purification selection that is more efficient than expected (Lopez *et al.*, 2018). Both these proviral integrations with intact ORFs are also found among populations with West African ancestry (Wildschutte *et al.*, 2016). It would be interesting for future studies to assess if these integrations possess a pathogenic burden or are beneficial exaptations for populations with these ancestries, particularly looking at how they might effect populations with different subsistence strategies (e.g. hunter-gatherer

versus farmer). Overlooking African populations, whom carry more HK2 integrations than non-Africans, has hindered an accurate calculation of recent HK2 activity in the human population until now. In the search for the putative pathogenicity of HK2 it is crucial to consider African populations.

While previous studies (Marchi *et al.*, 2013) have used simple neutral genetic models (Hartl & Clark, 1997) to simulate putative HK2 integration dynamics for the last 1MY, it would be best to consider a model with heterogeneous host mutation and recombination rates given that where an integration inserts itself, independent of whether it is reinfection competent or not, is under the control of host recombination and mutation when passing down in a Mendelian fashion. Moreover, if indeed HERVs possess a pathogenic burden as believed, accounting for selection pressures in the forms of purification and exaptation should be considered. By taking into account the demographic history of a population (Lopez *et al.*, 2018), the activity of HK2 in a sister taxon (Holloway *et al.*, 2019), and host restriction factors (Wolf & Goff, 2008; Hughes & Coffin, 2004), an elegant modelling of HK2 loss of activity could be possible.

Reference-free detection of archaic introgression

4.1 Background

The past decade has seen the industrial-scale sequencing and studying of ancient and modern genomes alike. Prior to this, studying the relationship between archaic hominins and anatomically modern humans was predominantly left to paleoanthropologists whom have proposed several evolutionary models for the emergence of our species *Homo sapiens* since the first half of the 20th century. The most popular hypotheses which have been in competition up until very recently are the multiregional evolution and the recent out of Africa model (RAO). Multiregional evolution proposed that humans evolved across the Old World from local ancestors through an interbreeding network since at least 1.8 million years ago (Wolpoff *et al.*, 1984), while RAO argued for an origin from a single region in Africa and then dispersing across the globe about 100,000 years ago (Stringer & Andrews, 1988). The turning point in this debate is often considered to be in the late 80s when Cann *et al.* (1987) conducted a study of mitochondrial DNA of present-day individuals where the results suggested a recent African origin for modern humans although future studies attempting to replicate the results showed that these mtDNA haplotype trees were flawed (Maddison, 1991; Templeton, 1992). However, several other mtDNA, Y-chromosome and autosomal DNA studies with reliable haplotype trees also corroborated the RAO model placing it as the consensus view at the start of the 21st century (Harding *et al.*, 1997; Ingman *et al.*, 2000; Jobling & Tyler-Smith, 2003).

Since the sequencing of Neanderthal (Green *et al.*, 2010) and Denisovan genomes (Reich *et al.*, 2010) we have been able to characterise the evolutionary relationship between archaic and modern humans using genetic data, subsequently allowing us to explore the hypothesis of gene flow among them (Reich *et al.*, 2011; Sankararaman *et al.*, 2012; Vernot & Akey, 2014; Sankararaman *et al.*, 2016; Vernot *et al.*, 2016). It is now understood that neither multiregionalism nor RAO can strictly explain the complex interactions between hominin taxa (Galway-Witham & Stringer, 2018). A number of intermediate models have been suggested where their main differences lie in how human genetic dispersal has taken place while mixing with other hominin groups

outside of Africa. Intermediate models have included the "Leaky Replacement" model, which is fundamentally RAO with limited hybridisation between modern and archaic populations. However, some of these models continue to assume an evolutionary history of humans diverging off a closely related core African population without accounting for the presence of multiple populations within Africa and their absorption through admixture before proximate extinctions of other *Homo* taxa prior the exodus from Africa. Others have made attempts to integrate genomic, archaeological, fossil and paleoenvironmental data to form models inclusive of African population structure (Lahr & Foley, 1998; Harding & McVean, 2004; Stringer, 2016; Scerri *et al.*, 2018; Henn *et al.*, 2018). Our understanding so far of the relationships between archaic and modern humans is best depicted in Figure 1.5. The crucial differences between proposed intermediate models are the varying amount of gene flow between regions and split times of modern humans (Henn *et al.*, 2018).

As research continued and progressed in this field, the idea of admixture between archaic and modern humans has become an indispensable event to take into account in models addressing human evolution. Evaluating admixture between non-Africans and archaic hominins has been comparatively more straightforward given the availability of both Neanderthal and Denisova genomes (Green *et al.*, 2010; Reich *et al.*, 2010) (See Introduction 1.2.1).

The sequencing of the non-African archaics has led to irrefutable interest in wanting to understand what relationship they had with modern humans and moreover what possibly connects us to them. This yearning has manifested in a number of questions about our origin as a species, but it has also prompt questions on whether this phenomenon has occurred not just in Eurasia but also in Africa (Hammer *et al.*, 2011; Lachance *et al.*, 2012; Hsieh *et al.*, 2016a; Skoglund *et al.*, 2017) given that *Homo* populations, including those that may be extinct modern humans (Rightmire, 2009; Grun *et al.*, 1996; Porat *et al.*, 2010; Hublin, 2017), have lived longer alongside each other within Africa than outside (See Introduction 1.2.2).

In order to answer these outstanding questions in the field, a number of tools have been developed to identify and characterise signatures of archaic admixture, predominantly depending on linkage disequilibrium and variant density. Patterson's *D* statistic, for example, measures excess sharing of derived alleles between two sister populations (ingroup) and an outgroup. If neither of the ingroup populations received genetic material from the outgroup, they should share about the same number of derived alleles with the outgroup. While this method can detect whether there is an asymmetrical variant sharing between archaic and modern humans, it is unable to pinpoint segments in the genome that are of introgressed archaic origin (Patterson *et al.*, 2012). Moreover, demographic scenarios such as ancient population structure can generate similar results (Eriksson & Manica, 2012; Theunert & Slatkin, 2017).

We can consider recombination and mutations as the basis of understanding an expected length of an introgressed tract in relation to the time since the admixture event. Introgression is distinct from incomplete lineage sorting (ILS) in that it should leave behind longer tracts with the former being more recent (Liang & Nielsen, 2014). The difference in the size of the tracts should provide a way to tell whether a shared tract with an archaic hominin is in fact introgression or the genomic relic from an earlier common ancestral population.

4.1.1 Methods to infer archaic introgression

Linkage disequilibrium based methods: S^* and Sprime

Prior to the availability of reference genomes following the sequencing of the full Neanderthal and Denisova genomes, Plagnol and Wall (2006) sought to take advantage of the logic that putatively introgressed tracts would have had a limited number of generations (e.g. ~ 2000 , assuming admixture occurred 40,000-50,000 years ago) to be broken down by recombination, while still retaining a set of strongly associated variants. Those tracts could be identified by linkage disequilibrium (LD) where variants in an archaic segment should be strongly associated with other archaic variants in the genome, in other words introgressed archaic variants should be found in high LD. Plagnol and Wall came up with the S^* statistic (Figure 4.1) which extracts this particular information through a scoring scheme that searches for single nucleotide polymorphisms (SNPs) that are in highest LD and which are derived.

Vernot & Akey (2014) applied S^* to 379 European and 286 East Asian genomes where they scanned whole genomes in 50-kb windows under the assumption that a putative date of admixture would have been 2,000 generations ago (50,000 years under a generation time of 25 years (Plagnol & Wall, 2006)). This was an adjustment made to S^* to look beyond a few coding (Plagnol & Wall, 2006; Wall *et al.*, 2009) and non-coding regions (Hammer *et al.*, 2011) as previously done.

Among the high scoring regions for the S^* , the authors retained only the top 1% as putative introgressed candidates for further analyses. Approximately 40 gigabases (Gb) of potentially introgressed sequence across all genomes was retrieved using this method. They then calculated how likely these high scoring candidates were to be of Neanderthal origin by comparing each of the against the high coverage Altai Neanderthal genome (Prüfer *et al.*, 2014), and estimated a 'Neanderthal match p -value'. This archaic p -value is estimated by first calculating the number of variants on a detected segment which are also present in the Neanderthal sequence divided by the number of variants shared by both Neanderthal and humans in that same segment. The resulting value is then compared to similar segments in non-introgressed portions of the genome. These similar segments are chosen from the portion of the genome

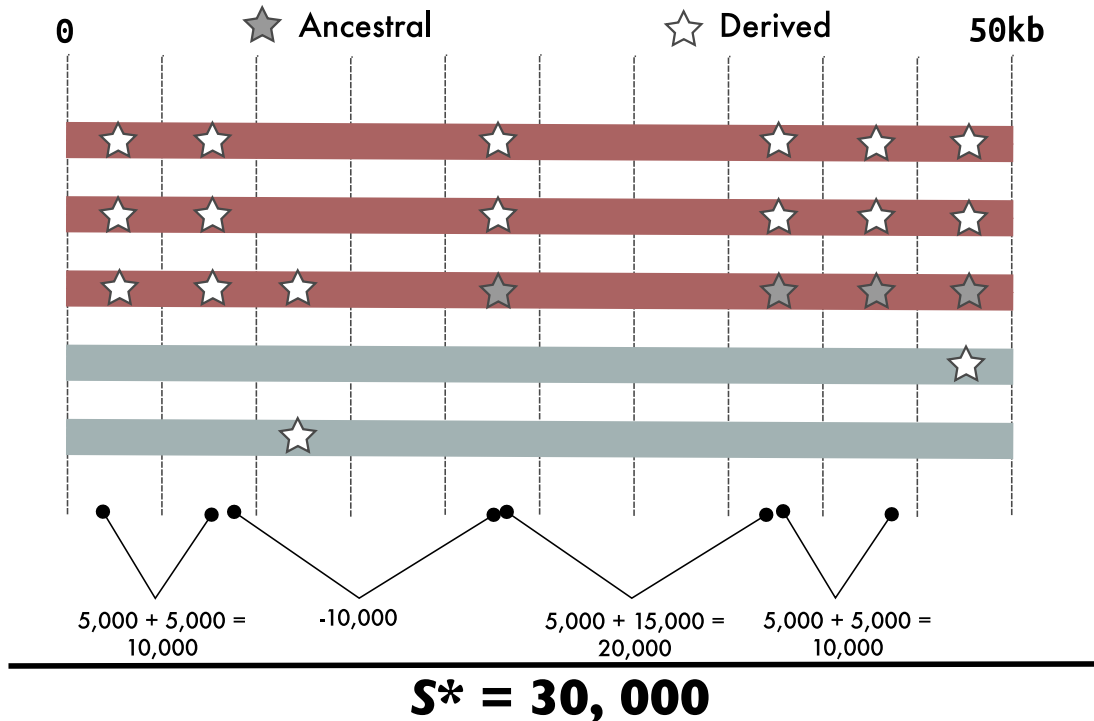


Fig. 4.1.: Schematic for S^* scoring. Here we see phased chromosomes and a test population that is represented in red and the outgroup in grey. S^* attempts to search for the most optimal sums of scores for an overall subset of SNPs at a given locus. It rewards fully linked pairs of sites, in other words where two successive SNP positions do not differ, and the increase in that reward is proportional to the distance between the positions. Consequently, S^* values increase with increasing LD within a window. The SNP positions that provide the optimal score are taken note of and can then be used to provide the delimited region that gives that optimal score. The lower bottom of the figure depicts an example of how S^* goes about calculating the most optimal sums of scores. Seven positions within a window of 50,000 bp and only SNPs not present in the outgroup are considered. Sites within the haplotype are labelled as a grey star (i.e. ancestral) or a white star (i.e. derived), those highlighted in red (target) are calculated to give the most optimal solution and that no other set of SNPs gives a higher score. The first 10,000 bp displays higher linkage than the following 10,000 bp where there is a change in genotype leading to a penalty of -10,000. We can see that in the calculation that follows the positions from about 30,000 bp onwards are rewarded for being linked and outweigh any penalty from before. The scoring function works in rewarding linked sites with a score of 5,000 plus the distance between sites, penalties are given -10,000 for up to 5 mismatches in the above example. Regions with more than 5 mismatches or no linked sites are $-\infty$, or essentially no score. The total S^* score for this example is 30,000.

not covered by regions that are putatively introgressed (i.e. segments that did not have a significant S^* value when compared to simulations of 80 different demographic models with admixture). The archaic p -value is the comparison between the putatively introgressed region and a similar non-introgressed segment which has been conditioned on 1) number of variant sites, 2) average frequency of variant on the segment, 3) standard deviation of frequency of variants on the segment and finally 4) length of the haplotype. The authors demonstrate that they can recover Neanderthal introgressed segments with a 5% FDR using this approach. Prior to calculating an archaic p -value, the authors showed that S^* was able to identify ~30% of all true positives (74 Mb/240 Mb) when only comparing to simulations based on demographic inferences and not using any archaic reference for validation.

This study also elucidated that the average length of introgressed tracts found was about 57 Kb, with Europeans hosting approximately 20-25 Mb of Neanderthal sequence per genome while East Asians retain about 25-30 Mb of Neanderthal DNA per genome. This is consistent with a study by Wall *et al.* (2013) that had also found more archaic segments in East Asians than in Europeans. Through simulations the authors (Wall *et al.*, 2013) explored if their data were consistent with a two admixture pulse model rather than a single pulse in a scenario where the first pulse took place into a Eurasian ancestor and a second pulse took place in Asians. Meanwhile, other studies have suggested that East Asia's smaller population sizes might have led to less-efficient purifying selection and therefore a larger retention of Neanderthal segments in East Asians could possibly be due to drift (Sankararaman *et al.*, 2014).

Two years later, Vernot *et al.* (2016) published a work focusing on the recovering of archaic segments from Melanesians. The authors recovered which archaic fragments were of either Neanderthal or Denisovan or shared origin. Data from 27 unrelated Papuans were used and S^* was modified to run specifically at the individual level. 107 Yoruban individuals were used as a population outgroup with the logic being that this African population is expected to have a very low amount of archaic ancestry than the target group, in this case Papuans. The comparison with Yorubans provides a way of calculating a null distribution for how many variants are shared with archaic individuals in the absence of archaic introgression which provides an S^* threshold for evaluating whether a haplotype is archaically introgressed or within the confines of shared variation. Similar to the 2014 study (Plagnol & Wall, 2006), S^* candidate segments were compared to non-introgressed fragments in the genome, which in this case was taken from Yorubans in 50-kb windows. The information collected on these non-introgressed fragments were a) number of variant sites, b) length of the haplotype, c) number of variants shared with an archaic genome (Neanderthal or Denisovan), d) total number of sites variable on the fragment and in the archaic species, and e) the number of unmasked bp in the 50-kb window. For every S^* candidate, corresponding Yoruban haplotypes were selected with exact same number of variants, length, and

number of unmasked bp in the region. A distribution is then generated with the expected archaic match percentage when there is no considerable Neanderthal introgression in a population conditioned on the characteristics of a putatively introgressed segment. From this distribution an archaic match p -value is obtained. Vernot and colleagues (Vernot *et al.*, 2016) then used these match p -values to develop a likelihood method to estimate the proportion of Neanderthal, Denisovan, and non-introgressed sequence amongst their S^* candidates. Briefly, the archaic match p -values are converted into posterior probabilities which allowed the authors to identify introgressed regions at an FDR of their choice whilst assigning them labels of whether they are "Neanderthal", "Denisovan" or "Ambiguous" (i.e. suggesting that they may be a shared derived haplotype in for both archaics).

They found about 51 Mb of archaic sequence in Europeans, 55 Mb in South Asians, 65 in East Asians, and 100 Mb in Oceanic populations. Prior to using archaic match p -values, the authors estimated that S^* should be able to recover about 60% of the archaic sequence in a population (Europeans–30.6 Mb; South Asians–33 Mb; East Asians–60 Mb).

Browning *et al.* (2018) later introduced an optimised S^* -like method called Sprime which no longer performs detection on sliding windows but rather surveys whole chromosomes and accepts fine-scale genetic maps as input in addition to whole genome sequence genotypes. The authors begin with the premise that if a pair of alleles within a target sample are in fact archaic-specific and from a single introgressed haplotype, then there should be no change in introgression status between these two alleles unless there has been a recombination event. In detail, like S^* , Sprime performs using a pairwise score where the score can take a positive value when there are no mismatches (i.e. mismatches should only be due to recombination events in an introgressed haplotype), a negative value is there are 2-5 mismatches or $-\infty$ if more than 5 mismatches. However, in Sprime the number of mismatches (i.e. number of genotype changes) allowed between adjacent alleles when scoring a putative introgressed segment is based on the number of samples used. Under the premise that in a larger number of target samples a higher number of recombination events (D) will be observed for high frequency introgressed haplotypes (n), Sprime penalises only if the putatively introgressed haplotype is rare (n is small) and there are too many recombination events (D is large).

On the other hand, Sprime also rewards a pair of alleles by considering local mutation and recombination rates. This is done by considering the ratio of mutation rate to recombination rate per region defined in the genetic map in order to avoid mistakenly finding or overlooking introgressed regions. For example, we might observe two opposite scenarios related to detecting false positives and false negatives: a) the combination of a high mutation rate with a low recombination rate, which may mistakenly lead

to finding many high-LD pairs of alleles in the region but possibly not because of introgression; b) alternatively regions characterised by a high recombination rate and a low mutation rate, which would result in the inability to find high-LD pairs, whether archaic or not. Sprime scores are positive when two adjacent alleles are in high-LD and the local rate of mutation per centiMorgan, or m , is low.

The accuracy of the reported putative archaic-specific alleles using Sprime was over 93% across the range of allele frequencies that were simulated from different curated demographic histories. Using Sprime the authors detected archaic introgressed segments in worldwide non-African populations from the 1000 Genomes Project (1KGP), Papuans from the SGDP (Mallick *et al.*, 2016), and individuals from the UK10K project (The UK10K Consortium *et al.*, 2015). Overall Browning *et al.* (2018) detected, with very high match rates to an archaic reference (i.e. 0.72–0.74), 1.36 Gb of introgressed genome across 19 European, Asian, and American populations from the 1KGP but only about 304 Mb introgressed segments across the genomes in UK10K study although the authors note that this may be due to the methods of generating the dataset. Lastly in Papuans from the SGDP, they found 239 Mb of the genome covered by one or more detected segments amounting to an average proportion of haplotype carrying a detected segment at a position is 1.48%. They found evidence that Asians carried Denisovan introgression as previously reported (Prufer *et al.*, 2014; Sankararaman *et al.*, 2016; Qin & Stoneking, 2015; Skoglund & Jakobsson, 2011), and furthermore, their findings supported two waves of Denisovan admixture, one from a population closely related to the Altai Denisovan into East Asian and another more distantly related to the Altai Denisovan in Papuans and South Asians.

Probabilistic methods: Hidden Markov model and conditional random field

Another option to using LD-based methods in detecting introgressed sequences is to incorporate parametric assumptions into a probabilistic framework. Two methods that do this are hidden Markov models (HMM) and conditional random field (CRF) with the concept that the genome is made up of SNPs, each of which is a hidden random variable with two states: either human or archaic. These methods integrate what we know about the biological processes such as human recombination and mutation and what we understand about the demographic events since the human exodus from Africa, on the assumption that we would not expect to find Neanderthal or Denisovan contributions in West Africans. With this parametric information one can calculate the probability that a given sequence is of archaic origin and if it is present in modern humans due to admixture. Both Prufer *et al.* (2014) and Seguin-Orlando *et al.* (2014) made use of HMM methods to search for archaic contribution in both recent and ancient human populations. The main difference between the two approaches was that, respectively, one provided a priori chosen parameters whereas the other gathered the relevant parameters from a reference dataset.

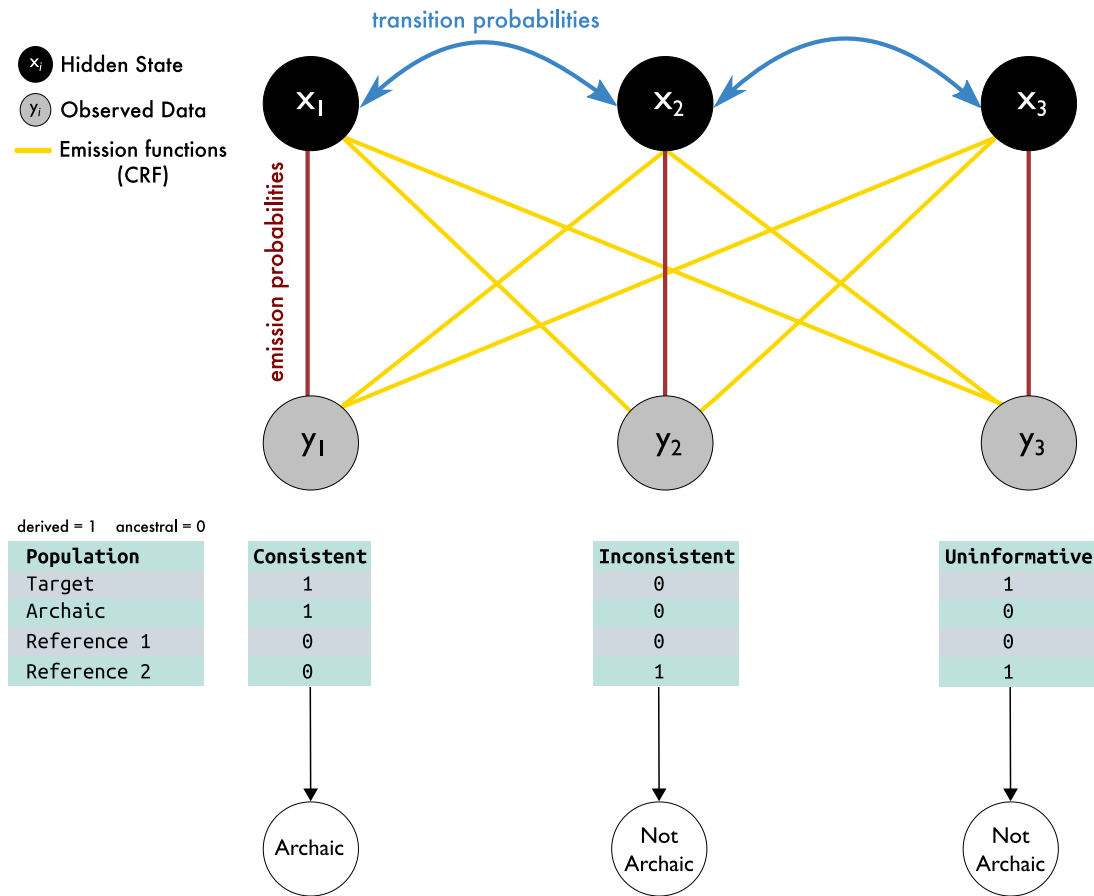


Fig. 4.2.: Overview of Hidden Markov model and conditional random field framework. Here is a depiction of how probabilistic models, such as HMM and CRF, can estimate the ancestry (x_i) of a SNP in a genome sequence (i). The possible hidden states for x_i are either introgressed (archaic) or not introgressed (not archaic). In the example, y_i is a matrix composed of individuals from a target population, an archaic population, and two outgroup populations that act as references for modern human variation. These are considered at three SNP positions (below model). Sites which are consistent with introgression (x_1) are determined by the derived allele of the target population also present in the archaic population but not in the reference. In the case of x_2 , the derived allele is only in one of the reference populations suggesting that the derived allele is modern human variation and therefore inconsistent with introgression. In x_3 , both target and reference populations share the derived allele but absent in the archaic population which is uninformative. Hidden states (ancestry) are connected to the observed data (y_i) through emission probabilities (red) for HMM or emission functions (vertical red and diagonal yellow) for CRF. These emission functions in a CRF model can be used to evaluate whether a site is consistent or inconsistent with introgression, like in HMM. Additionally they can also score whether a haplotype overall is closer to the archaic sequence than to the reference haplotypes. Horizontal connections between x_i and x_{i+1} denote transition probabilities (HMM) or transition functions (CRF) which model linkage between ancestral states along the genome. These parameters depend on the recombination, admixture proportion, and time of admixture.

Prüfer *et al.* (2014) divided the genome into 5-kb bins and counted in each bin the number of derived variants observed in a pool of archaic reference genomes (e.g. Neanderthal and Denisovan). They used 107 Yorubans as a comparison of whether a site was consistent with introgression, that is to say that if a variant was absent in Yorubans but present outside Africa, it was considered as a putatively introgressed site. However, if the variant was present in both Yorubans and their test sample, the two samples were closely related to each other and therefore not consistent with introgression. Windows that were made up of primarily introgression consistent sites and were at least 0.0005 Morgans (i.e. the expected switch rate for expected admixture 2,000 generations ago) long had a higher likelihood of being introgressed. This allows for the recovering of segments that are longer (i.e. > 10-kb) which effectively avoids the detection of short windows that might be due to ILS. Furthermore, the authors set a scoring scheme that would generate emission probabilities for observing whether a window would be consistent in the human state or in the archaic state (Figure 4.2). Consistent sites were defined as such if the test haplotype in Sub-Saharan Africans are all ancestral, in the target population it was derived, and if at least one of the archaic alleles in the test haplotype was derived as well. The HMM parameters were 1) the ancestry switch rate, 2) the prior probability for archaic ancestry (set to 0.01) and finally, 3) the probability that a locus is of archaic ancestry if all SNPs are "consistent" and in turn, that a locus is of human ancestry if all SNPs are "inconsistent" (set to 0.99).

This method was initially trained using Yorubans as reference to calculate the enrichment of introgressed sites that Europeans and Asians had in relation to them. What was observed was a 100-fold increase in the number of windows with posterior probability of being archaic in Non-Africans compared to Yorubans. Prüfer and colleagues (Prüfer *et al.*, 2014) also estimated that there is a 70-100% enrichment of archaic ancestry in East Asians compared to Europeans.

In the same vein, Seguin-Orlando (Seguin-Orlando *et al.*, 2014) also divided the genome into fragments deemed archaically introgressed or of human origin. Sites with a derived variant that were found to be variable in Neanderthal and Non-Africans but absent from Africans were labelled as consistent with introgression. Alternatively, sites where a variant was variable in Neanderthals, Non-Africans, and Africans was considered as not introgressed.

The emission probabilities of the model were first trained with fixed transition values that had been given in a study on models of local ancestry by Gravel (2012). This provided them with a higher probability (0.0155) of observing consistent sites while in the archaic state and a lower probability (6.67×10^{-9}) while in the human state. Transition probabilities are then estimated using maximum likelihood on the same set of sites as used to calculate emission probabilities. Specifically, the rate of transition,

αNH , from one state to another can be converted into admixture dates by calculating this equation:

$$NH = m(t - 1) \text{ per Morgan}, \quad (4.1)$$

where the product of admixture proportion is m and time of admixture in generations is t .

Using this method, Seguin-Orlando and colleagues (Seguin-Orlando *et al.*, 2014) found the approximate time of admixture to be around 16,600 years ago, when testing an ancient individual that was alive about 36,200 to 38,700 years ago. This date of admixture, considering the date of the sample, would be in-line with what other studies suggested using only present-day samples (37,000-86,000 years ago) (Sankararaman *et al.*, 2012; Green *et al.*, 2010; Reich *et al.*, 2010).

Another approach similar to HMM is CRF which models the ancestry of a set of proximate SNPs along windows of the genome. The ancestry state of each SNP, like in HMM, is also binary: modern or archaic. However, CRF models can contain more than one emission function which can incorporate other forms of information related to the data for example, LD, haplotype structure and allele configurations from multiple samples. The parameters are then calibrated through training using simulations with particular demographic assumptions ranging from divergence dates between humans and the archaic to effective population sizes of the test population.

Sankararaman *et al.* (2014) used a CRF model in their study where the first emission function provided a high probability of being archaically introgressed if a variant was found in Non-Africans and in the Neanderthal reference genome but absent in Africa; akin to what the aforementioned studies had done with their HMM methods. Additionally, they incorporated a second emission function where the distance of a putatively introgressed haplotype to the Neanderthal genome and the genomes of Africans was considered. This is to say that it would deem variants from regions which are very similar to Neanderthals but absent in Africa as more likely to having been introgressed. The last feature they included was an emission function, much like the transition matrix in a HMM, trained on simulated data to expect a putative haplotype corresponding to an admixture event about 2000 generations ago (i.e. an expected length of 50-kb). Using this CRF model the authors detected in the 1000 Genomes Project 15% of archaically introgressed sequence with 99% precision. The more archaic sequence the authors were able to detect, the lower their precision (e.g. for 38% of the archaic introgression they detected, their precision decreased to 98%). This study

found 1.17-1.20% of the autosomes in Europeans have Neanderthal ancestry and that between 1.37-1.40% of East Asian autosomes have Neanderthal ancestry.

Two years after developing this CRF model Sankararaman and colleagues (Sankararaman *et al.*, 2016) applied their method to the Simons Genome Diversity Project (Mallick *et al.*, 2016) which includes 257 high-coverage individuals from 142 world-wide populations. Here both the Neanderthal and Denisovan genomes were used in the analysis and they identified 1.06% of autosomal DNA in Europeans as being of Neanderthal origin while in East Asians this proportion was slightly higher at 1.40%. Furthermore they found a small although still significant amount of Denisovan contribution in East Asians, estimated to be 0.06%.

While inarguably powerful methods, so far HMM and CRF have predominantly relied on two things: a priori chosen parameters which are based on well-established demographic assumptions and a reference dataset. Recently an HMM method was released by Skov *et al.* (2018) which depends neither on demographic assumptions nor on an archaic reference to detect introgression. This model bases its logic on the observation that an archaic tract introgressed into a population should display a high density of variants not found in populations which have not experienced introgression. Like the previous methods reviewed here, Skov and colleagues (2018) focus on a scenario where admixture has occurred between a deeply divergent archaic population (e.g. Denisovan) and a modern human population (ingroup - Papuans) but where admixture has only occurred with the ingroup and not with the outgroup (e.g. Yoruba). By removing variants that are shared with the outgroup, they inspect the density of the remaining variants in the ingroup which are, essentially, private SNPs. Because this method does not require demographic assumptions or an archaic reference and does not require phased data, it can detect introgression with just the genome samples provided. In their analysis they use Sub-Saharan Africans from the 1000 Genomes Project as an outgroup. Skov and colleagues (2018) demonstrate, with 89 Papuan genomes, that they are able to recover more Denisovan introgressed segments (77 Mb out of 164.23 Mb of archaic sequence in Papuans) than previous methods (Vernot *et al.*, 2016; Browning *et al.*, 2018) possibly because this approach does not rely upon the similarity to the Altai Denisovan genome reference, ultimately demonstrating the advantage of a reference-free detection method. In addition to detecting introgressed fragments, this method can also infer admixture proportions as well as divergence time of human and archaic populations—a useful tool which can potentially be exported for exploring introgression in African populations as well as for studying ancient admixture in other species.

4.2 Introduction

Simulating several demographic scenarios has corroborated how simulations can help our understanding of the demographic history of a population (Gravel, 2012). In the case of Africa, for example, there have been studies that simulated important events related to ancient and more recent migrations such as the Bantu expansion and pastoralist movements from East Africa (See Chapter 5). Simulations can assist our understanding of African pre-history, in order to make more appropriate demographic assumptions about the history of populations in the continent. However, our knowledge of what may have occurred since the divergence of modern human and other archaic forms is limited by both the fossil record and availability of modern genomes (Schlebusch & Jakobsson, 2018). Through simulations we can explore events that may explain the signals observed in populations which diverged earlier from other groups (e.g. Pygmies and San) (Behar *et al.*, 2008; Karafet *et al.*, 2008). However, without more archaeological data and more genomes, both ancient and modern, those explorations might overall be of limited scope and self-fulfilling.

To better elucidate both the power and limitations of simulations in regards to searching for archaic introgression in Africa, this chapter will present the results of the analyses of several simulations based on different demographic models that one might encounter when exploring different demographic assumptions. I will show how demographic and biological processes can not only confound the detection of but also mimic archaic introgression signals when using LD-based methods without an archaic reference.

4.3 Materials and Methods

4.3.1 Coalescent whole-genome simulations

Using MaCS (Chen *et al.*, 2009a) I simulated 19 demographic models that were based on models used by Plagnol & Wall (2006) as well as the Hsieh *et al.* (2016a) and Hsieh *et al.* (2016b) demographic inferences. I simulated scenarios with and without recombinational heterogeneity. When accounting for recombinational heterogeneity I used HapMap genetic maps when simulating each chromosome. For the mutation rate we used Scally & Durbin (2012) rate of 10^{-9} per base pair per year. Because the simulations were made with the intention to better understand how each demographic parameter is capable of influencing S^* , I adjusted a number of parameters as described in Table 4.1. Effective population size, dates of divergence, dates of admixture and migration proportions were adjusted directly from Plagnol and Wall (Plagnol & Wall, 2006) command-line for ms (Protocol S1, 2006) particularly in simulations where

exponential growth took place in Africa. Furthermore we took into account the parameters used in Hsieh *et al.* (Hsieh *et al.*, 2016a) specifically Supplementary Figure S10 of their manuscript and Hsieh *et al.* (Hsieh *et al.*, 2016b) demographic inferences made on Central African hunter-gatherers, specifically their Table 1 and Table S1, when simulating population structure with logistic growth. Each simulated human population was of a sample size of 7 and 1 chromosome was simulated for the archaic population as has been done in previous simulations exploring archaic admixture (Vernot *et al.*, 2016).

The scripts for the simulations we produced can be found at <http://github.com/santaci/archaia/>.

4.3.2 Performing S^* and Sprime

S^* estimates were performed with a pipeline developed by Vernot *et al.* (2016) which can be found at <http://github.com/bvernot/freezing-archer>. I used simulated Eurasians as the outgroup and simulated Africans as the target group of S^* . I calculated S^* on each of our simulations in 50-kb windows with a step size of 10-kb. I compared S^* scores for each of the different simulations with pairwise comparisons (Wilcoxon) as well as multigroup testing (Kruskal-Wallis). The top candidates were originally defined using the null hypothesis simulation (i.e. absolute score greater than 400,000) but because there were some simulations that did not meet the null hypothesis cut-off, I also took the top 1% as reported in previous studies (Hsieh *et al.*, 2016a).

4.3.3 TMRCA estimates

We used a previously described method to calculate TMRCA for the simulated genomes in 10-kb windows for both random regions and for top S^* candidates (Thomson *et al.*, 2000; Hudson, 2007). This method computes the average coalescent time from derived sites assuming that mutations have taken way in a Poisson distribution. From that value we estimate TMRCA in years by using mutation rate in years based on pedigree studies (Scally & Durbin, 2012). The scripts produced to calculate TMRCA can be found on <http://github.com/santaci/archaia/>.

4.4 Simulations

Here I have, as an initial step, explored the behaviour of the reference-free method S^* that was first introduced by Plagnol & Wall (2006) and then applied in a number of recent studies looking at whole-genome sequences (Lachance *et al.*, 2012; Vernot &

Akey, 2015; Hsieh *et al.*, 2016a; Browning *et al.*, 2018). The model used here as a basis for the simulations has been shown to be simple enough to allow a precise estimation of parameters but sophisticated enough to capture characteristics of each simulated population (Plagnol & Wall, 2006).

The simulations included biological parameters such as a mutation rate of 10^{-9} per base pair per year (Scally & Durbin, 2012) and, assuming a putative admixture date comparable to non-Africans with Neanderthal (Hammer *et al.*, 2011) of 2000 generations, and a generation time of 20 years which is consistent with the lifespan of chimpanzees and gorillas (Plagnol & Wall, 2006; Matsumura & Forster, 2008; Teleki *et al.*, 1976). This generation time is likely to be relevant for the life-history of early modern humans as well and originally used by Plagnol & Wall (2006).

Additionally, I incorporate demographic features of well-known events in the history of modern humans, such as recent population growth for the simulated populations as well as a bottleneck in non-African populations (Europeans and Papuans in my dataset) to mimic the effect of the Out of Africa expansion with an additional bottleneck in Papuans, to highlight the colonisation process of Oceania (Plagnol & Wall, 2006; Yang *et al.*, 2014). Furthermore, I consider the differentiation in population structure and growth types (Hammer *et al.*, 2011; Scerri *et al.*, 2018) by simulating populations with limited carrying capacity (logistic growth) and asymmetric migration as could have been seen in hunter-gatherer populations and which previous studies have estimated through demographic inference (Hammer *et al.*, 2011; Hsieh *et al.*, 2016a; Excoffier *et al.*, 2013). The effective population sizes used are based on the original Plagnol & Wall (2006) ancestral effective population size. A sample size of seven individuals (i.e. 14 haplotypes) per population is simulated to match the sample sizes available for African populations from the SGDP (Chapter 5).

I simulate 26 two-population model scenarios (Table 4.1) to explain in detail how different divergence and admixture times, recombination rates, demographic events, and migration influence S^* results (Figure 4.3).

4.4.1 Admixture Proportion

S^* statistics were designed to detect regions of archaic introgression and so I initially explored what differences can be observed when S^* looks at two scenarios that are identical in all other parameters but different in whether archaic admixture is present or not (Sim Null and Sim 2). In Figure 4.4 we observe that a scenario with no admixture has lower absolute values for S^* than those that did include admixture. Moreover, we see an appreciable skew of the overall S^* distribution when admixture is present, with top S^* values being larger in the admixed than non-admixed (Null) scenario.

Tab. 4.1.: Parameters of the 26 simulations that were generated and analysed. The first column numbers the simulation modelled and the subsequent columns delineate the status of other parameters used. The demography column refers to the basis model used (Figure 4.3B-E) and which is the target population in the simulation. If a recombination map was used to simulate local recombination rate it was labelled as variable, whereas if a single recombination rate was used across the simulations, it was labelled constant. The sample column reports the number of diploid individuals simulated in each model. Dates of divergence between human and archaic as well as between Africans and Eurasians are given in years in separate columns. Date of admixture is also provided in years followed by a final column that reports proportion of archaic admixture simulated in the model. Simulations with no archaic admixture have "NA" under admixture date and 0 under admixture proportion (Null, 3, 11, 19). A: Archaic; H: Human Af: African; Eu: European.

Models	Split (A-H) (1)	Split (Af-Eu) (2)	Admixture Date (3)	Admixture (%)	Recombination	Demography
Beta 1	700 KYA	128 KYA	NA	0	Constant	B
Beta 2	700 KYA	128 KYA	32.2 KYA	2	Constant	B
Beta 3	700 KYA	128 KYA	32.2 KYA	4	Constant	B
Null	700 KYA	128 KYA	NA	0	Variable	B
1	700 KYA	128 KYA	32.2 KYA	2	Variable	B
2	700 KYA	128 KYA	32.2 KYA	4	Variable	B
3	700 KYA	128 KYA	NA	0	Variable	C
4	700 KYA	128 KYA	32.2 KYA	2	Variable	C
5	700 KYA	128 KYA	32.2 KYA	4	Variable	C
6	700 KYA	128 KYA	100 KYA	2	Variable	B
7	1 MYA	128 KYA	100 KYA	2	Variable	C
8	1 MYA	128 KYA	100 KYA	2	Variable	B
9	1 MYA	128 KYA	35 KYA	2	Variable	C
10	1 MYA	128 KYA	35 KYA	2	Variable	B
11	350 KYA	128 KYA	NA	0	Variable	B [†]
12	350 KYA	128 KYA	35 KYA	2	Variable	B
13	350 KYA	128 KYA	35 KYA	4	Variable	B
14	350 KYA	128 KYA	100 KYA	2	Variable	B
15	700 KYA	128 KYA	35 KYA	4	Variable	B
16	350 KYA	128 KYA	35 KYA	40	Variable	C
17	350 KYA	128 KYA	35 KYA	40	Variable	B
18a	700 KYA	350 KYA	50 KYA	4	Variable	B
18b	700 KYA	350 KYA [§]	50 KYA	4	Variable	B
19	700 KYA	350 KYA	NA	0	Variable	B
20	700 KYA	128 KYA	NA	0	Variable	D
21	700 KYA	128 KYA	50 KYA	2	Variable	D
22	700 KYA	128, 45 KYA [*]	NA	0	Variable	E
23	700 KYA	128, 45 KYA [*]	40 KYA	6	Variable	E

[†] Means that the growth was doubled for Africa.

[§] Signifies uneven gene-flow between Europe and Africa.

^{*} Signifies the Eurasian and Papuan divergence.

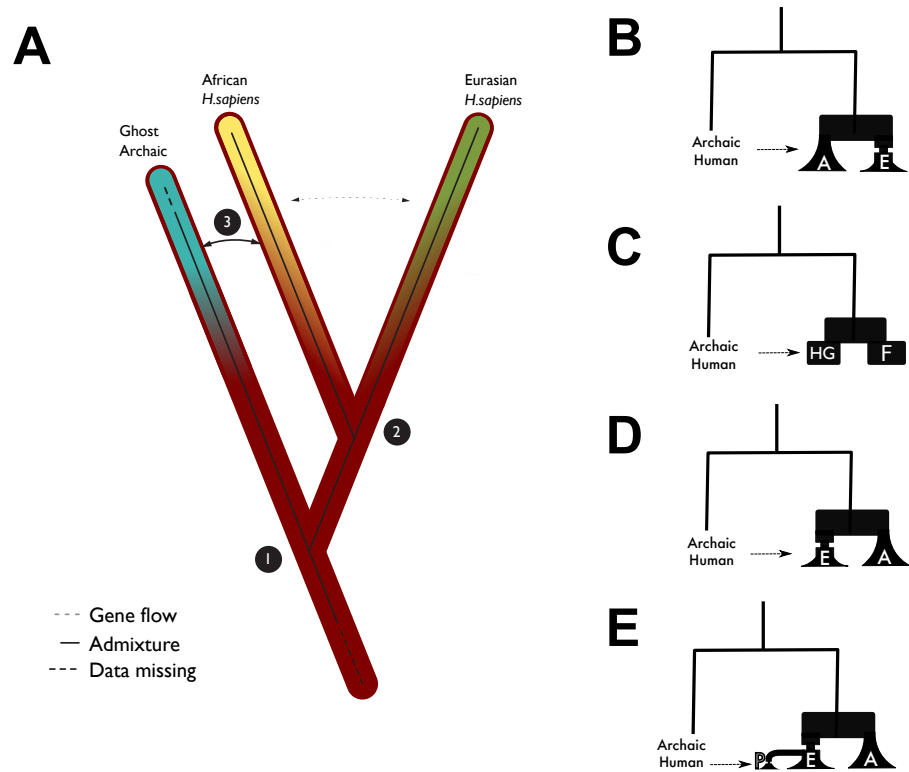


Fig. 4.3.: Demographic models used in simulations. The schematic of (A) the base population model showing the main points where parameters varied in each of the simulations ran. Referred to parameters can be found in Table 4.1. Parameter (1) specifies the date of divergence between humans and the archaic, (2) controls the split time between Africans [A] and Eurasians [E] and (3) dictates the date where admixture between an archaic and African humans occurred. Illustration based off of schematic from Galway-Witham & Stringer (2018) (Galway-Witham & Stringer, 2018). Schematics representing the (B-E) demographic events explored in the simulations. Table 4.1, under the "Demography" column, indicates which simulation was run with a simulated African population under (B) exponential growth and a Eurasian population with a bottleneck, (C) a hunter-gatherer [HG] African population and a farmer African population under logistic growth, (D) a model like (B) but where only Eurasians may receive archaic introgression and finally (E) a model where a simulated Papuan [P] population passes through two bottlenecks and may undergo archaic admixture.

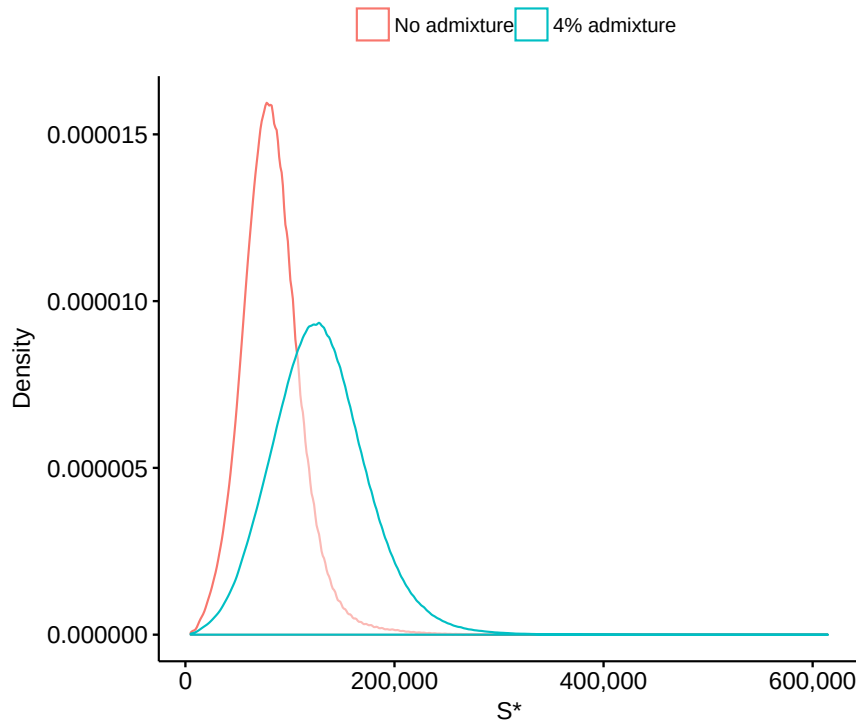


Fig. 4.4.: Admixture versus no admixture. Density plots displaying skew in the full distribution of S^* scores when comparing a population simulated under the null hypothesis (red) versus a population simulated under the null hypothesis with 4% admixture (blue).

I then proceeded by comparing all the other scenarios in a pairwise fashion by focusing specifically on two aspects: the overall S^* distribution and the top values of the S^* distribution. The way the top S^* candidates have been selected in African whole genomes varies across studies. For example, Lachance and colleagues (Lachance *et al.*, 2012) selected the top 0.5% of the distribution ranked by S^* score. Others have also determined a cut-off by evaluating the S^* p -values from a null distribution of simulations accounting for a putative demographic history without admixture (Hsieh *et al.*, 2016a; Vernot & Akey, 2015; Vernot *et al.*, 2016). By building a generalised linear model to the grid of S^* quantiles from these simulated whole-genome sets they estimated S^* percentiles for each putative introgressed haplotype. Putative introgressed haplotypes with a S^* score in the 99th percentile of the null simulations. This cut-off often corresponds to an absolute S^* score of about 100,000-150,000 in non-African populations with introgression (Vernot & Akey, 2015; Vernot *et al.*, 2016; Browning *et al.*, 2018).

Given this variability, I too used simulations to guide my decision in determining the cut-off to be used to identify the top S^* distributions. In the simulations where no admixture had been introduced, S^* scores would consistently remain below the absolute value of 500,000 irrespective of the demographic model explored, with the exponential growth being the only demographic model that boosted absolute values past 400,000, more than any other scenario without admixture (Figure 4.4). On this account, and

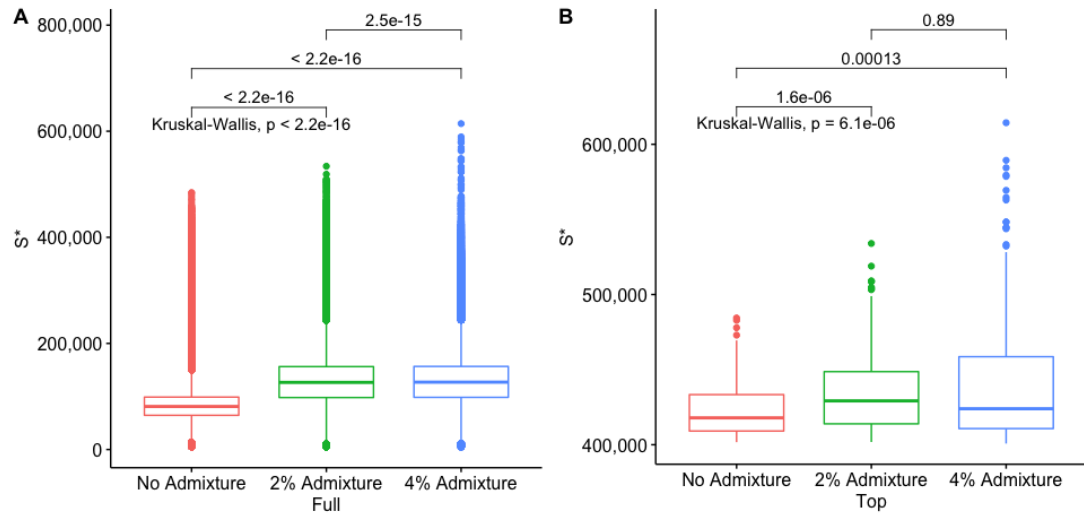


Fig. 4.5.: Admixture proportions. Pairwise comparison of simulations without admixture and with admixture in different percentages. (A) Full distribution of S^* scores and, (B) the top distribution of the S^* scores (> 400,000) when admixture is present.

similar to recent studies looking at archaic introgression in African populations (Hsieh *et al.*, 2016a), I set to use 400,000 as the cut-off for S^* scores when defining the top of the S^* distributions. In simulations where S^* scores do not reach the 400,000 value (e.g. simulations without exponential growth and/or without admixture), the full distribution is reported considering that when the cut-off is set to 100,000 for top candidates, the resulting distribution was not significantly different from the full distribution (i.e. from 100,000 more than $\frac{3}{4}$ of S^* detected candidates were still present).

What I observed across the different models was that the proportion of simulated admixture affected in a statistically significant way the full S^* distributions (Figure 4.5A). In fact a 4% introgression generated a significantly higher full S^* distribution than a 2% introgression. However, when comparing top S^* of simulations varying in the amounts of introgression, there were no statistically significant differences between distributions (Figure 4.5B). This suggests that S^* is sensitive to the presence of archaic admixture (Figure 4.5A and 5B) but less so to the proportion of admixture, particularly when looking at top detected candidates (Figure 4.5B). Although other studies have suggested that the skew observed in the tail of a S^* distribution is correlated with the extent of archaic introgression (Lachance *et al.*, 2012), this "tail" best fits the range of scores found within the full distribution of the simulations presented here rather than the top scores.

4.4.2 Recombination

In my attempt at exploring how different parameters can affect the outcome of the S^* statistics, I run a set of simulations that were variations of the null hypothesis

(Table 4.1) applying parameters first used by Plagnol & Wall (2006). A scenario with no admixture, in theory, should give low absolute scores yet S^* , being a linkage disequilibrium based method, will be sensitive to local mutation and recombination rates as well as other demographic events which can potentially mimic the signal of introgression even when none is present (Hsieh *et al.*, 2016a).

I first explored the changes in signal when having a constant recombination across the genome (Figure 4.6) versus a variable one based on the HapMap recombination map (The HapMap Consortium *et al.*, 2010). Like previous studies have pointed out (Hsieh *et al.*, 2016b), it is important to incorporate recombinational heterogeneity when simulating demographic null models. The skew corresponding to variation in recombination rate can be seen for both full and top S^* distributions. For this reason, I settled on using a variable recombination rate, as this is also most biologically lifelike (Myers *et al.*, 2005) and accounting for genomic recombination can help mitigate confounding factors when we look at real data.

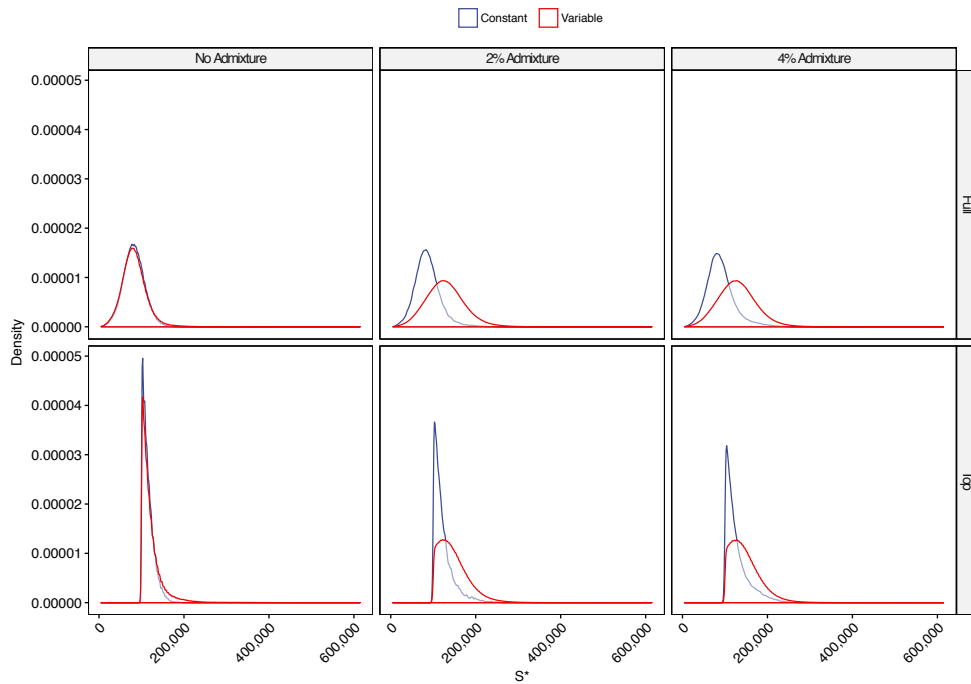


Fig. 4.6.: Constant versus variable recombination. Models with admixture and without admixture with either constant or variable recombination rate. (Above) Full distribution of S^* scores when admixture is present and, (below) the top distribution of the S^* scores (i.e. $> 100,000$ instead of $400,000$ to visualise differences in no admixture scenarios).

4.4.3 Demography

In my review of demographic inferences that explored archaic introgression outside Africa (Hammer *et al.*, 2011; Lachance *et al.*, 2012; Hsieh *et al.*, 2016a) and the history of hunter-gatherers and farmers in Africa (Kim *et al.*, 2014; Lopez *et al.*, 2018),

I extracted a combination of parameters to be used for the simulations of African populations (Figure 4.3 B-E; Table 4.1). We found that population growth in particular can greatly influence S^* (Figure 4.7). By simulating models where the growth rate was incremented (Sim 11) or decreased (Sim 3), it was possible to comparatively evaluate how statistically significant their S^* values were. Generally, it can be observed that by increasing recent population growth in a simulation one could mimic the signals observed in admixture (Figure 4.7). Notably, the feature that most stands out between a simulation with a recent growth expansion (blue) and one with a logistic growth (yellow) is the density peak at the low range of S^* scores when there is no recent growth (Figure 4.7).

A scenario where there has been recent exponential growth in a population would yield an excess of low-frequency alleles in comparison to the expectation under a constant population size because of the influx of new mutations introduced by new births. Population substructure, on the other hand, can have several different effects on the allele frequency spectrum, which may include an increase in both medium and high-frequency alleles (Allendorf, 1986; Kingman, 1982). Given the nature of S^* to focus on derived alleles, a higher absolute value of S^* may be more an indication of growth than of admixture. In Figure 4.6, we can observe that simulations with logistic growth, meant to reflect the density-dependent carrying capacity of a structured population, tends to give weak S^* values in comparison to the exponential growth scenarios.

In order to better explore how admixture and growth influence the behaviour of the S^* statistic, we compared to a demographic model with an exaggerated amount of admixture (e.g. 40%) but experiencing a logistic growth rate after African and Eurasian divergence (Sim 16). We can see that S^* candidates of a non-admixture scenario are significantly higher than that of an archaic admixture scenario with logistic growth (Figure 4.8). Furthermore, we could only evaluate the full distribution of S^* scores for these models because the logistic growth simulation (Sim 16) did not meet our 400,000 score cut-off to compare top S^* scores.

Under an exponential growth scenario, S^* scores tend to be inflated, suggesting that S^* is more sensitive to the variation produced under a growth model, with and without admixture. Admixture does not appear to drive S^* signals of putative candidates as much as exponential growth given that a 40% admixture event is unable to produce S^* scores 400,000 or greater if the growth is not exponential.

In scenarios where bottlenecks were modelled, I observed that S^* scores tended to be low overall. The scenarios simulated were meant to represent a Eurasian population (1-bottleneck) and a Papuan population (2-bottlenecks). Figure 4.9A and 9B demonstrate that bottleneck scenarios without admixture do not seem to visibly skew the full

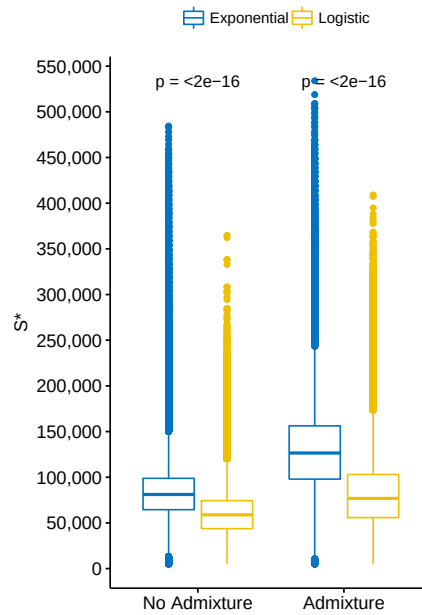


Fig. 4.7.: Exponential versus logistic growth. Pairwise comparisons of full distribution of S^* scores for simulations with exponential growth (blue) or logistic growth (yellow) in scenarios without (left) and with (right) admixture. P -values indicate how S^* distributions are statistically significantly different from each other when comparing growth models regardless of whether archaic introgression is present or absent.

distribution of S^* scores. However, we note that in a pairwise comparison of the two bottleneck scenarios without admixture (Figure 4.9C), a significant difference is observed between the full distributions with the 2-bottleneck scenario displaying a tail with higher S^* scores. Yet because these distributions without admixture do not meet the 400,000 absolute score cut-off, we did not compare the top scores. For scenarios where 2% and 6% admixture were introduced, respectively, S^* scores showed a visible skew in score (Figure 4.9A,B). This is particularly highlighted when comparing the full score distributions of simulated Europeans (1-bottleneck) and simulated Papuans (2-bottlenecks) with respective admixture proportions (Figure 4.9D). The top S^* scores for both populations undergoing admixture (Figure 4.9E, density plot) show contrasting distributions where the tail of a 1-bottleneck scenario with 2% admixture has a wider distribution than the 2-bottleneck scenario with 6% admixture.

The latter scenario displays a slight bimodal distribution amongst the top scores and the boxplots comparing the top S^* scores in Figure 4.8E show the 2-bottleneck scenario having a lower S^* median but with a few higher scoring regions in comparison to a 1-bottleneck scenario with 2% admixture. It is also noteworthy that although admixture proportion did not seem to make a significant difference for top candidates when observing a fixed demography with only changes in admixture contribution (Figure 4.9B), there is consistently a significant difference between full S^* distributions (Figure 4.9A). However, it is also crucial to point out that even without admixture present different

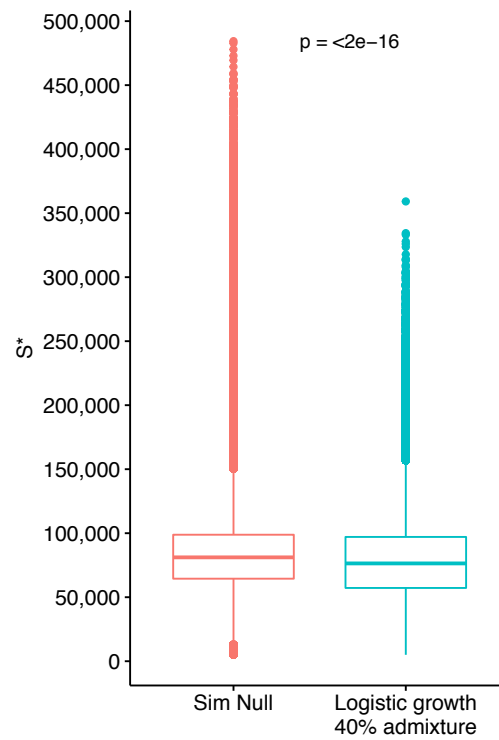


Fig. 4.8.: Admixture under logistic growth. Pairwise comparison of the full distribution of S^* scores for the null hypothesis model which contains no admixture (Sim Null) versus a simulation of a scenario with 40% of archaic admixture but with logistic growth in the target population (Sim 16).

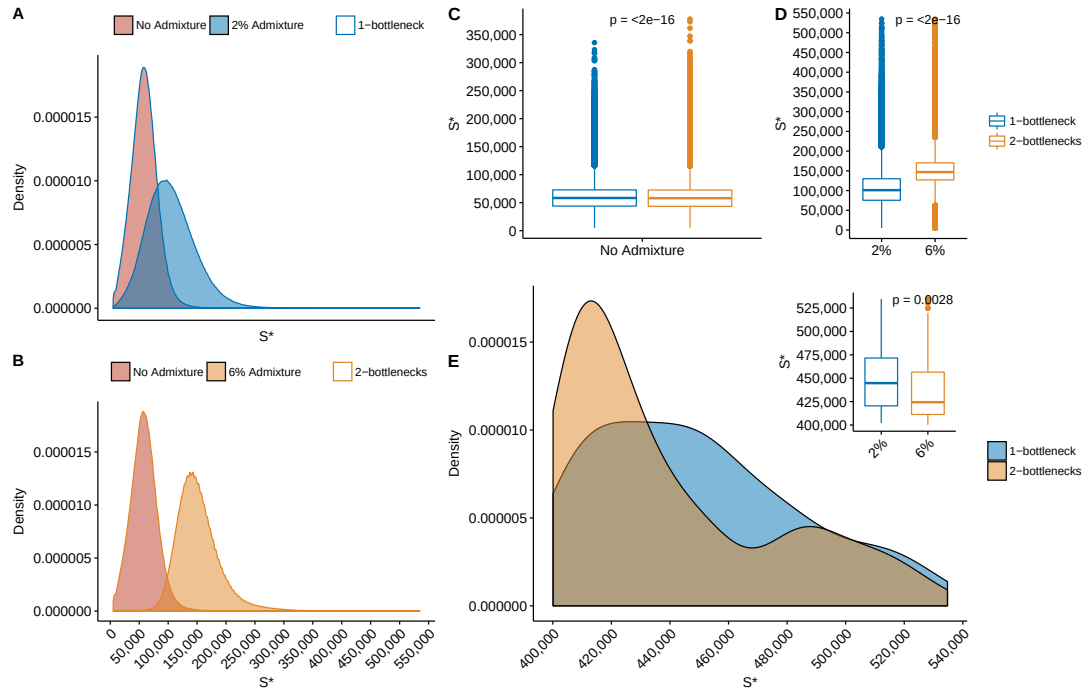


Fig. 4.9.: S^* scores for bottleneck scenarios with and without admixture. (A) Density plot of S^* scores for a 1-bottleneck scenario (Demographic model D) with and without admixture. (B) Density plot of S^* scores for a 2-bottleneck scenario (Demographic model E) with and without admixture. (C) S^* score pairwise comparison between both bottleneck scenarios in absence of admixture. (D) Pairwise comparison of all S^* scores for both bottleneck scenarios with admixture. (E) Bottom density plot displaying distributions for top S^* scores in both bottleneck scenarios with admixture. Upper right-hand boxplot showing a pairwise comparison for the top S^* scores with 2% admixture, 1-bottleneck and 6% admixture, 2-bottlenecks scenarios, respectively. Top scores can only be obtained for admixture scenarios.

demographic scenarios can present statistically different S^* distributions (Appendix Figure A.1A).

4.4.4 Divergence and Admixture Dates

Previous studies have pointed out that a higher FDR is observed for scenarios where there has been a recent divergence time with the introgressed archaic population (< 500,000 years ago) (Lachance *et al.*, 2012). This falls in line with the difficulty of distinguishing ancestral shared variation from admixture or gene-flow particularly when using LD-based methods that rely on an expectation of ancestral LD versus admixture LD (Theunert & Slatkin, 2017; Pickrell *et al.*, 2014; Moorjani *et al.*, 2011). Having already shown that top S^* scores are less sensitive to differing admixture proportions (Figure 4.5B), differences between the top scores of both bottleneck scenarios (Figure 4.8E) may be explained by dates of divergence with the archaic and the date of admixture.

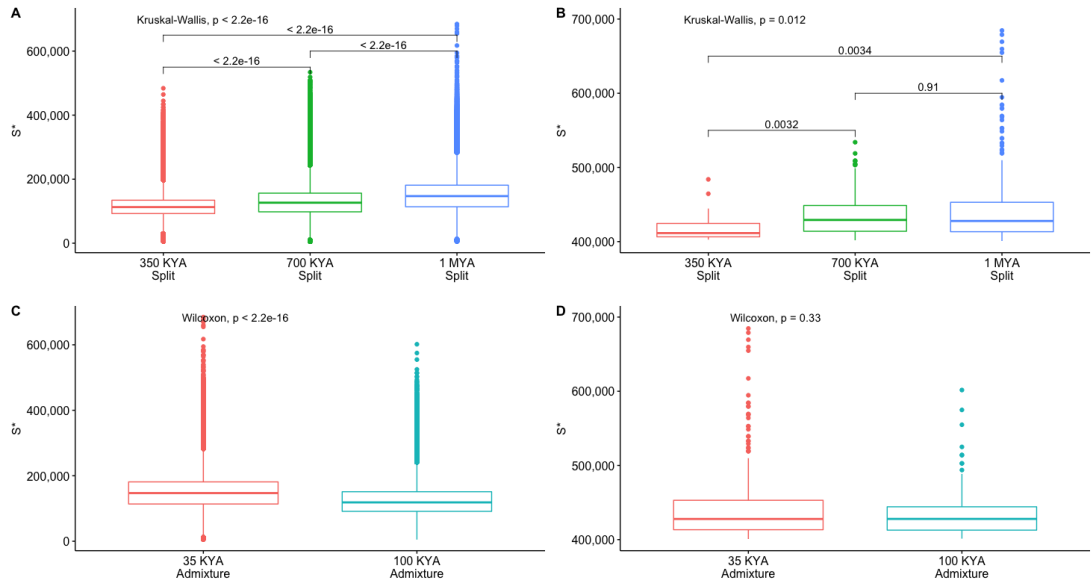


Fig. 4.10.: *Pairwise comparisons and multigroup comparison of simulations with differing split times and admixture times. (A) Full distribution of S^* scores and (B) the top distribution of the S^* scores ($> 400,000$) for scenarios where 2% admixture took place at 35 KYA but archaic and human divergence took place at 350 KYA, 700 KYA, and 1 MYA, respectively. (C) The full distribution of S^* scores and (D) the top distribution of the S^* scores for scenarios where splits took place 1 MYA but 2% admixture between humans and an archaic took place at different time points (35 KYA and 100 KYA).*

Notably, we observed that the divergence time between the archaic and humans generated the largest difference in both the full distribution of scores (Figure 4.9A) and the top scores of our simulations (Figure 4.9B). There was also a difference between the full distributions of admixture dates which were statistically significant (Figure 4.9C). On the other hand, the top of the distribution was not strongly influenced by the date of admixture (Figure 4.9D).

In light of these results, an explanation for the significant differences in the top S^* distributions for both bottleneck scenarios with admixture (Figure 4.9D-E) suggests that the number of bottlenecks and their respective dates may be a stronger influence in S^* than both admixture proportion (Figure 4.5) or date of admixture (Figure 4.10C-D).

4.4.5 TMRCA

Canonically, TMRCA has been used in other studies to support the hypothesis of archaic introgression by building on the expectation that genomic regions which are detected by S^* ought to have a wider TMRCA distribution than other random regions in the genome (Lachance *et al.*, 2012; Hsieh *et al.*, 2016a; Vernot *et al.*, 2016). This is consistent with the idea that some introgressed loci will be under selection and therefore display a younger TMRCA, whereas there will also be detected regions under genetic drift that exhibit older TMRCA culminating in a wide-joint distribution.

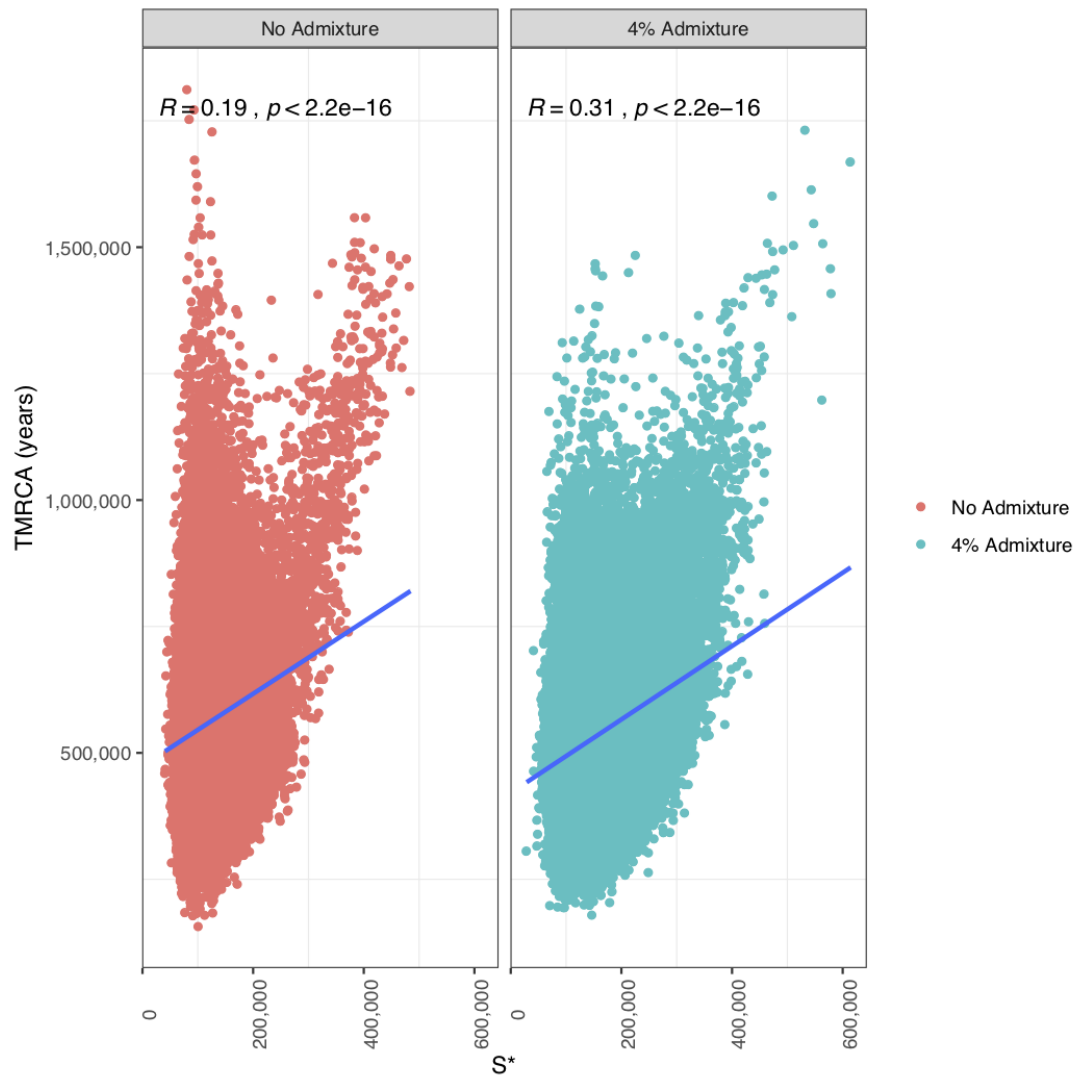


Fig. 4.11.: *TMRCA versus S^** . Positive correlation coefficient (R) between S^* and TMRCA (in years) in scenarios without admixture (left) and with 4% admixture (right).

Here I show that there is some correlation between S^* and TMRCA with the age of candidate regions incrementing slightly as S^* scores increased in both scenarios with no admixture (p -value < 0.01) and with 4% admixture (p -value < 0.01) (Figure 4.11). Notably, the no admixture scenario (Sim Null) displays the oldest S^* candidates as low scoring. In contrast, it can be observed that an admixture scenario (Sim 1) yields candidates with high S^* scores which have a stronger correlation with TMRCA.

In both full and top distributions of S^* , there are significant differences between scenarios with and without admixture for the TMRCA (Figure 4.12A) and absolute score (Figure 4.12B). TMRCA in a no admixture scenario demonstrates that a higher S^* score does not necessarily yield older haplotypes than a lower S^* score. This is particularly exemplified by the tails of the S^* candidate distributions (Top) which show that scenarios with and without admixture can yield similar high scoring regions

(Figure 4.12 B) but do not necessarily equate to being older in an admixture scenario (Figure 4.12 A). Nevertheless, it is worth stressing that the trend is not particularly strong in either scenario and both regressions explains little of the variability in the data (Figure 4.11).

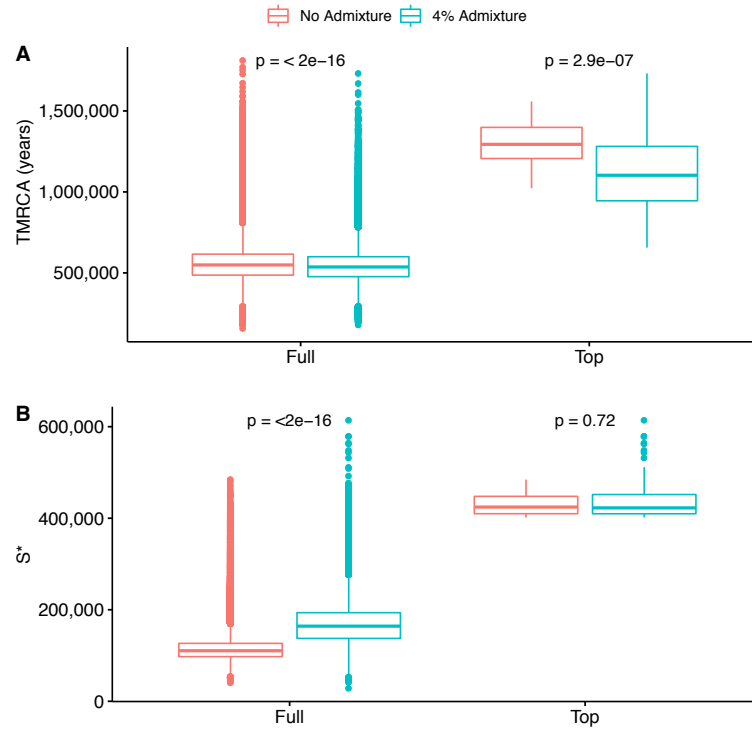


Fig. 4.12.: *Pairwise comparison of exponential growth simulations with admixture and without admixture for both full and top distributions of S^* candidates. (A) TMRCA in years of candidate regions and (B) S^* scores for candidate regions.*

As seen with S^* scores (Figure 4.7), TMRCA of S^* candidates fluctuates with other demographic events independent from admixture (Figure 4.13). In particular, the TMRCA of S^* candidates tend to be overall higher in a scenario where a population undergoes exponential growth compared to one that is limited by logistic growth regardless of the presence of admixture (Figure 4.13A). However, using a threshold of 400,000 for top S^* candidates in both growth scenarios with admixture suggests that top S^* will are not significantly different in TMRCA despite different forms of growth.

In addition to growth, bottlenecks can also influence S^* candidates with respect to TMRCA despite the absence of admixture (Figure 4.14A). Like with the difference in absolute score (Figure 4.8), TMRCA for S^* candidates of different bottleneck scenarios will be significantly different from each other. This is also observed in the TMRCA for top S^* candidates from scenarios differing in the number of bottlenecks and in admixture proportion (Figure 4.14). It has already been observed that S^* top candidates are not

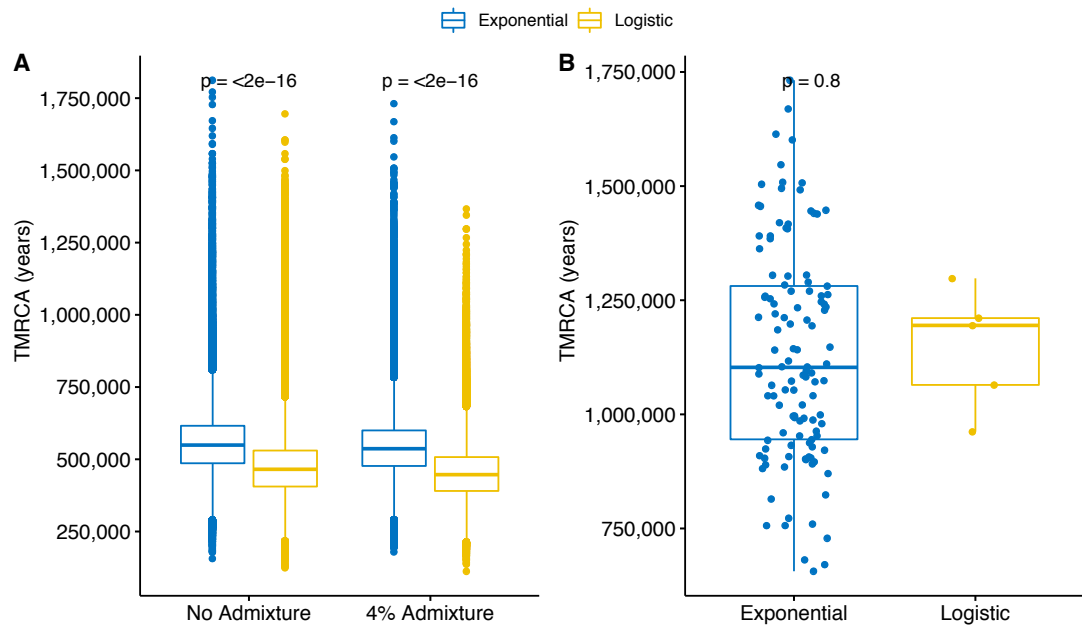


Fig. 4.13.: *Pairwise comparison of TMRCA in years for S^* for exponential (blue) and logistic (yellow) scenarios. (A) Comparison of the growth scenarios in absence of (left) and with admixture (right) for full S^* distribution. (B) Comparison of TMRCA in years for top S^* scores (> 400,000) when admixture is present. Under a no admixture scenario, top scores can only be obtained for exponential growth.*

strongly sensitive to admixture proportion score-wise (Figure 4.5B) and that differences in TMRCA among top S^* candidates can be misleading even when comparing models with admixture to one without admixture (Figure 4.12A).

4.5 Discussion

The decision to explore LD-based methods rather than the HMM or CRF approaches reviewed above stems from the understanding that the latter methods recover archaic introgression with more precision but only in possession of either 1) an archaic reference (Prufer *et al.*, 2014; Seguin-Orlando *et al.*, 2014; Sankararaman *et al.*, 2014; Sankararaman *et al.*, 2016) or 2) a clear understanding of the demographic assumptions (Seguin-Orlando *et al.*, 2014) or 3) large sample sizes (Skov *et al.*, 2018; Durvasula & Sankararaman, 2018).

However, when I explored how differences in demographic scenarios affect the performances of the S^* statistics, the results suggest that S^* is most likely sensitive to demographic perturbations of the allele frequency spectrum, in particular from bottlenecks followed by what type of growth a population has undergone (Figure 4.7, 4.8, 4.9). Recent exponential growth models consistently gave stronger values for S^* with or without archaic introgression, presumably because S^* scoring is more sensitive to the excess

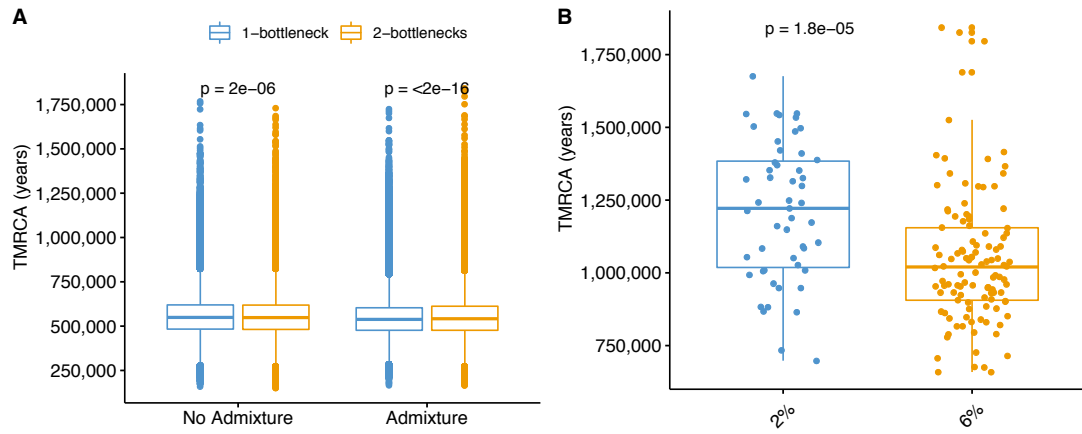


Fig. 4.14.: TMRCA in years for bottleneck scenarios with and without admixture. (A) Pairwise comparisons between the full distribution of S^* candidates for a 1-bottleneck scenario (Demographic model D) and a 2-bottleneck scenario (Demographic model E) with and without admixture. **(B)** Pairwise comparison of TMRCA in years for top S^* candidates for both scenarios with 2% admixture, 1-bottleneck and 6% admixture, 2-bottlenecks, respectively. Top scores could only be obtained for admixture scenarios.

low-frequency alleles in a population—a phenomenon also pertinent to archaically introgressed populations.

Similarly, a recent bottleneck scenario should also result in an excess of low-frequency alleles because of an increase in heterozygosity as a result of a reduction in the population size. Europe and Asia have had population bottlenecks accompanied by recovery which has led to a higher density of low-frequency alleles when compared to Africans (Lopez *et al.*, 2018). Both populations have experienced archaic introgression, which also increases low-frequency alleles and therefore measuring low-frequency alleles alone is prone to some degree of uncorrected detection of introgressed regions. Figure 4.7 and Figure A.1A suggest that demography alone is capable of skewing S^* scores.

Other studies (Theunert & Slatkin, 2017) have proposed looking at the site frequency spectrum (SFS) in combination with linkage disequilibrium to distinguish admixture from ancestral population structure. Theunert and Slatkin (Theunert & Slatkin, 2017) show that scenarios with ancestral structure consistently display an increase of both medium- and high-frequency alleles in comparison to recent archaic admixture scenarios. Specifically, they demonstrate that in a scenario with ancestral structure and no archaic admixture there should be no strong deviation between shared derived or ancestral alleles when comparing the ratio of linkage disequilibrium for two populations (i.e. when those alleles are conditioned on those of a third population). Moreover, the authors demonstrate that bottlenecks do not affect the decay of linkage disequilibrium when comparing to scenarios without a bottleneck despite a higher density in low-frequency alleles. To provide an example, if no admixture had taken place in non-Africans but there was ancestral structure at some point in time, then derived and

ancestral alleles of both Africans and non-Africans (i.e. those which are shared with Neanderthal alleles) should not deviate much from each other in a ratio comparing LD of both populations (e.g. in a ratio they should be close to 1).

This is harder to emulate in a scenario for Africa given that we do not know how many demes may have been present at the time when archaic admixture may have putatively occurred. Furthermore, previous studies (Eriksson & Manica, 2012; Theunert & Slatkin, 2017) have demonstrated that D -statistics, SFS, double conditioned frequency spectrum (dcfs), as well as linkage disequilibrium methods, have difficulties distinguishing between what is ancestral structure and what is admixture. This is particularly true in complex scenarios where there is asymmetrical gene flow between populations and when populations have strong sweeps of selection (Theunert & Slatkin, 2017). In trying to simulate the best fit demographic models available from previous studies (Yang *et al.*, 2014; Hsieh *et al.*, 2016b; Vernot & Akey, 2015; Plagnol & Wall, 2006), I was able to produce S^* distribution medians (Figure A.1B-f) that rank similarly to real data (Figure A.1C-f). However, there is a noticeable difference between these distributions which brings one to ask whether the lack of fine-scale detail in these models simulated limits studies that take model-based inference approaches to confidently assess the possibility of archaic admixture in Africa (Hsieh *et al.*, 2016a; Henn *et al.*, 2018).

This suggests that LD-based methods, while useful, are still limited when it comes to discriminating events within intermediate scenarios where both population structure and admixture may exist. Archaic references provide a way of conditioning LD-based methods in order to distinguish whether there is an excess of derived alleles shared with an archaic or if the amount can be attributed to ancestral structure. Contrary to the claim of earlier studies (Lachance *et al.*, 2012), our results suggest that in scenarios where there have been bottlenecks and/or recent exponential growth, LD-based methods like S^* are not robust enough and can provide confounding results. That is to say that S^* for a population with logistic growth and admixture will output weaker candidates than a population with exponential growth and no admixture. Moreover, we do not observe S^* sensitivity to the amount of introgression in top candidates. We also show (Figure 4.12) that S^* tends to pick up sequences with older TMRCA regardless of whether admixture is present or not. Moreover, TMRCA amongst top S^* scores can be misleading even when comparing scenarios with and without admixture (Figure 4.12A; Figure A.2A). Together this would suggest that TMRCA is not a favourable comparator for S^* when trying to evaluate admixture among different demographic histories (Figure A.2B-C).

The results from Figure 4.9C-D would suggest, that perhaps top S^* scores are not always the optimal candidates for introgression but rather other characteristics of the detected regions within the full distribution may be more telling of introgression. A number of other studies (Vernot & Akey, 2015; de Manuel *et al.*, 2016) attempt to resolve this by comparing with whole-genome simulation, or an archaic reference when available, to

calculate the statistical significance of an S^* score (Storey *et al.*, 2003; Williamson *et al.*, 2007). In this study, notably, it is shown that a threshold between S^* and TMRCA may be the best way to extract true introgressed candidates (Figure 4.11). However, with no other similar scenario to compare with or without a robust demographic inference for a population, using S^* on real data can be misleading and difficult to extract true positives from S^* detected candidates.

Other methods (Durvasula & Sankararaman, 2018) have used deep learning algorithms to examine the weights associated with archaic introgressed features, such as the number of private SNPs in a population followed by the skew of a distribution which is composed of pairwise differences between a focal haplotype and all other haplotypes. In a scenario of archaic admixture, a stronger negative skew is expected when comparing modern and archaic haplotypes whereas in a scenario with little archaic ancestry, the distribution should be skewed closer to 0. S^* as an absolute score, however, resulted to be of very small weight when examining important features when detecting archaic ancestry in this particular study. Overall, we demonstrate that S^* results are compromised when populations have undergone different demographic events whether it be number of bottlenecks or different forms of growth. In S^* what is interpreted as putative admixture signals in one population cannot be compared against another with a very different demographic history. This can render detected candidates misleading in absence of a clear understanding of the demographic history for the population in question or an archaic reference to compare with.

Searching for signals of archaic introgression in Africa

5.1 Background

Our knowledge of what may have occurred since the divergence of modern humans and other archaic forms within Africa is limited both in the fossil record and in availability of modern genomes. Fewer genetic studies have addressed early human history in Africa than other regions of the world, as Europe for example (Sirugo *et al.*, 2019; Popejoy & Fullerton, 2016; Nielsen *et al.*, 2017; Slatkin & Racimo, 2016), despite the amount of fossil and archaeological findings connected with the emergence of anatomically modern humans (AMH) in Africa (Campana *et al.*, 2013; Schlebusch & Jakobsson, 2018).

Although crucial for the story of human evolution, African whole genomes remain almost absent from both contemporary and ancient DNA datasets. An amalgamation of climatic, political and economic conditions has made it difficult to obtain both modern and ancient samples (Schlebusch & Jakobsson, 2018). Nevertheless, recent studies have made efforts to sequence ancient African samples to reconstruct African demographic history (Skoglund *et al.*, 2017; Gallego Llorente *et al.*, 2015; Schlebusch *et al.*, 2012). In relation to modern genomes, it has been pointed out that there are significant differences in structural variations when using a single human reference, such as the hg19 and GRCh38 builds. These builds are a conglomeration of the genomes of several individuals (Sherman *et al.*, 2019) and may hinder studies carried out on specific populations such as modern African populations which have been shown to have more structural variants—diversity that is often overlooked (Wang *et al.*, 2017; Sherman *et al.*, 2019).

In addition to a lack of the state-of-the-art tools suitable for studying human populations with larger genetic diversity as discussed in Chapter 4, there is also the bias of genomic research focusing on non-African populations, and therefore a relative disparity in number of genomes and populations sampled (Popejoy & Fullerton, 2016). Given the current situation it is not surprising then that the search for archaic admixture in humans has primarily concentrated outside of the African continent (Green *et al.*,

2010; Reich *et al.*, 2010). Canonically, African populations such as the Yoruba, have been consistently used as a proxy for a population having low or potentially no archaic introgression. While it may be indeed unlikely to observe archaic introgression from an archaic source very closely related to Neanderthal or Denisovans, there is no reason to believe that archaic introgression did not take place in Africa. Quite the contrary, some suggest a metapopulation scenario for Africa across time and space where admixture may have been a continuum across human (modern and archaic) populations (Povysil & Hochreiter, 2016) and that the recent out-of-Africa most likely consists of more than a single African origin for the founding population of non-Africans (Harding & McVean, 2004; Scerri *et al.*, 2018; Henn *et al.*, 2018). This would suggest that despite deep structure in Africa, there was ongoing gene flow between *Homo* populations across the continent. The fossil record in particular has shed light on the fact that archaic and pseudo-archaic or pseudo-modern forms were contemporaries of AMH as recently as 13 KYA (Harvati *et al.*, 2011; Tryon *et al.*, 2015).

5.1.1 The complicated demographic history of Africa

In all of the human evolution models discussed in the Introduction, processes within Africa have, for the most part, been left unrefined. *Homo sapiens* traits are assumed to have emerged across Africa but how this came to be remains unclear and, for the most part, unexplored. Despite its pivotal role in human evolution, population dynamics and history of *Homo sapiens* in Africa remain understudied. Recent investigations support a model where humans evolved from a highly structured population in Africa (Lahr & Foley, 1998; Scerri *et al.*, 2014; Scerri *et al.*, 2018; Skoglund *et al.*, 2017; Mazet *et al.*, 2016; Henn *et al.*, 2018) which point to a multiple origins model within Africa rather than the single region/population that was previously accepted (Henn *et al.*, 2011; Stringer, 2002; Schlebusch *et al.*, 2017; Schlebusch *et al.*, 2012).

Simulating various demographic scenarios has corroborated our understanding of the demographic history of populations (Gravel, 2012). In the case of African demographic history, modern genomes, albeit challenging, can still provide a window into the past. As such, there have been studies that simulate important events in early migrations such as the Bantu expansion and pastoralist movements (Marks *et al.*, 2015; González-Santos *et al.*, 2015; Tishkoff *et al.*, 2009; de Filippo *et al.*, 2012). This can assist our understanding of African pre-history in order to make more appropriate demographic assumptions about the populations on the continent. Henn *et al.* (2018) lay out distinct models of early modern human origins and corresponding support for them in the fields of morphology, archaeology and genetics. These models can provide guidance in model-testing and exploration of early demographic events. Nevertheless, without more archaeological data and more genomes, both ancient and modern, those explorations might overall be of limited scope and self-fulfilling, as the investigation of European

populations has revealed (Yang *et al.*, 2012; Fu *et al.*, 2015; Lazaridis *et al.*, 2014; Raghavan *et al.*, 2014; Fu *et al.*, 2016).

Population structure, migration, and admixture

Central for reconstructing human evolutionary history is clarifying the population structure of African populations. These ethnically diverse populations, some with substantial stratification connected to geography, show the highest levels of within-population genetic diversity among humans today (Tishkoff *et al.*, 2009). Those that speak languages belonging to the Bantu family are highly genetically homogenous which is attributed to the dispersal of sedentary farmers expanding out of west Africa starting from 4,000 BP and reaching southern Africa around 1,800 BP. This recent population expansion consisting of both the absorption and replacement of local indigenous communities of hunter-gatherers is also known as the Bantu Expansion. It is a crucial part of African demographic history which has created the strong genetic affinities observed in some African populations today but which also acts as a genetic veil limiting our understanding of sub-Saharan Africa before this event (Tishkoff *et al.*, 2009; Berniell-Lee *et al.*, 2009; Li *et al.*, 2014).

Of the African populations that remain most stratified both geographically and genetically are the central African Rainforest hunter gatherers (RHG, also referred to as Pygmies), the San populations of southern Africa (we note that several populations use the terms "Pygmy" and "San" when referring to themselves and such terms are used here without any derogatory connotation), and the Hadza hunter-gatherers from eastern Africa (Tishkoff *et al.*, 2009; Patin *et al.*, 2014; Schlebusch *et al.*, 2017; Pickrell *et al.*, 2012; Verdu *et al.*, 2009). Central African RHGs are divided by the Congo Basin, with isolated populations existing in the western and eastern parts (Verdu *et al.*, 2009). Data from several studies show support for central Africa being the place where the eastern and western branches of Bantu languages split suggesting that the Bantu expansion spread from western Africa through the central African rainforest and that this migration led to the split of the two language branches (Busby *et al.*, 2016; Patin *et al.*, 2017; de Filippo *et al.*, 2012). Recent research showed that the effective population size of central African RHG was as large as that of the ancestor of neighbouring agriculturalists for most of their evolutionary past (Lopez *et al.*, 2018). This suggests that central African hunter-gatherers had been more interconnected by gene flow than in the more recent past, much like what has been found in KhoeSan hunter-gatherers (Skoglund *et al.*, 2017; Kim *et al.*, 2014).

The San in particular have been shown to be the most divergent lineage of any other existing human population today (Schlebusch *et al.*, 2017) with a deep human population divergence time as far as 350,000-260,000 BP (Schlebusch *et al.*, 2012). This estimated divergence time coincides with the anatomical transition of archaic human

into modern human in the southern African fossil record such as Florisbad (Grun *et al.*, 1996) and the assemblages of Kathu Pan 1 (Porat *et al.*, 2010). Moreover, this time-frame overlaps with the contemporary existence of modern humans with other southern African hominids such as *Homo heidelbergensis* of Hoedjiekpunt (Berger & Parkington, 1995) and *Homo naledi* (Berger *et al.*, 2015). Whether these archaics provided any genetic contribution to modern human still remains unknown. However, Schlebusch and colleagues (Schlebusch *et al.*, 2012) suggest that the earliest diversification of modern humans involves ancestors of the Khoe-San in this epoch where archaic hominid and modern humans cohabited southern Africa. Consequently, the Khoe-San were used in the argument of a southern African origin (Henn *et al.*, 2011) given their high genetic diversity. Yet ancient remains from southern Stone Age hunter-gatherers and an individual from Ethiopia (Schlebusch *et al.*, 2012; Gallego Llorente *et al.*, 2015) have shown that 9-30% East African/Eurasian admixture around 1,500 -1,300 years BP may be responsible for both the elevated diversity and inflated past effective population sizes among present-day Khoe-San groups (Mazet *et al.*, 2016). However the deep split present among African populations has been confirmed by the analysis of Ancient Southern African HG and other African populations, the former showing no East African contributions (Skoglund *et al.*, 2017).

This East African admixture component present in all contemporary Khoe-San groups today is present at higher frequencies in Khoekhoe herders of southern Africa such as the Nama with as much as 23-30% admixture (Schlebusch *et al.*, 2012). Historical hunter-gatherers such as the Ju|'hoansi are the least admixed with only 9-14% East African/Eurasian admixture (Schlebusch *et al.*, 2012). Notably this admixture source was dated younger than 2,000 BP and appears to have been already admixed—69% East African and 31% Eurasian ancestry—supporting the estimate of Eurasian back-migration into East Africa around 3,000 year BP (Schlebusch *et al.*, 2012; Pickrell *et al.*, 2014; Pagani *et al.*, 2012; Skoglund *et al.*, 2017; Breton *et al.*, 2014). Together this would suggest that a migrating group with this admixed ancestry, most likely from East Africa, brought herding practices to southern Africa and was later assimilated by local southern African hunter-gatherer groups.

Eurasian gene flow appears to consist of several admixture events spreading from north to south along the Nile and Blue Nile (Hollfelder *et al.*, 2017) affecting most of northeastern and eastern Africa. This component has been suggest to be linked to the Middle East and the Arab migration (Hollfelder *et al.*, 2017). Interestingly, an ancient sample from a 3,100-BP individual of Luxmanda, Tanzania was shown to have about 38% ancestry related to farmers of the Levant and a remaining ancestry similar to the Mota individual from Ethiopia (~4,5000-year old) (Skoglund *et al.*, 2017). Additionally, the Sahel Belt (~3,000-1,000 BP) also acted as a corridor for human migration and genetic evidence suggests that there is a clinal differentiation between western and eastern Sahelian populations. Specifically, there appear to be strong signals of Eurasian

admixture into central and eastern (e.g. north Chadian populations) but not into western Sahelian populations (Triska *et al.*, 2015).

The Ari (Blacksmiths) and Hadza are two existing populations of local hunter-gatherers living in East Africa who have very limited contribution from possible back-migrations of farmers and herders from nearby African regions and Eurasia (Gallego Llorente *et al.*, 2015; Pagani *et al.*, 2012; Tishkoff *et al.*, 2009; Skoglund *et al.*, 2017). Although genetic signals of hunter-gatherer ancestry have been erased by migrations of herders and farmers, studies which have looked at ancient genomes (Skoglund *et al.*, 2017; Schlebusch *et al.*, 2012) suggest that a cline of relatedness existed along East Africa in a north-to-south direction. A genomic analysis comprising genomic material from ancient individuals from Southern and Eastern Africa, therefore lacking confounding genomic fragments derived by recent demographic events, revealed that a model depicting Southern Africa Khoe-San populations as basal to all the African populations is not fully supported, since the former show different relatedness to East and Western Africa (Skoglund *et al.*, 2017). Furthermore, different West African populations (Mende and Yoruba) show different relatedness to ancient Khoe-San, which is inconsistent with being derived from a homogeneous ancestral population which diverged from ancient southern Africans. This scenario would be compatible with either "archaic" admixture in Western Africans, affecting different populations heterogeneously, or with long term admixture which affected Western African groups with different magnitudes (Skoglund *et al.*, 2017; Montinaro & Capelli, 2018).

5.1.2 On the archaic side of things

In my review of bioinformatic tools used to identify archaic introgression, I have covered studies that have implemented such methods outside Africa. As mentioned in the previous chapter, these approaches rely on the use of an archaic reference (e.g. the genomes of Neanderthals and Denisovans) to filter out false positive candidates as well as identify false negatives (Chapter 4). Without using an archaic reference, the authors of those studies have collectively shown that methods like S^* and Sprime can identify around 30-60% of all true positive introgressed sequences at low false discovery rates. Simulated data for different demographic histories showed also that Sprime can identify archaic introgressed fragments, with an accuracy of 93% (Vernot & Akey, 2015; Vernot & Akey, 2015; Browning *et al.*, 2018). Similarly, Skov *et al.* (2018) have shown a low rate of false detection on both simulated and real data using an HMM method which did not rely on an archaic reference.

The sequencing of Eurasian Archaics has elucidated the evidence of admixture between archaic and modern humans which has been an invaluable contribution to understanding the events that took place outside of Africa. But this has not aided efforts in

answering whether archaic admixture happened in Africa despite the fact that *Homo* populations have lived longer alongside each other within than outside of Africa. An outstanding question in human evolution still remains: Did archaic admixture occur in Africa?

To be clear, when I refer to archaic admixture in Africa it is referring to recent admixture that has occurred between two populations that have been previously isolated for an extensive period of time—a deep divergence between the two as observed between Eurasian AMH with Neanderthal and Denisovans. While some studies suggest they have found evidence for archaic admixture in Africa (Wall *et al.*, 2009; Hammer *et al.*, 2011; Lachance *et al.*, 2012; Hsieh *et al.*, 2016a), few have considered similar work in context of other possible models such as admixture between archaic hominins now ancestral to AMH (Povysil & Hochreiter, 2016) or admixture with a now ghost (extinct) African population(s) of *Homo sapiens* (Skoglund *et al.*, 2017). Here I review the studies that have explored the possibility of archaic admixture in African populations and the claims made to date.

Early studies initially implemented the LD-based method S^* on African populations using genotype data for both coding and non-coding regions (Plagnol & Wall, 2006; Wall *et al.*, 2009; Hammer *et al.*, 2011). In particular, a study looking at 61 noncoding autosomal regions was able to infer about 2% of genetic material introgressed into contemporary Mandenka, Biaka, and San around 35 kya from an archaic population that split from the ancestors of modern humans approximately around the same time as Neanderthal and Denisovans, ~700 kya (Hammer *et al.*, 2011). The authors use an inferential approach to test whether the data they have for two hunter-gatherer populations (Biaka and San) and an agricultural population (Mandenka) fit a scenario of no admixture or low levels of admixture under two possible models: a two-population or a three-population model. Under a two-population model they test for archaic introgression by comparing S^* values from their data to values estimated using parameters under a no admixture scenario. In a complementary fashion, the authors also introduce three summary statistics for an approximate-likelihood method which consists in first identifying the two most divergent sequences for a locus and establishing two groups where the rest of the sequences will cluster to correspondingly. The protocol then calls for estimating the distribution of the fraction of shared polymorphisms between the two groups (D_1), the ratio of the number of differences between them (D_2), and finally the size of the smaller of the two groups (D_3). Each one is meant to represent the time of introgression, the time of the archaic-split, and finally the proportion of admixture, respectively. This distribution of summary statistics is calculated by simulating several ancestral recombination graphs (ARGs) (Griffiths & Marjoram, 1997). In using both S^* and these three summary statistics (D_1, D_2, D_3), the authors rejected the null hypothesis that an ancestral population with no admixture gave rise to anatomically modern humans. Instead both their inferential methods identify three exceptionally

long haplotypes (Table 1) at low-frequency among African hunter-gatherers (central African Pygmies and San), signalling these regions as putatively introgressed from an archaic population. Specifically, the authors suggest that central Africa may have been the place of origin for an extinct archaic hominin that admixed with modern humans in light of finding all three haplotypes among the Mbuti of Democratic Republic of Congo.

With the advances in sequencing technology whole-genome sequencing became accessible enough to sequence multiple individuals of several populations. Lachance *et al.* (2012) were the first to implement S^* on whole-genomes from African Hunter-Gatherer populations: Rainforest Hunter-Gatherers from Cameroon, click-speaking Hadza and Sandawe from Tanzania. Consistent with previous studies (Plagnol & Wall, 2006; Hammer *et al.*, 2011), they find evidence of archaic admixture in all three populations with candidate loci corresponding to a time to recent common ancestor comparable to those observed in Europeans from Neanderthal introgression. Through simulations, the authors came to the conclusion that S^* was robust enough to detect admixture and differences in amounts of introgression. When using non-Africans as proof of concept, they found that genomic regions with high S^* scores tended to coincide with Neanderthal-specific SNP enrichment. Based on this, the authors took the top $\sim 0.4\%$ of regions detected by S^* for each hunter-gatherer population and comparing them to all the other regions finding that the TMRCA of these top candidates were significantly older.

An ensuing study by Hsieh *et al.* (2016a) corroborated the evidence for archaic admixture in Africa by addressing the confounding effects of the demographic history of central African RHGs. S^* P-value distributions were calculated for simulations based on two models inferred from a previous study (Hsieh *et al.*, 2016b) that incorporated both isolation and gene flow with neighbouring farming populations. Four neutral whole-genome simulations, each model simulated with two different recombination maps (HapMap YRI and African-American [ASW]), were then compared with S^* P-value distributions from the observed data. The authors had originally estimated a false discovery rate (FDR) of 19% to 68% in top candidates when first testing their approach on the four neutral whole-genome simulation sets. Despite the large range in FDR, this approach represented an improvement to Lachance *et al.* (2012) by accounting for sequences that may have extreme S^* values but not statistically significant if one considers the effects of demography and biological processes such as mutation and recombination rates. Only the significant top 1% in this P-value distribution were chosen as candidate introgression loci, resulting in a total of 265 candidate loci covering ~ 20 Mb in length. From these putatively introgressed regions, Hsieh *et al.* (2016a), using a variant of D_3 from Hammer *et al.* (2011), inferred at least one admixture event with low amounts of introgression around 9,000 years ago.

In a recently released preprint, Durvasula & Sankararaman (2018) developed a machine learning method, ArchIE, which makes use of training datasets to calculate a set of features that are potentially informative of introgression. A prediction about archaic local ancestry can then be made for any given genomic window by using a binary logistic regression model based on the set of computed parameters. Results are then summarised for what is indicative of archaic admixture for each haplotype. The authors found that ArchIE weighs firstly the number of private SNPs followed by the skew of the distance vector in the underlying logistic regression model. The authors applied this method on 50 Yoruban individuals from the 1000 Genomes Project after training their logistic regression predictor using the parameters from a dataset with confirmed Neanderthal introgression in non-African populations. They implemented this under the assumption that the predictor is expected to be sensitive to introgression events from populations that shared ancestral population structure with Yorubans. Their results suggest that the archaic ancestry in Yoruba is best explained by admixture with an archaic ghost population more than the alternative scenarios of either the presence of Neanderthal ancestry from non-African back-migration or from admixture with a now extinct AMH population. In total, about 258 Mb of potentially introgressed sequences in the Yoruba were recovered using this method, which spanned several protein coding regions at high frequency in the population (Durvasula & Sankararaman, 2018).

In another recent preprint, Ragsdale and Gravel (Ragsdale & Gravel, 2018) explore classic statistics, as well as less familiar tests, to measure introgression. These tests were used to infer a demographic model with archaic introgression within a likelihood framework in the absence of an archaic reference genome. By implementing this approach of joint statistics on intergenic data from the 1000 Genomes Project, the authors found that the Luhya in Webuye, Kenya and the Yoruba of Ibadan, Nigeria exhibited approximately 6 and 8% archaic admixture, respectively. Moreover, this study shows that the commonly used model of human demographic history, derived from single-site allele frequency spectrum (AFS) and corroborated by LD decay curves, tends to fit the real data well but significantly underestimates the levels of LD among rare alleles. They show that by modelling archaic introgression worldwide, including African admixture with an archaic population that split off around 460-540 kya, this discrepancy in the levels of LD among rare alleles is resolved.

Despite these various set of results, the fine-scale understanding of the nature of these possibly recurrent admixture events in Africa still remains unresolved.

Back to the basics: The landscape of recombination and genetic variation in African populations

In the absence of aDNA from an archaic hominin or ghost population in Africa, we must rely on modern genomes to uncover whether archaic introgression took place.

Demographic history alone is not sufficient when using modern genomes as "windows" into the past. We must also take into account how biological processes affect the observed genomic diversity in order to develop tools that aid us in answering questions about our genetic history. Consequently, it is crucial to be aware of the differences in the genomic landscape of different populations. For example, recombination maps have been built for populations with European-ancestry (deCODE, HapMapCEU) (Kong *et al.*, 2010; The HapMap Consortium *et al.*, 2010), West African-ancestry (HapMapYRI) (The HapMap Consortium *et al.*, 2010) and mixed-ancestry (AAMap, AfAdm map) (Hinch *et al.*, 2011; Wegmann *et al.*, 2011). These studies have led to the conclusion that comparatively, West Africans have more recombination hotspots across the genome than Europeans and therefore crossovers are more evenly distributed leading to shorter LD distances in West Africans (Hinch *et al.*, 2011; The HapMap Consortium *et al.*, 2010). Moreover, these studies have been able to show that recombinations appear to be concentrated at hotspots which correlate with a particular ancestry (The HapMap Consortium *et al.*, 2010; Wegmann *et al.*, 2011); the activation of such hotspots being attributed to specific *PRDM9* variants (Berg *et al.*, 2011). Considering that African populations show the highest levels of genetic diversity in both between and within-population, it is essential to note that recombination landscape across Africa may differ and therefore LD-based methods developed to search for introgression in non-African populations must be adjusted for these differences. For example, if we wanted to use S^* to explore archaic introgression in West Africans, we must account for the shorter extent of LD and therefore expect shorter introgressed haplotypes than what is seen in Eurasians even if admixture putatively took place at a similar date. Mutation rate too has been shown to differ between populations and still remains without a final consensus (Narasimhan *et al.*, 2017; Ragsdale & Gravel, 2018).

5.2 Introduction

Methods that do not require an archaic reference can be used to exploit modern whole-genome data but still have their caveats such as high false positive rates. There have been but a few studies focusing on Africa in attempts to detect archaic sequences and potentially introgressed regions (Wall *et al.*, 2009; Hammer *et al.*, 2011; Lachance *et al.*, 2012; Hsieh *et al.*, 2016a; Durvasula & Sankararaman, 2018; Ragsdale & Gravel, 2018). Previous studies focusing on archaic admixture outside of Africa have consistently used Africa as a proxy population that did not experience "archaic introgression", which has essentially meant "did not experience contact with Eurasian Archaics" such as Neanderthal and Denisova (Green *et al.*, 2010; Reich *et al.*, 2010; Sankararaman *et al.*, 2014; Sankararaman *et al.*, 2016; Prufer *et al.*, 2014; Durvasula & Sankararaman, 2018). Those that have taken up the question of archaic introgression in Africa have had to do so in absence of an African archaic reference. Some of those studies have

used LD-based methods such as S^* to search for signals and have equated the identified regions, which display a fairly old time to recent common ancestor (TMRCA) and long haplotypes, as evidence of archaic introgression (Lachance *et al.*, 2012), although TMRCA does not always equate to population divergence when considering selection and complex demographic models (Henn *et al.*, 2018; Hsieh *et al.*, 2016a). Others have, similarly, run S^* on African hunter-gatherer populations and then chosen top candidates of archaic introgression by comparing against simulations with a set of demographic assumptions (Hsieh *et al.*, 2016a) to account for confounding effects on TMRCA due to selection. Nevertheless, both of these studies emphasise the importance to better characterise the nature of admixture in Africa and, more generally, the demographic history of early African populations.

In order to investigate the validity of the reports of archaic introgression in African populations I explored S^* and Sprime on African individuals from the Simons Genome Diversity Project (SGDP) (Mallick *et al.*, 2016), a dataset recently made available and which has not been systematically investigated yet for the presence of archaic introgression in Africa. In searching for signatures of archaic introgression, I have taken in consideration what I have learned about the behaviour of S^* under different demographic scenarios (Chapter 4). In addition, I have investigated the effect of sample size on the detection of introgressed regions.

5.3 Materials and Methods

5.3.1 Whole genome sequencing data from SGDP

Papuan data were obtained from publically available Simons Genome Diversity Project. Variant calling and phasing were all performed by Mallick and colleagues (Mallick *et al.*, 2016). We downloaded phased variants PS2 which declared to be phased using SHAPEIT (Delaneau *et al.*, 2012). These variants were phased without a reference panel with all samples together and it was suggested by Reich and colleagues to have better phased than PS3 (Mallick *et al.*, 2016). We chose to use PS2 instead of PS1 because of our interest in populations that are not available in the 1000 Genomes (Table 1) which was used as a reference panel to improve phasing in PS1 (Mallick *et al.*, 2016). All pairwise comparisons (Wilcoxon tests) reported were Bonferroni-corrected. All multiple group comparisons were one-way ANOVAs (Kruskal-Wallis tests).

5.3.2 Performing S^* and Sprime

We downloaded S^* software package from <https://github.com/bvernot/freezing-archer> and Sprime from <https://faculty.washington.edu/browning/sprime.html>. HapMap YRI and African Enriched recombination maps were retrieved from https://www.well.ox.ac.uk/~anjali/AAmap/maps_b37.tar.gz.

5.3.3 Papuans

On Papuan whole-genomes from SGDP I performed both S^* and Sprime with similar protocol to Browning *et al.* (2018). For example, we have performed S^* and Sprime with Papuans as the target population but as our outgroup have used all Africans (44 individuals from 19 populations) from the SGDP panel instead of the 107 Yorubans from the 1000 Genomes Project. S^* was calculated on each individual in 50 Kb windows with a step size of 10. For Sprime calculations I used the default settings for the mutation rate and used the HapMap (The HapMap Consortium *et al.*, 2010) genetic map. We performed both S^* and Sprime on different sample sets (i.e. 2, 6, 15) to observe how both methods perform in detecting candidates when there are differences in sample size. We computed candidate detection overlaps by finding overlapping SNPs identified by S^* , Sprime, and what was confirmed by Li *et al.* (2013) to be introgressed.

5.3.4 African populations

African individuals were subset into different groups in order to maintain comparable samples sizes ($N=6$). SGDP individuals per grouping can be found in Table S5.1. I performed S^* and Sprime on each African grouping using 42 Eurasians as an outgroup (Table S5.2). S^* was calculated on each individual in 50-kb windows with a step size of 10. Sprime was performed with default settings for mutation rate and using two different genetic maps, HapMapYRI (The HapMap Consortium *et al.*, 2010) and African Enriched (Hinch *et al.*, 2011). Top candidates were chosen based on the threshold of 400,000 for both S^* and Sprime. Sprime detected candidates were primarily used for downstream analyses such as detecting shared and unique haplotype regions and when marking overlap detection for candidate loci of previous studies. Candidate loci for archaic introgression in Africa from previous studies and this study can be found uploaded on <http://github.com/santaci/archaia/>.

5.3.5 TMRCA estimates

TMRCA calculations were performed on Papuans for different sub-samples on the whole genome, S^* top candidates and Sprime top candidates. TMRCA estimates for African groupings were performed on positions calculated by S^* and Sprime (HapMap YRI & AE maps). Estimates were calculated using the Hudson method (Hudson, 2007). This method computes the average coalescent time from derived sites assuming that mutations have occurred under a Poisson distribution. From that value we estimate TMRCA in years by using mutation rate in years based on pedigree studies (Scally & Durbin, 2012). Sites were considered ancestral or derived using the Ancestral Allele information retrieved from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/. The scripts produced to calculate TMRCA can be found on <http://github.com/santaci/archaiia/>.

5.3.6 Comparison with Denisovan and Neanderthal genomes

D-statistics

D-statistics were computed on all 44 African individuals from the SGDP, 1 English, and 1 Papuan individual. The reconstructed common ancestor (RCA), referred to as Ancestral in this study was used as the outgroup. Altai Neanderthal, Vindija Neanderthal and Altai Denisovan WGS variants in VCF format were retrieved from the Max-Planck repository. The python software POPSTATS (Skoglund *et al.*, 2015) was used to perform D-statistics: <https://github.com/pontussk/popstats>. D-statistics is a type of f_4 -statistics which tests whether two pairs of populations are symmetric in relation to one another and quantifies if any resulting asymmetry is due to admixture (Green *et al.*, 2010). Moreover, it is meant to normalise for heterozygosity. Briefly, if two samples P_1 and P_2 are consistent with descending from a common ancestral population, then in a biallelic site it is expected for them to exist in equal frequency. In the set of $P_1, P_2, P_3, Ancestral$ the ancestral allele is "A" and the derived allele is denoted as "B" when observed in P_3 . P_3 denotes an archaic sample - either Denisova, Altai Neanderthal or Vindija Neanderthal. Patterns of ABBA or BABA are the ones expected in equal frequency. A formal test for this consists of counting ABBA and BABA sites across n base pairs in the genome where all samples are aligned. The counting is normalised by the total number of observations where $C_{ABBA}(i)$ and $C_{BABA}(i)$ are indicator variables existing as 0 or 1 depending on which pattern is observed at base i .

$$D(P_1, P_2, P_3, Ancestral) = \frac{\sum_{i=1}^n [C_{ABBA}(i) - C_{BABA}(i)]}{\sum_{i=1}^n [C_{ABBA}(i) + C_{BABA}(i)]} \quad (5.1)$$

5.4 Calibrating detection of archaic introgression

5.4.1 When Size Matters: Sample size in Papuans

Simulations are undoubtedly powerful tools to explore unknown demographic scenarios and can provide a comparative framework for putative results in order to make sense of data we have. Similarly, we can make use of information on how S^* behaves in "real" modern genomes which we know have experienced introgression. Such genomes can provide insights on how experimental design, in addition to demographic and biological factors, can also affect results.

Present-day Papuans and Melanesians have inherited 5% of their genome from Denisovans (Reich *et al.*, 2010) and previous studies have been able to recover archaic segments from modern Oceanic genomes (Vernot *et al.*, 2016). I used previous work (Vernot *et al.*, 2016; Browning *et al.*, 2018) done on Papuans as proxy to understand how S^* behaves when there is true introgression. The Simons Genome Diversity Project (Mallick *et al.*, 2016) (SGDP) has whole genomes for 15 Papuan individuals which I used in determining how sample size impacts S^* scores. Some populations in genome diversity sets like SGDP (Mallick *et al.*, 2016) and The 1000 Genomes Project Consortium (2015) (1KG) are better represented than others and therefore understanding the stochasticity of sample size differences is important when evaluating a method that relies on variant density. Using real data that has already been validated for archaic introgression is crucial when it comes to confirming candidates detected by S^* whilst estimating the false positives among them. This can later be used to determine how, and if, I can extrapolate S^* behaviour when detecting introgression in other populations. Furthermore, I also use these samples to oversee the overlaps between S^* and an enhanced version of S^* , Sprime (Browning *et al.*, 2018) to compare detection on an individual-by-individual basis (S^*) versus on a population basis (Sprime).

S^* and Sprime

During the time this study was in progress, Browning and colleagues (Browning *et al.*, 2018) published a study introducing a new S^* -like method, Sprime. Sprime was able to detect two pulses of archaic Denisovan admixture in Papuans from the SGDP (Mallick *et al.*, 2016). The new approach was developed for increased power and was shown to give higher accuracy than S^* in their study. I compared both methods on the same very dataset as they did to characterise how sample size affected S^* and Sprime scores. Since the 2016 S^* was meant to target individuals and used only two haplotypes for determining linkage disequilibrium when producing a score, it had the advantage of not requiring large sample sizes. Nevertheless, Browning *et al.* (2018) demonstrate that

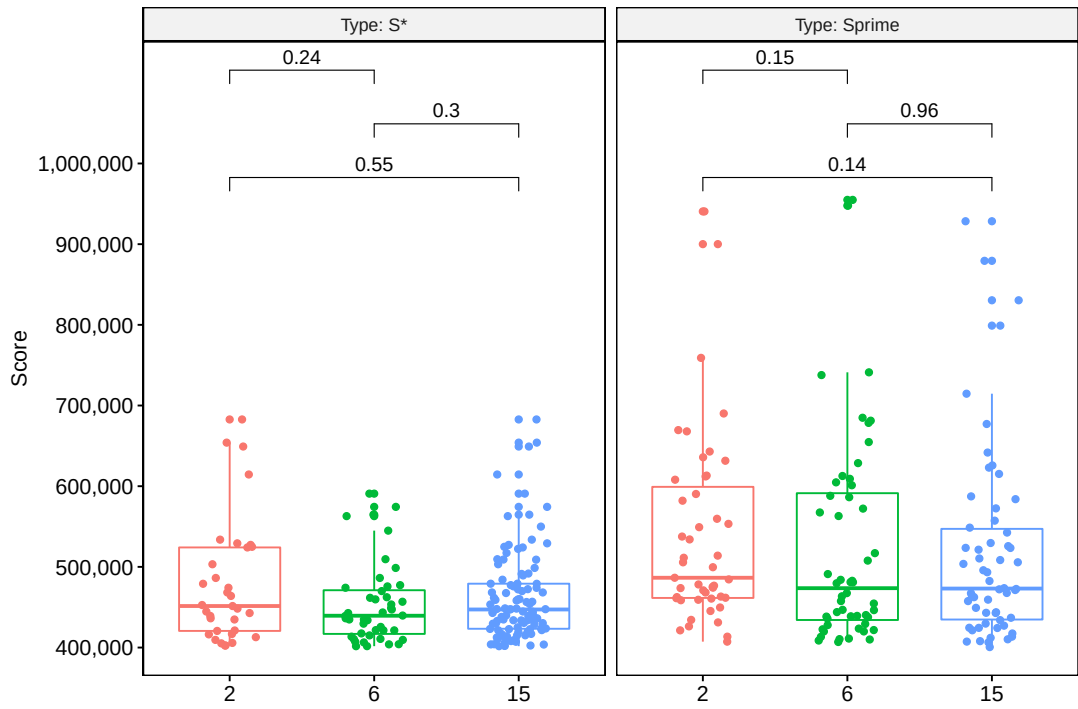


Fig. 5.1.: Score and sample size. Pairwise comparisons between different sample sizes for top S^* and Sprime candidates. Top candidates are haplotypes that surpass the 400,000 score threshold.

with more sample power and by avoiding windowing Sprime is able to detect accurately at a higher frequency across larger regions than S^* .

We implemented Sprime on Papuans in a similar fashion to the Browning *et al.* (2018) analysis (see 5.3). I compared S^* to Sprime and searched for overlaps in the scoring scheme, detection of false positives, and contrast how the two approaches perform with different sample sizes from the same population in regards to scoring, TMRCA, and putative fragment length.

Sample size

I decided to focus specifically on the top candidates detected by S^* and Sprime, respectively, based on what was observed in the simulations performed in Chapter 4 and what has been done in other studies (Hsieh *et al.*, 2016a; Hsieh *et al.*, 2016b). Because S^* scores on an individual basis, sample size does not play a vital role in absolute scoring. However introducing more individuals can change the top candidates by presenting more high scoring loci (i.e. loci with strong linkage disequilibrium). Figure 5.2 exhibits that neither S^* or Sprime display a statistically significant difference between subsamples among the values of top-scoring candidate regions. In line with previous investigations, I further supplemented my observations by including the expected time to most recent common ancestor (TMRCA) of top-scoring S^* and Sprime

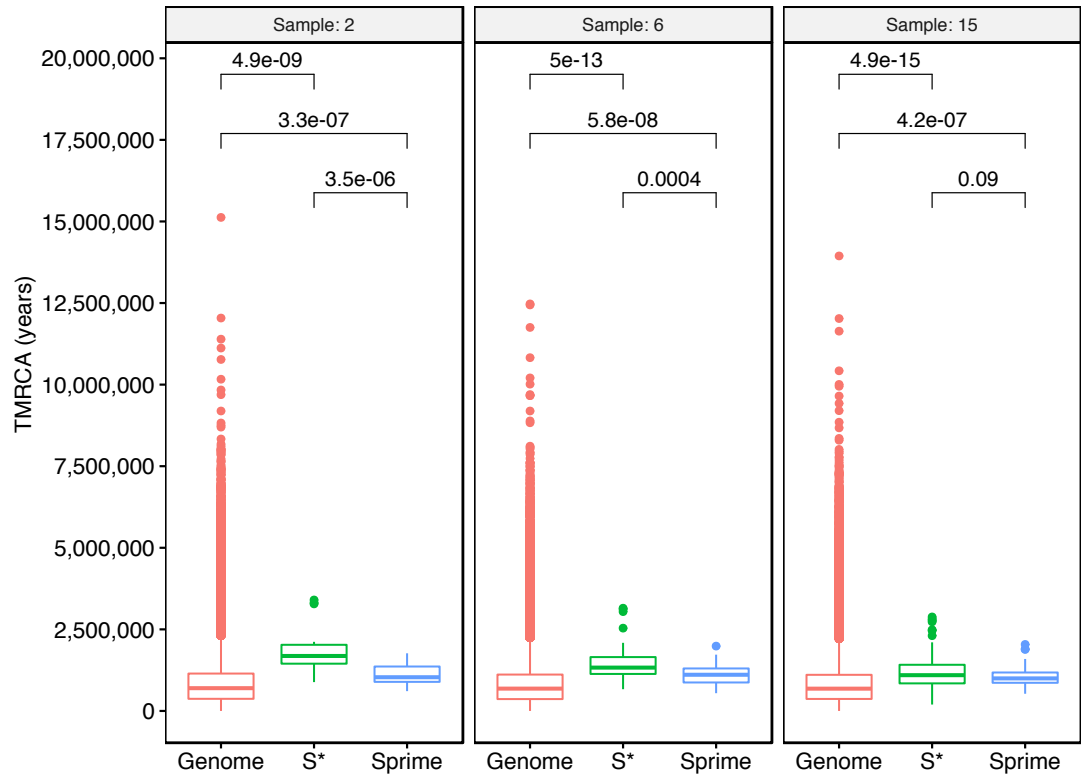


Fig. 5.2.: TMRCA, S^* , Sprime and whole genome. Pairwise comparison of the time to most recent common ancestor between whole genome, top candidates of S^* and Sprime segments across different sample sizes.

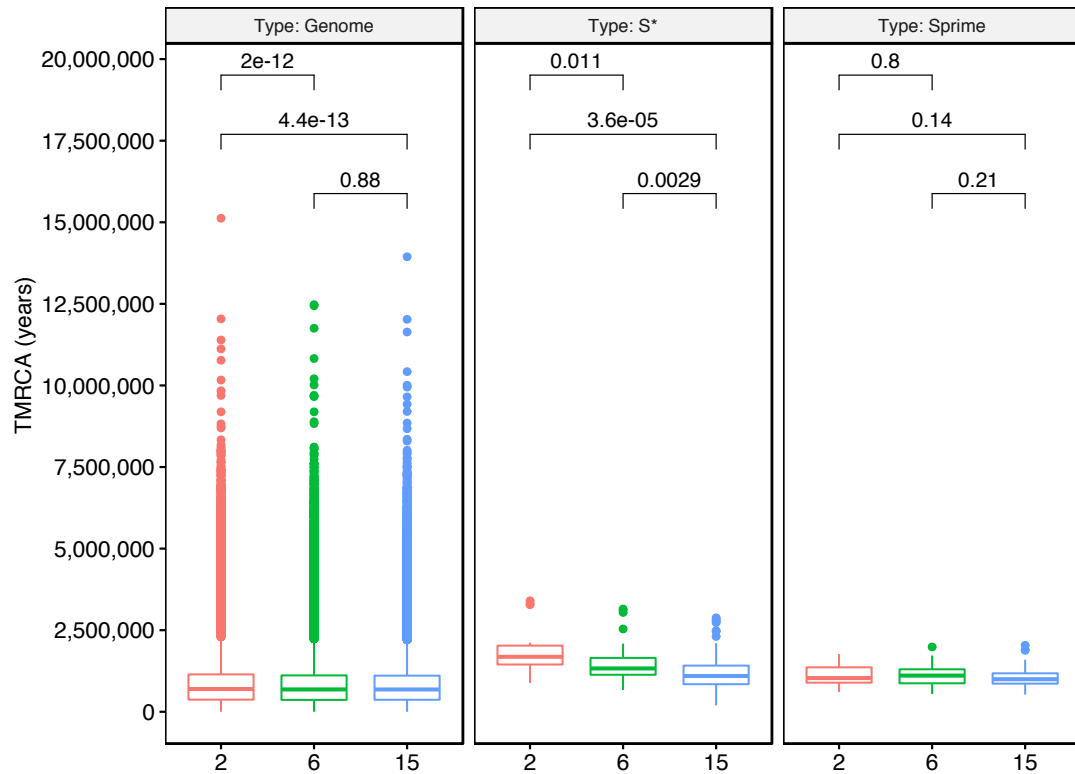


Fig. 5.3.: TMRCA and sample size. Pairwise comparison of the time to most recent common ancestor between different sample sizes for estimates across the genome and top candidates detected by S^* and Sprime.

fragments as well as for the whole genome in different subsamples (e.g. 2, 6, 15). Generally the TMRCA differences between the whole genome and top S^* regions were statistically significant as were the differences between the whole genome and Sprime regions (Figure 5.2). Only in a sample size of 15 were top S^* and top Sprime candidate ages similar to each other ($p=0.09$) (Figure 5.2). All subsamples for whole genome and for S^* top candidate ages were significantly different from each other. In contrast, Sprime had similar aged haplotypes across all subsamples (Figure 5.3). On the other hand, top Sprime candidate scores and the whole genome TMRCA revealed to be closer when comparing samples sizes 2 and 6 but significantly different from each other when the sample size was increased to 15 (Figure 5.1 , Figure 5.2). Part of the higher scoring of Sprime versus S^* can be due to the larger haplotypes that Sprime detects, as S^* is run on windows of fixed length, in this case 50-kb long.

Characterising false positives

We extracted the overlapping SNPs between S^* and Sprime and compared them to those found by Browning *et al.* (2018) and found that by increasing sample size more candidates detected in the original study could be recovered. Regardless of sample size, Sprime detected consistently a subset of S^* candidates. Notably the top 1% of

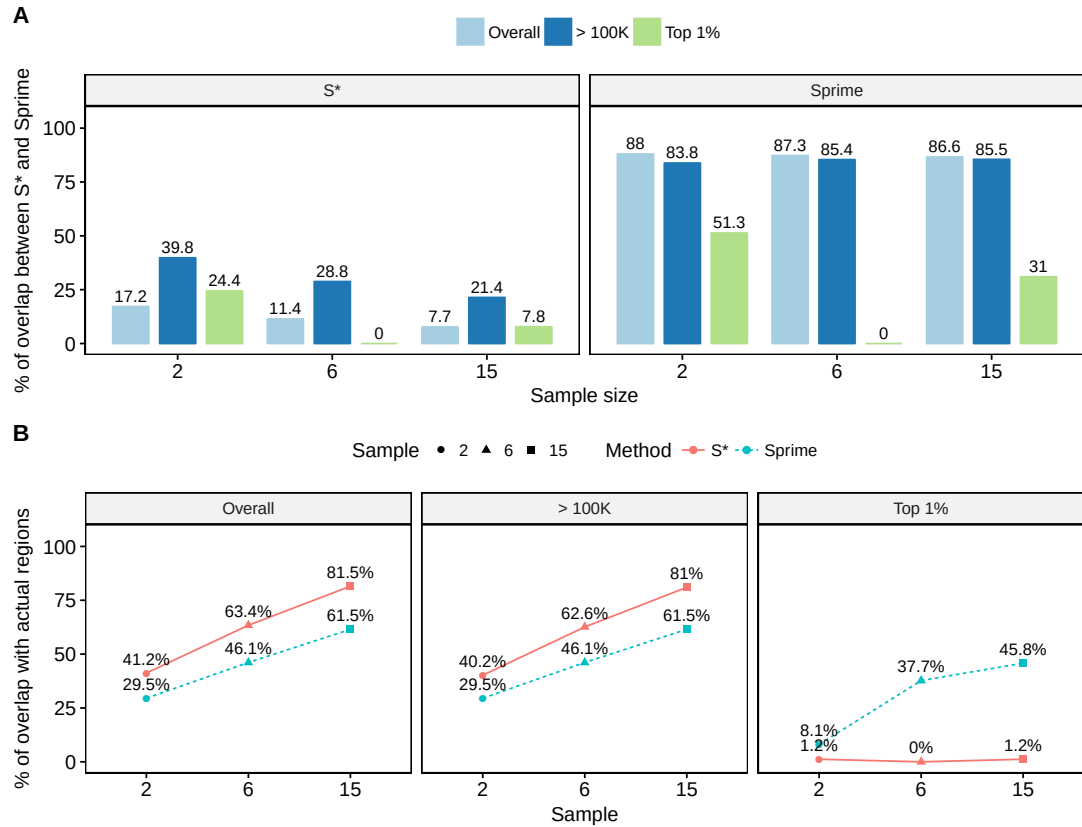


Fig. 5.4.: Comparison of S^* and Sprime on Papuan WGS from SGDP Data. (A) Overlap between S^* and Sprime for different sample sizes in overall candidates, candidates that were above the score of 100,000, and in top candidates. Example: In a sample size of 2, 17.2% of S^* region overlap with Sprime detected regions overall and conversely, 88% of Sprime regions overlap with regions detected by S^* overall. (B) Overlapped with regions retrieved by Browning *et al.* (2018); Overall: all regions recovered; >100k: regions with an absolute score higher than 100,000; Top 1%: top 1% scoring fragments.

the identified regions for each respective method has limited overlap (Figure 5.4 A). However, detection using different number of subsamples showed a relative increment of almost 20% in detecting true archaic regions overall (Figure 5.4 B) yet among top 1% candidates, the incremental difference of finding true archaic regions was relatively limited.

In my analyses using S^* and Sprime I was unable to recover all the introgressed haplotypes detected by Browning *et al.* (2018). One possible explanation is that Browning and colleagues used, as an outgroup, Yoruba individuals from the 1KGP (n=108) while I used all African individuals (n=44) from the SGDP. It is therefore possible that, due to back migration, some of the African populations I have included, such as East Africans and North Africans, have Neanderthal introgression (Gallego Llorente *et al.*, 2015; Pagani *et al.*, 2012; Skoglund *et al.*, 2017; Schlebusch *et al.*, 2012) and therefore those regions have not been detected in my Sprime analysis (Figure S7.4).

The behaviour of true positive candidates in relation to Sprime scores and TMRCA estimates can be seen in the correlation plot for Papuans (Figure 5.5). Introgressed haplotypes have been found to be dispersed in both Sprime scores and TMRCA (Figure 5.5 - Introgressed) whereas regions that had no match with either archaic reference tended to be low scoring in Sprime (< 400,000) with a negative correlation to TMRCA (Figure 5.5 - No Match) than truly introgressed regions, although not statistically significant. Given the results shown in Figure 5.5, an older TMRCA does not seem to be indicative of introgression (Figure 5.5; left) and can in fact be misleading when a detected candidate is not introgressed (Figure 5.5; right) as there appears to be TMRCA overlap among introgressed and non-introgressed regions. In line with simulations in Chapter 4 and results in Figure 5.5, I used an Sprime cutoff score of 400,000 since this score, while losing true signals of introgression, seems more likely to render true positive candidates than using a TMRCA threshold.

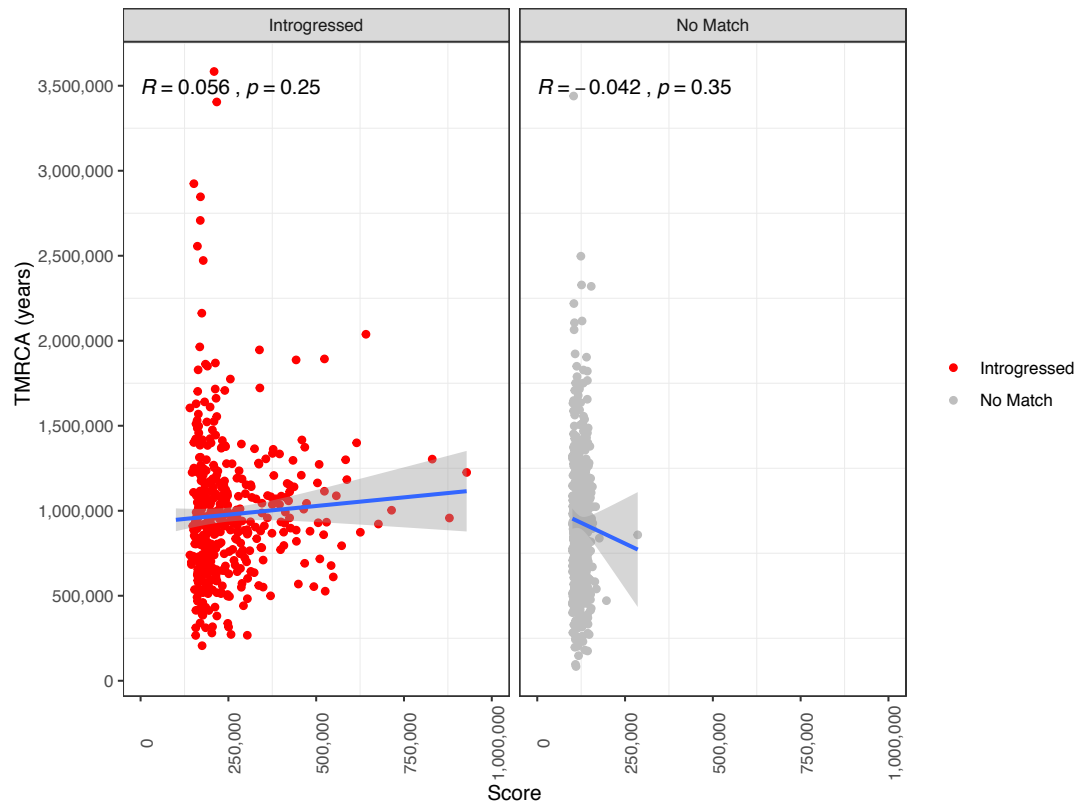


Fig. 5.5.: TMRCA versus Sprime in Papuans. Correlation coefficient (R) between Sprime score and TMRCA in Papuans for detected haplotypes which have been confirmed to be introgressed from Neanderthal and/or Denisova (left) and candidates which were found to have no match with either archaic reference (right).

We did not correlate S^* matches with TMRCA given that Figure 5.4 A explains that true introgressed regions detected by Sprime are already a more accurate subset of S^* .

5.5 Detection of putative archaic introgression in Africans

Forty-four high-coverage whole genome sequencing data for 19 African populations from the SGDP Dataset were analysed (Table 1). I performed S^* and Sprime on these samples to detect putatively introgressed regions per individual (S^*) and population or per defined geographical grouping (Sprime). African populations were grouped based on genetic similarity demonstrated from earlier studies (Tishkoff *et al.*, 2009; Montinaro & Capelli, 2018) in order to maintain comparable sample sizes ($N=6$). Individuals included in the subset are listed in Table S7.1. For comparison purposes some analyses did not include the groupings North and South since they included fewer than 6 individuals each.

Tab. 5.1.: Grouping of the 44 African individuals from the Simons Genome Diversity Project.

Group	Populations	Individuals
Pygmies	Mbuti	4
	Biaka	2
	<i>N</i>	6
San	Khomani	2
	Ju 'hoansi	4
	<i>N</i>	6
West	Yoruba	3
	Gambian	2
	Mandenka	3
	Mende	2
	Esan	2
	<i>N</i>	12
East	Luhya	2
	Masai	2
	Dinka	3
	Luo	2
	BantuKenya	2
	Somali	1
	<i>N</i>	12
South	BantuHerero	2
	BantuTswana	2
	<i>N</i>	4
North	Saharawi	2
	Mozabite	2
	<i>N</i>	4

I initially investigated how the distribution of top S^* scoring fragments compare across African populations on all the available samples as grouped in Table 1, including the Papuans as reference for a population with archaic introgression (Figure 5.6). Although the distributions were not significantly different, a visual inspection of Figure 5.6 suggests that S^* candidates for all African populations seem not only to be higher

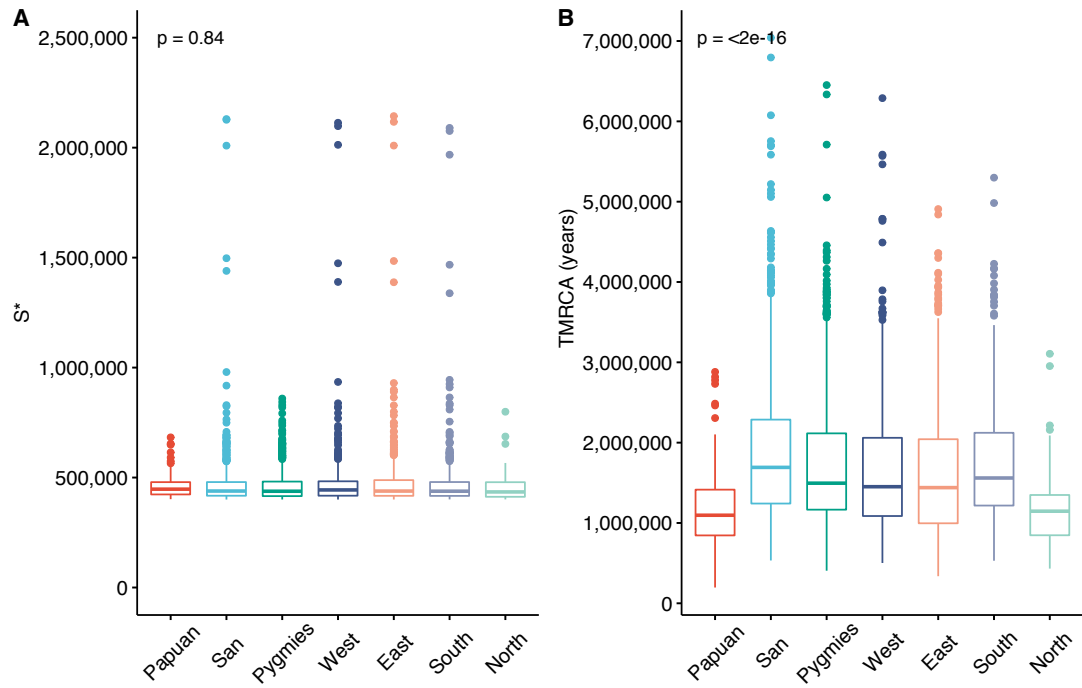


Fig. 5.6.: S^* and TMRCA across all Papuan and African SGDP samples. (A) Multiple comparison of top S^* candidates for Papuans ($n=15$) and African populations ($n=44$) from SGDP Data. Top S^* candidates are haplotypes that surpass the 400,000 score threshold.

in score but higher in the number of detected archaic haplotypes when compared to Papuans. However, a few African populations have distributions significantly lower than the Papuans such as the North consisting of Mozabite and Saharawi populations ($p < 0.001$, Wilcoxon test).

In order to take in consideration the possibility that variation in sample size might affect the distribution of detected regions, I also ran analyses on a subset of six per grouping and have focused the rest of my analyses on this subset. Top detected candidates are shown below with calculated S^* score (Figure 5.7A) and TMRCA (Figure 5.7B).

Although the results for the whole dataset and the resampled one are similar for top score, I had initially observed a significant difference among the full distributions for S^* scores observed for the whole dataset ($p < 0.001$, Kruskal-Wallis). Based on those initial results I tried to correct for biases that could be introduced due to populations with unequal sample sizes. Nevertheless I do recognise that for top S^* scores, as shown in Figure 5.1 unequal sample sizes can yield comparable results.

The subsampled grouping suggests that the distribution of S^* scores among the tails are comparable to each other as well (Figure 5.7A). However, TMRCA of the S^* score for each respective grouping suggests significantly different ages especially when one sees

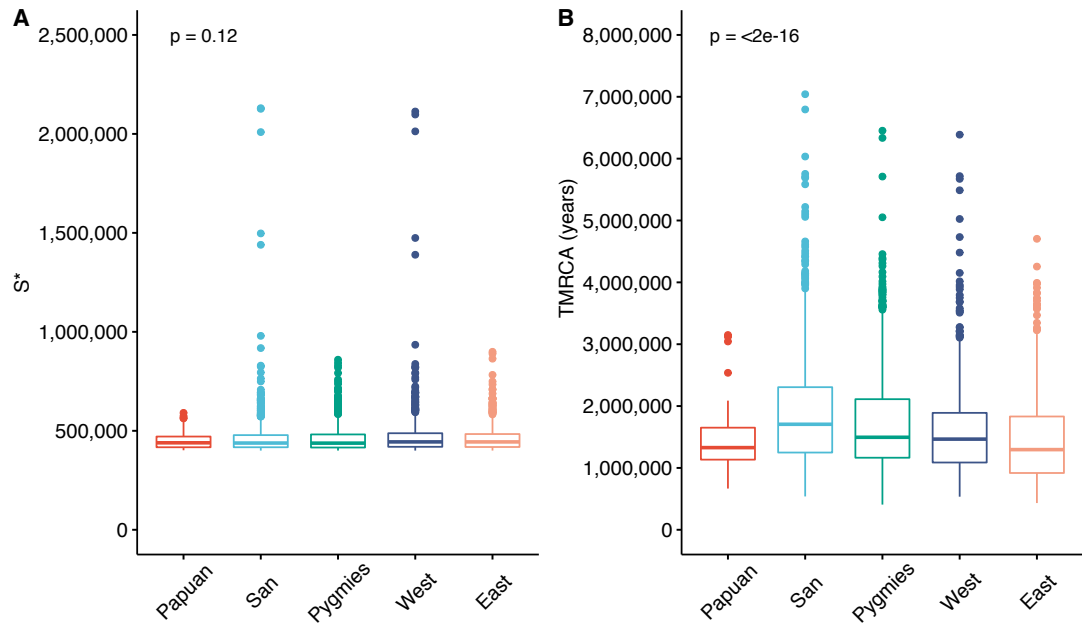


Fig. 5.7.: S^* and TMRCA across subsamples Papuan and African SGDP samples. (A) S^* score and (B) TMRCA multiple comparison (Kruskal-Wallis test) of top S^* candidates for Papuans and African populations. Groupings were subset to N=6 (Table S5.1; Supplementary Table). Top S^* candidates are windows that surpass the 400,000 score threshold.

the difference in age between detected regions in Papuans and in African populations (Figure 5.7B). If one considers the full distribution of S^* detected candidates regions (Figure 5.8A), African S^* scores are shifted towards a higher median S^* and are characterised by a long tail of S^* scores (Figure 5.7A). Consistently, it can be observed that TMRCA for higher scoring S^* regions tend to be older across all populations despite differences in demographic histories (Figure 5.8B). It is relevant to note that even the full distribution of S^* scores is different between Africans and Papuans. No major differences were observed among the top S^* scores for the African groups analysed here ($p=0.069$, Kruskal-Wallis test).

At the time these analyses were performed, Sprime had been newly released and shown to be more accurate than its predecessor, S^* (Browning *et al.*, 2018). In addition to S^* , I therefore also ran Sprime on the African samples from the SGDP. Using the same cutoff score of 400,000, both top Sprime and S^* detected introgressed candidate regions show that Papuan candidates do not necessarily have a higher TMRCA when compared to other populations (Figure 5.7B; Figure 5.9B) although Papuans have been shown to have archaic introgression (Vernot & Akey, 2015; Browning *et al.*, 2018; Reich *et al.*, 2010). All African groups tend to display older detected haplotypes than Papuans however the median TMRCA for Papuans is older (Figure 5.7B; Figure 5.9B).

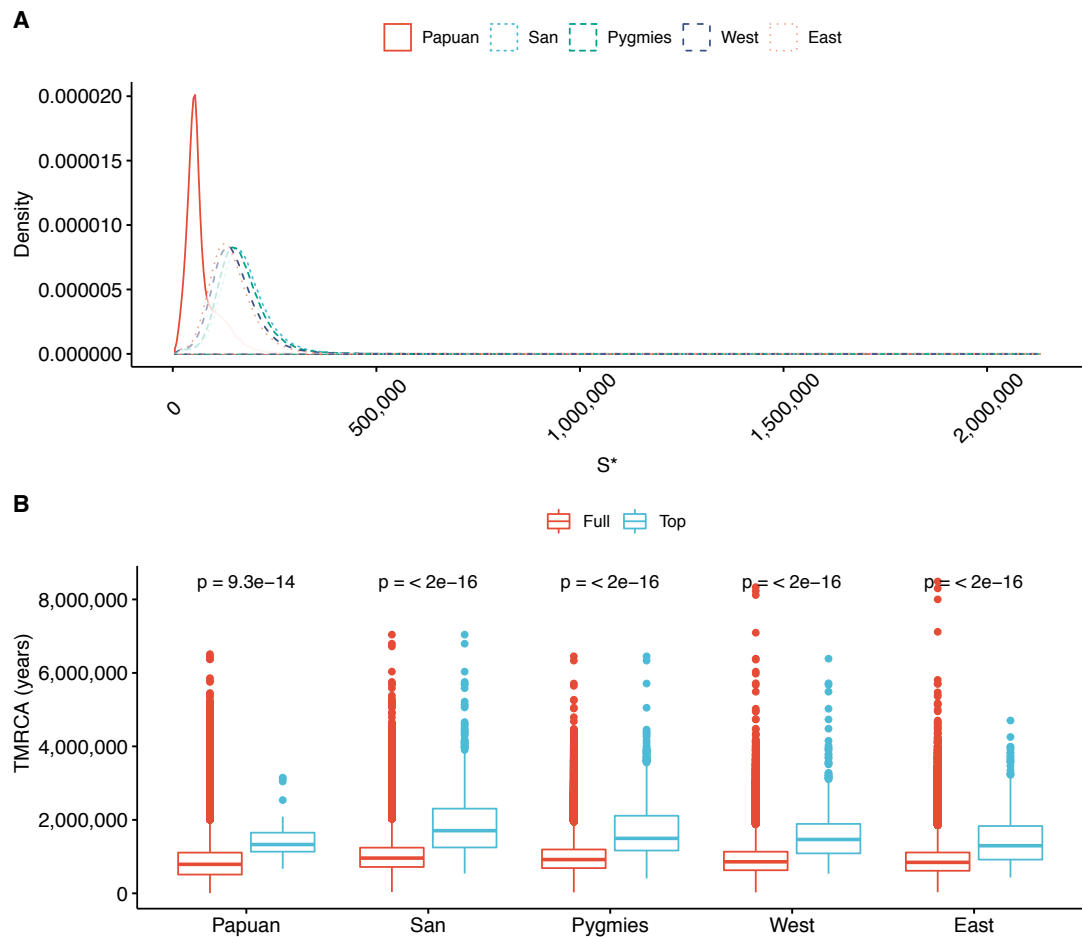


Fig. 5.8.: S^* and TMRCA distributions. (A) Density plot for the full distribution of S^* scores for each grouping and (B) TMRCA pairwise comparison for all S^* candidates versus top S^* candidates. Top S^* candidates are haplotypes that surpass the 400,000 score threshold.

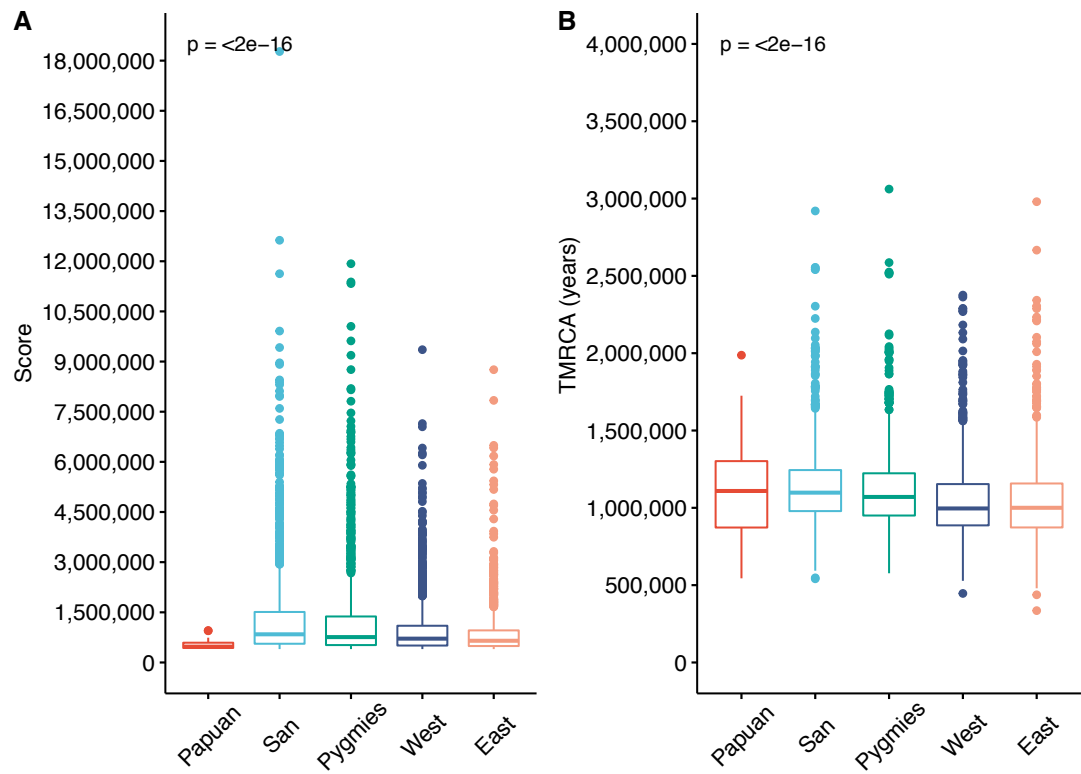


Fig. 5.9.: *Sprime* and TMRCA. (A) *Sprime* score and (B) TMRCA multiple comparison (Kruskal-Wallis test) of top *Sprime* candidates for Papuans and African populations. Groupings were subset to N=6 (Table X; Supplementary Table). Top *Sprime* candidates are haplotypes that surpass the 400,000 score threshold.

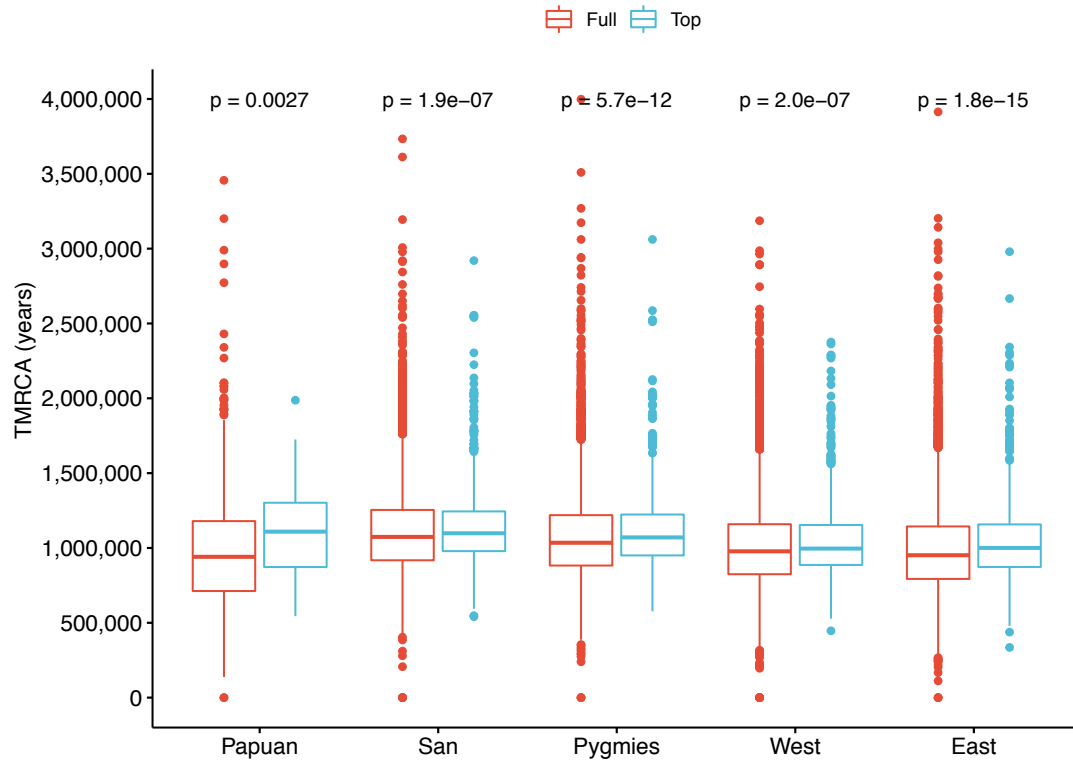


Fig. 5.10.: TMRCA: Full versus Top. TMRCA pairwise comparison for all Sprime detected haplotypes and top haplotypes detected by Sprime per grouping (N=6).

It appears that top Sprime candidates do not necessarily yield the oldest haplotypes (Figure 5.10) which falls in line with our findings from Chapter 4 in regards to the relationship of TMRCA and S^* score (Figure 4.11). This observation also fits in with the finding that true Papuan introgressed regions do not display an appreciable correlation between Sprime and TMRCA (Figure 5.5). However, top Sprime candidates do tend to have an older median TMRCA in comparison to the full distribution of detected candidates in all populations (Figure 5.10).

In comparing the findings of both S^* and Sprime it is clear the respective results for each grouping changes considerably especially when comparing the top detected regions (Figure 5.7A; Figure 5.9A). I have already established in Figure 5.4A that Sprime candidates are a subset of S^* candidates but which have been selected with higher accuracy (Figure 5.4B). With Sprime, Papuans score the lowest with both a lower scoring median as well as a shorter tail of detected haplotypes (Figure 5.9A). It is crucial to also point out that given the longer haplotypes detected by Sprime, TMRCA calculations can vary. In the case of S^* where detection is performed across windows of equal length, there is less variation in the count of derived alleles. I stress this as one of the limitations of the Hudson method for estimating TMRCA and suggest that in future studies other approaches to calculate TMRCA across whole genome sequences

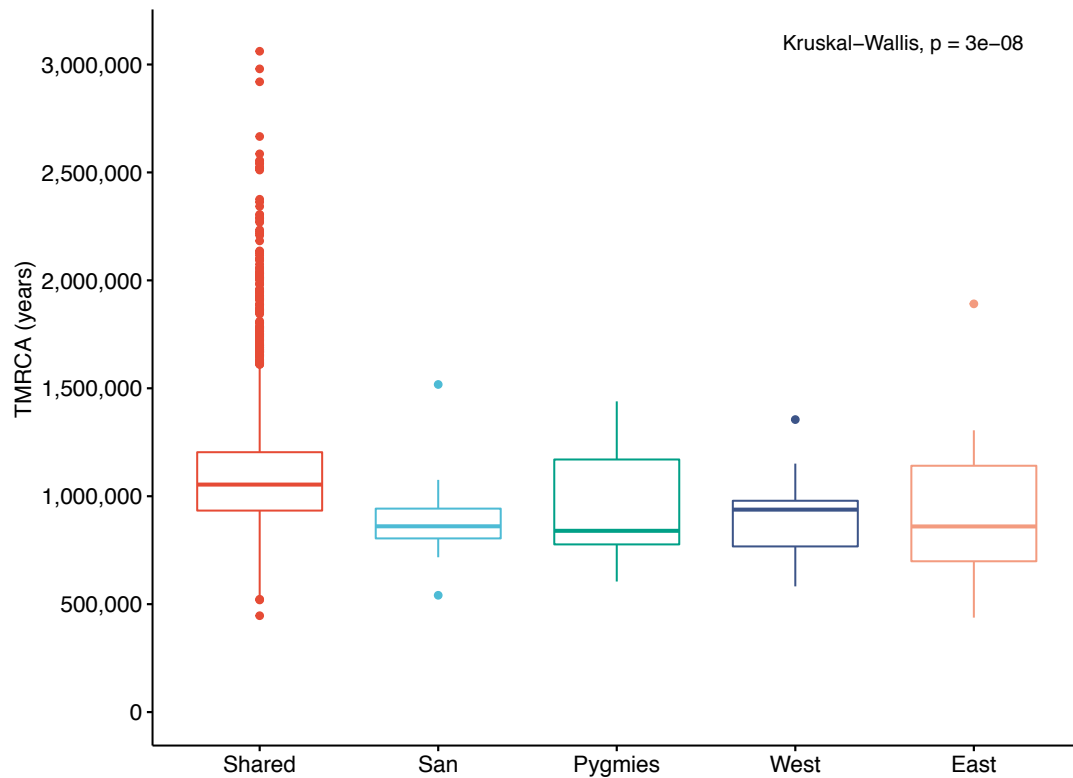


Fig. 5.11.: TMRCA: Top shared regions. TMRCA multiple comparison of top Sprime candidates for shared regions (red) and unique regions per African grouping.

should be implemented even if the Hudson (Hudson, 2007) method attempts to control it by including length of the regions considered.

Nevertheless, when comparing between populations under either method, TMRCA for detected haplotypes in Africans are consistently older than those in Papuans (Figure 5.7B; Figure 5.9B). Using Sprime, both Papuan and African candidate ages for the full distribution and for the top display significantly different age distributions (Figure 5.10).

To better characterise the behaviour of the potentially introgressed regions, I investigated the differences in TMRCA between top candidate regions detected by Sprime (Figure 5.11). Shared haplotypes and unique East haplotypes were dated with the oldest TMRCA. The San displayed significantly younger TMRCA medians compared to shared haplotypes (Figure 5.12A-C). With the exception of the West and the East (Figure 5.12F), which were the two populations that displayed an older median and the oldest haplotypes, respectively, TMRCA estimates for shared haplotypes among the top candidate regions tended to be older than unique haplotypes (Figure 5.12A-E) although not statistically different. Unique haplotypes per African grouping fit in with what is presently understood about African demographic history; for example in Fig-

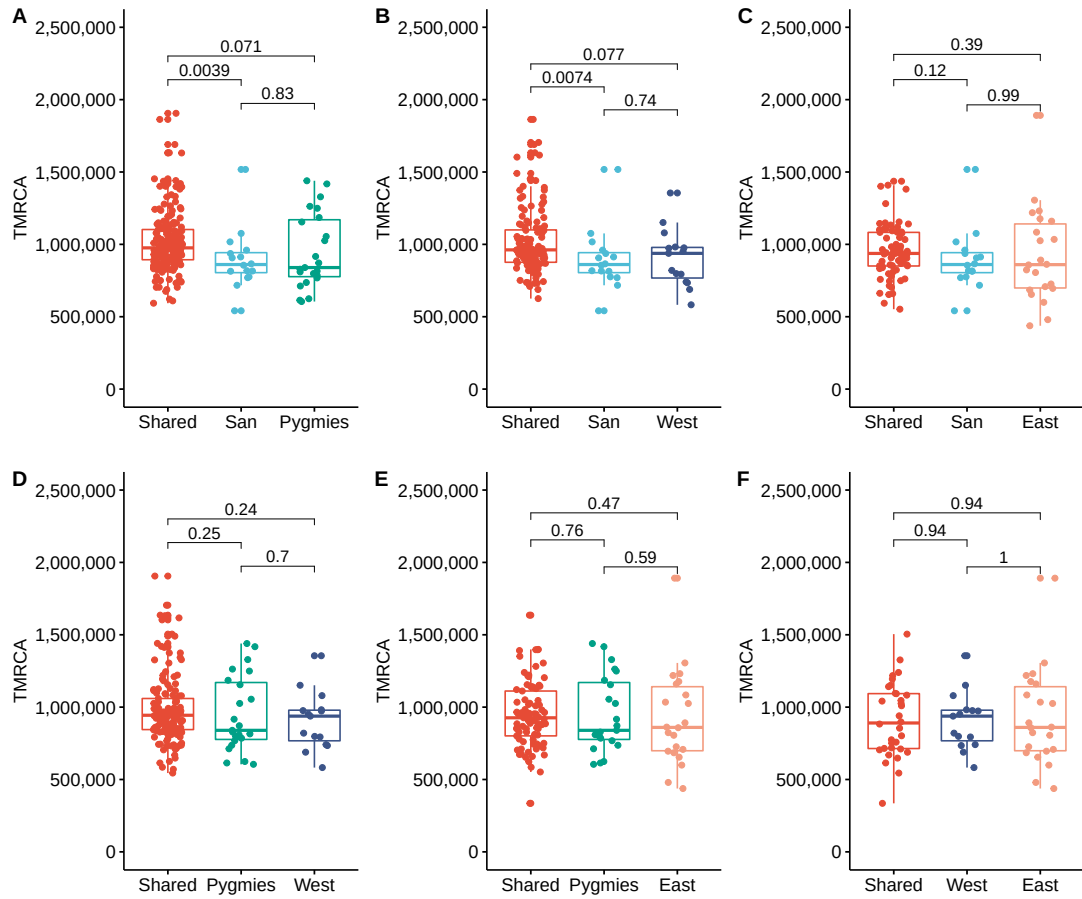


Fig. 5.12.: *TMRCA: Shared versus unique regions.* Paired TMRCA (years) estimates, (Wilcoxon tests) for shared (red) and unique top Sprime candidate regions in San, Pygmies, West and East groupings.

ure 5.12B,D,F we observe that West shares older haplotypes with Pygmies consistent with gene flow from the Bantu expansion into central Africa (Tishkoff *et al.*, 2009; Berniell-Lee *et al.*, 2009; Li *et al.*, 2014), although shared haplotypes do not seem to be significant older or younger than unique ones ($p > 0.05$) except in the case of the San.

To assess whether Sprime is capable of retrieving sites that had been detected in previous studies, I compared the regions detected here with the candidate loci from available studies that looked at archaic introgression in African populations (Plagnol & Wall, 2006; Wall *et al.*, 2009; Hammer *et al.*, 2011; Lachance *et al.*, 2012; Hsieh *et al.*, 2016a; Durvasula & Sankararaman, 2018). In particular, I found that with Sprime I was able to retrieve the majority of candidate loci from previous studies (Table 2). Often these candidate loci were shared haplotypes despite being originally found in a single population. Among Hsieh *et al.* (2016a) candidate loci, only 8.5% of them were not completely shared by all African groupings in my analysis (Table 2, middle row). Sprime was unable to recover only 7 of the 256 candidate regions which were originally found in central African RHG (Hsieh *et al.*, 2016a) (Table 2, bottom row).

Tab. 5.2.: Putatively introgressed regions detected in other studies in both genic and non-genic regions. The "non-fully shared" section corresponds to Hsieh *et al.* (2016a) candidate regions which in this study were not shared by all African populations. Bottom section lists loci unrecovered by Sprime. All coordinates correspond to hg19 reference positions for each chromosome. The "Gene | Locus" column denotes the genic region it is in or the known chromosomal arm notation from a previous study. Columns for each geographical or population grouping indicate whether these previously detected candidate regions in other studies are present (+) in this study or absent (-).

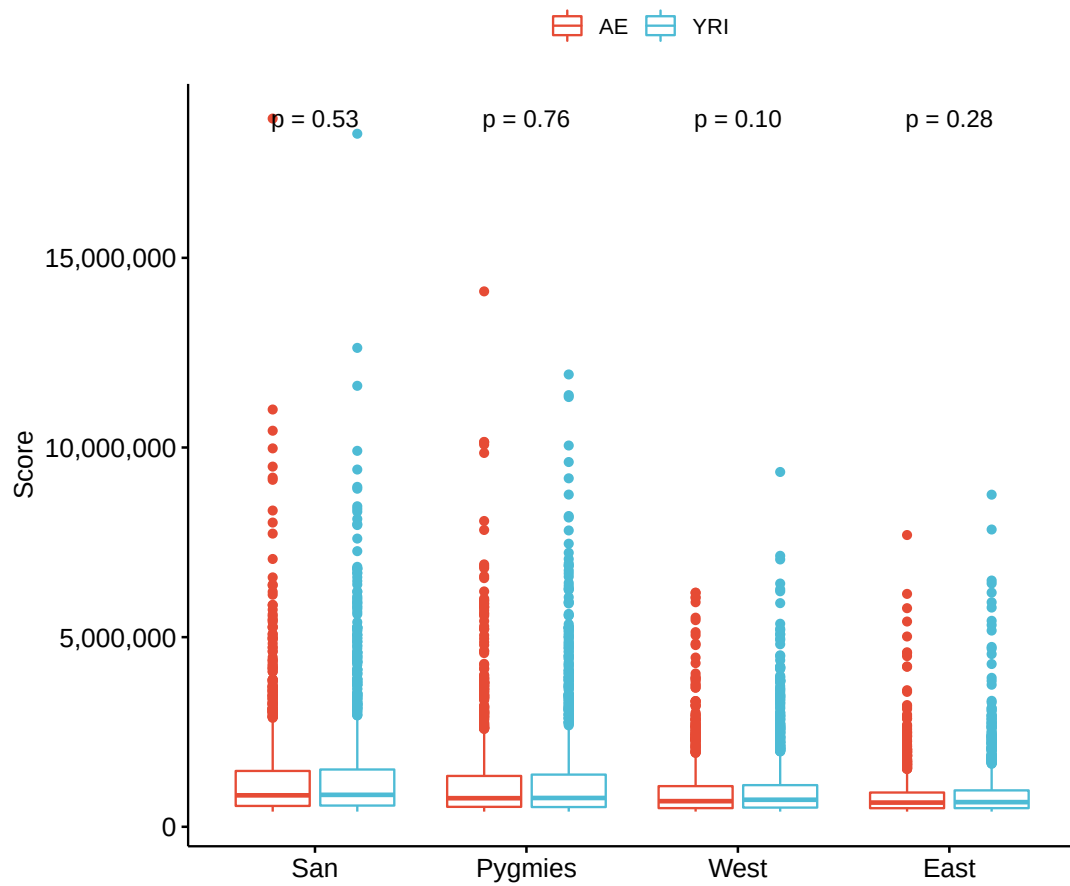
	Chromosome	Start	End	Gene Locus	San	Pygmies	West	East	Study
Genic & Non-genic	4	20731704	21950254	KCNIP4	+	+	+	+	(Durvasula & Sankararaman, 2018)
	4	179361853	179361854	4qMB197	-	+	-	-	(Hammer <i>et al.</i> , 2011)
	5	82400738	82649061	XRCC4	+	+	+	+	(Plagnol & Wall, 2006; Wall <i>et al.</i> , 2009)
	6	136552412	136570237	MTFR2	+	+	+	+	(Durvasula & Sankararaman, 2018)
	8	116426250	116680110	TRPS1	+	+	+	+	(Durvasula & Sankararaman, 2018)
	10	26333496	26361600	MYO3A	+	+	+	-	(Hsieh <i>et al.</i> , 2016a)
	10	68953370	69010247	CTNNA3	+	+	+	-	(Hsieh <i>et al.</i> , 2016a)
	13	93169285	93292517	GPC5	+	+	-	+	(Hsieh <i>et al.</i> , 2016a)
	13	108697052	108697053	13qMB107	+	-	-	-	(Hammer <i>et al.</i> , 2011)
	15	29993798	30260964	TJP1	+	-	-	-	(Plagnol & Wall, 2006; Wall <i>et al.</i> , 2009)
	15	48623363	48634275	DUT	+	+	+	+	(Plagnol & Wall, 2006; Wall <i>et al.</i> , 2009)
	16	82069029	82132041	HSD17B2	+	+	+	+	(Durvasula & Sankararaman, 2018)
	17	29422327	29701173	NF1	+	-	+	-	(Durvasula & Sankararaman, 2018)
	17	43497210	43750940	ARHGAP27	+	+	-	+	(Hsieh <i>et al.</i> , 2016a)
	18	62567942	62567943	18qMB60	-	+	+	-	(Hammer <i>et al.</i> , 2011)
	21	17962645	17962646	MIR125B2	+	+	+	+	(Durvasula & Sankararaman, 2018)
Non-fully shared Hsieh <i>et al.</i> (2016a) haplotypes	1	73617423	73668215		+	+	-	-	(Hsieh <i>et al.</i> , 2016a)
	2	82250664	82309165		+	+	-	+	(Hsieh <i>et al.</i> , 2016a)
	2	107766741	107809724		+	+	+	-	(Hsieh <i>et al.</i> , 2016a)
	2	202489878	202541708		+	+	-	+	(Hsieh <i>et al.</i> , 2016a)
	3	144870134	144968209		+	+	+	+	(Hsieh <i>et al.</i> , 2016a)
	3	190667597	190741231		+	+	+	-	(Hsieh <i>et al.</i> , 2016a)
	4	39563067	39618627		-	+	+	+	(Hsieh <i>et al.</i> , 2016a)
	4	159578723	159680528		+	+	+	-	(Hsieh <i>et al.</i> , 2016a)
	5	29575558	29635655		+	-	+	+	(Hsieh <i>et al.</i> , 2016a)
	5	52644425	52709322		+	+	-	-	(Hsieh <i>et al.</i> , 2016a)
	6	86929266	87012195		+	+	+	-	(Hsieh <i>et al.</i> , 2016a)
	6	122065689	122109224		+	+	+	-	(Hsieh <i>et al.</i> , 2016a)
	7	75442269	75533898		-	+	+	-	(Hsieh <i>et al.</i> , 2016a)
	7	83671947	83746791		+	+	-	+	(Hsieh <i>et al.</i> , 2016a)
Unrecovered	8	95972718	96023983		+	+	+	-	(Hsieh <i>et al.</i> , 2016a)
	8	97747709	97811234		+	+	+	-	(Hsieh <i>et al.</i> , 2016a)
	12	47917203	47965081		-	+	+	-	(Hsieh <i>et al.</i> , 2016a)
	14	48166800	48228866		+	+	-	+	(Hsieh <i>et al.</i> , 2016a)
	1	50808705	50863203		-	-	-	-	(Hsieh <i>et al.</i> , 2016a)
	4	72843246	72873961		-	-	-	-	(Hsieh <i>et al.</i> , 2016a)
	5	8559982	8601614		-	-	-	-	(Hsieh <i>et al.</i> , 2016a)
	6	29951049	29971015		-	-	-	-	(Hsieh <i>et al.</i> , 2016a)
	17	39228142	39411030		-	-	-	-	(Hsieh <i>et al.</i> , 2016a)
	19	21041908	21154604		-	-	-	-	(Hsieh <i>et al.</i> , 2016a)
	19	40200085	40281324		-	-	-	-	(Hsieh <i>et al.</i> , 2016a)

5.6 A Tale of Two Maps: HapMap YRI & African Enriched Recombination Maps

To determine whether the top candidate haplotypes detected across African groupings and their corresponding ages would vary if one used different genetic maps when performing Sprime, I compared two African-ancestry specific recombination maps: one that focuses on West-African-ancestry (HapMapYRI) (The HapMap Consortium *et al.*, 2010) and one that is derived by detecting hotspots correlated specifically with African-ancestry (African Enriched) (Hinch *et al.*, 2011). Score-wise Sprime performs similarly when using either genetic map (Figure 5.13). In Chapter 4, I showed how a variable and constant recombination rate affect the behaviour of S^* distributions in simulated data. In a similar vein, here I assess how these two genetic maps that are African-ancestry specific can affect the behaviour of Sprime, a tool reliant on measuring linkage disequilibrium and therefore sensitive to recombination as shown in previous studies with S^* (Hsieh *et al.*, 2016a).

Interestingly, when I estimated the TMRCA for the top score regions, TMRCA calculated for the top candidates detected using the African Enriched genetic map showed younger haplotypes than with HapMap YRI (Figure 5.14) excluding Pygmies. TMRCA differences between the two genetic maps were only statistically significant for Pygmies ($p < 0.05$) and the East ($p < 0.01$). This is in line with the general observation of a lack of strong correlation between Sprime scores and TMRCA. The length of the haplotypes detected by Sprime did not vary significantly between the two genetic maps (Figure 5.15). This suggests that the antiquity of the haplotypes are not due to simply differences in size of the fragments. The regions detected using both maps were comparably the sample haplotypes (Supplementary Data Repository), but the number of polymorphisms counted is strongly variable depending on where a haplotype begins and on the local recombination rate.

Contrary to what could be expected about putatively introgressed haplotypes in African populations, we observe strikingly long haplotypes—some which are too long to be biologically plausible—rather than the expected haplotypes of 50-kb if archaic admixture took place 40,000-50,000 years ago in Africa (Sankararaman *et al.*, 2016; Sankararaman *et al.*, 2014) or shorter considering the faster recombination rate in African populations (Hinch *et al.*, 2011; Wegmann *et al.*, 2011). This obviously raises the issue of how reliable the identified introgressed haplotypes are and if some conspicuous limitation is present in the approach. One explanation for these unexpected lengths is that the outgroup used to filter out existing variation serves poorly in accounting for the shared human derived alleles. In other words, because non-African populations are but a small subset of African variation, we are unable to filter out all that is human-derived. While



[H]

Fig. 5.13.: *Sprime: HapMap versus AE maps.* Pairwise comparison of scores for top *Sprime* candidates detected using the African Enriched (AE; red) and HapMap YRI (YRI; blue) genetic maps.

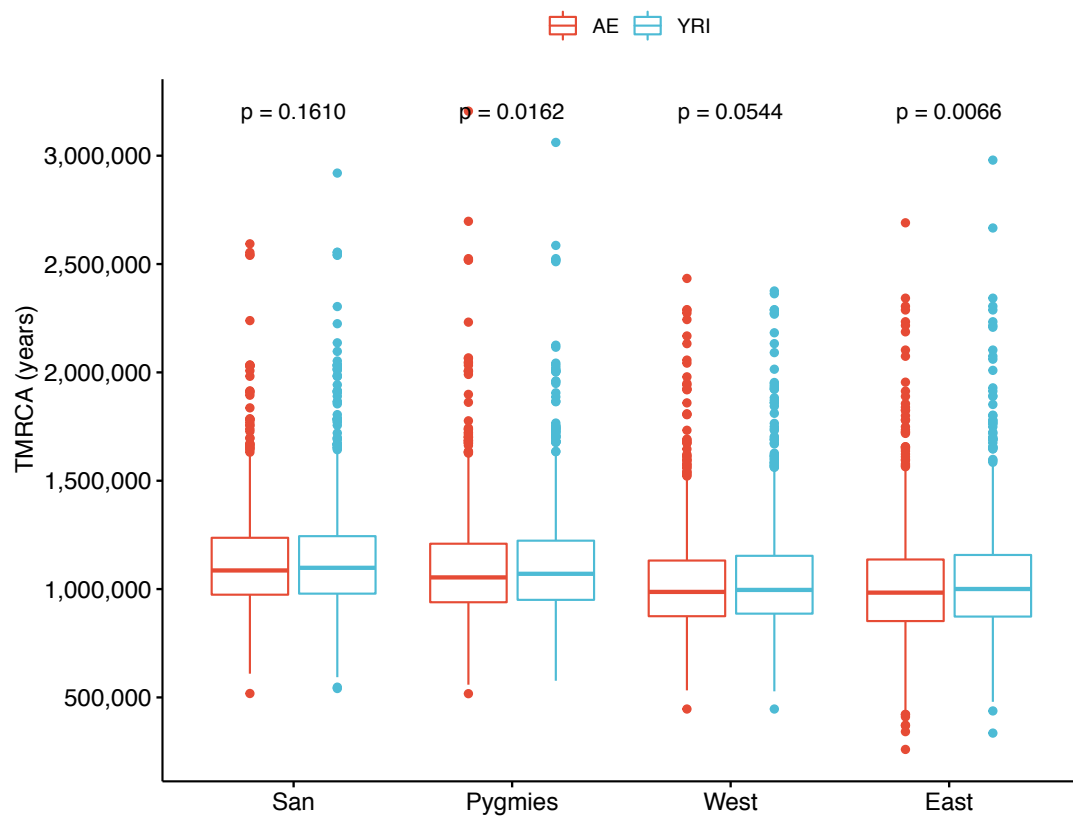


Fig. 5.14.: *TMRCA: HapMap versus AE maps.* Pairwise comparison (Wilcoxon test) of TMRCA for top Sprime candidates detected using the African Enriched (AE; red) and HapMap YRI (YRI; blue) genetic maps.

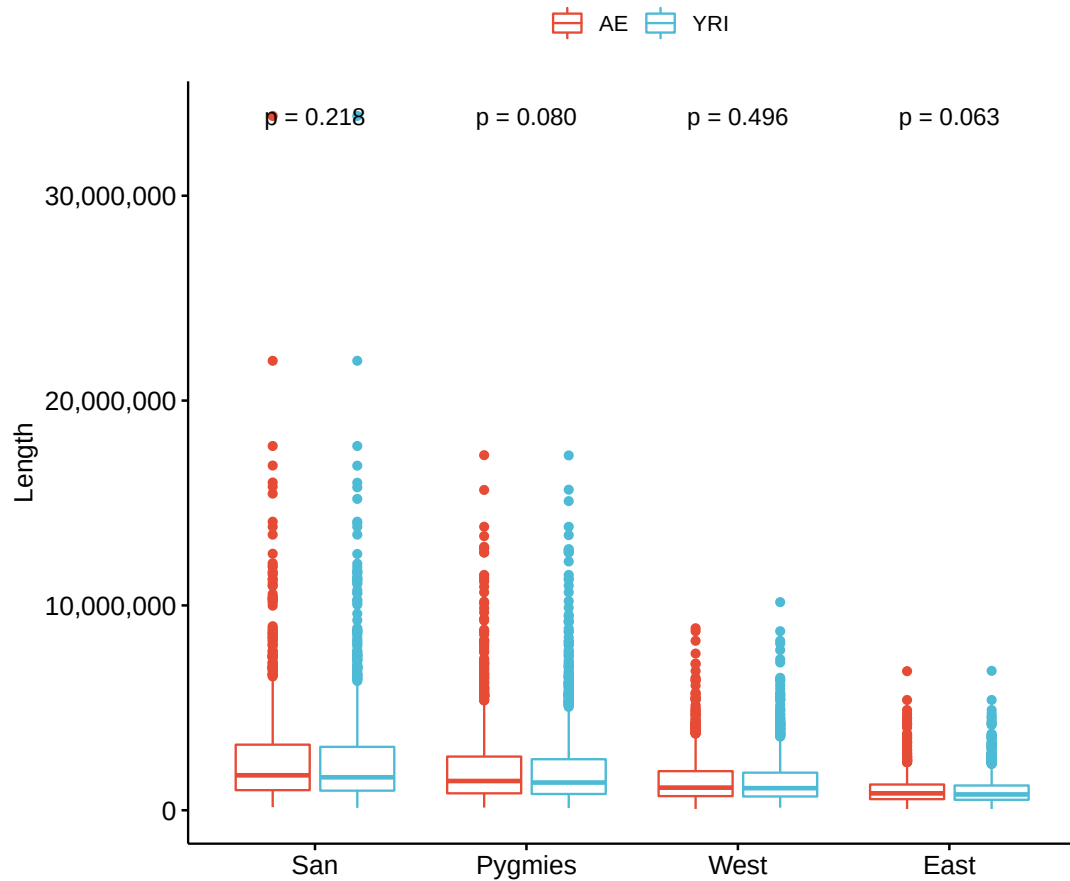


Fig. 5.15.: Length: HapMap versus AE maps. Pairwise comparison of lengths (bp) for top Sprime candidates detected using the African Enriched (AE; red) and HapMap YRI (YRI; blue) genetic maps.

it is possible to look solely at private SNPs for each population as other studies have done (Skov *et al.*, 2018), the high amount of variation within populations in addition to recent admixture across Africa does not lend itself for finding archaic introgression easily. Despite Yorubans being the proxy outgroup population to seek for Neanderthal and Denisovan introgression in non-Africans, no corresponding outgroup population exists to search for archaic introgression in Africans. In a personal communication, Skov (Skov, 2018) showed that even by using additional African individuals from the 1000 Genomes Project to alternate as outgroups for each other, this approach did not manage to efficiently filter out shared SNPs, especially for populations such as the San and Pygmies.

5.7 *D*-statistics: Archaic ancestry in Africa

In order to investigate the relationship of African populations with known archaic samples (Neanderthal and Denisovans) and potentially shed light on the result observed for the S^* and Sprime introgression signals, I decided to perform a series of analyses based on *D*-statistics for several combinations of both modern and archaic individuals. *D*-statistics provide a simple yet powerful test for a deviation of bifurcating evolutionary history and is therefore suitable to test for introgression. Performing this test was motivated by the different signals which on one hand suggest that putatively introgressed regions are unexpectedly long with old TMRCA (Figure 5.9B; Figure 5.15) but at the same time a great deal of these old haplotypes are shared across several African populations (Figure 5.11). Under a hypothetical scenario of archaic admixture in Africa, one might expect archaic populations in Africa to be more closely related to non-African archaics such as Neanderthal and Denisovans. In this case, I can use *D*-statistics to disentangle whether admixture has occurred recently ($\sim 50,000$ years ago) in one or more populations or if the haplotypes I observe with S^* and Sprime are artefacts from not having a proper outgroup in combination with signatures of incomplete lineage sorting (ILS).

In agreement with previous studies, I found that non-Africans in comparison to Africans share more alleles with Neanderthals (Figure 5.16). Interestingly not all African populations are similarly related to Neanderthal, with some showing closer affinity to Neanderthal than others. I ran *D*-statistics comparing African populations to Vindija Neanderthal and found that indeed a few individuals from North Africa, Mozabites and Saharawi are closer to Neanderthals than other Africans. One possible explanation for such results is the occurrence of back-migration into African from non-African populations, which would have operated as a vehicle for introducing Neanderthal signatures into African populations (Schlebusch *et al.*, 2012; Pickrell *et al.*, 2014; Pagani *et al.*, 2012; Skoglund *et al.*, 2017; Breton *et al.*, 2014). Notably, African populations that are closer to European populations (Appendix Figure A.3) were also the ones that tended to share more alleles with Neanderthals (Figure 5.16).

I repeated the same set of analyses using the a different genome of Eurasian Archaics, a representative sample of the so-called Denisovans (Reich *et al.*, 2010). As Denisovans share a common recent ancestor with Neanderthal earlier than with humans, the relationships of African populations versus Denisovans is expected to broadly mirror the one observed for the Neanderthals.

However, my results pointed out that some African groups, such as the San and some Mbuti individuals, share more alleles with Denisovan not only when compared to other African groups but even when compared to some non-Africans (i.e. English)

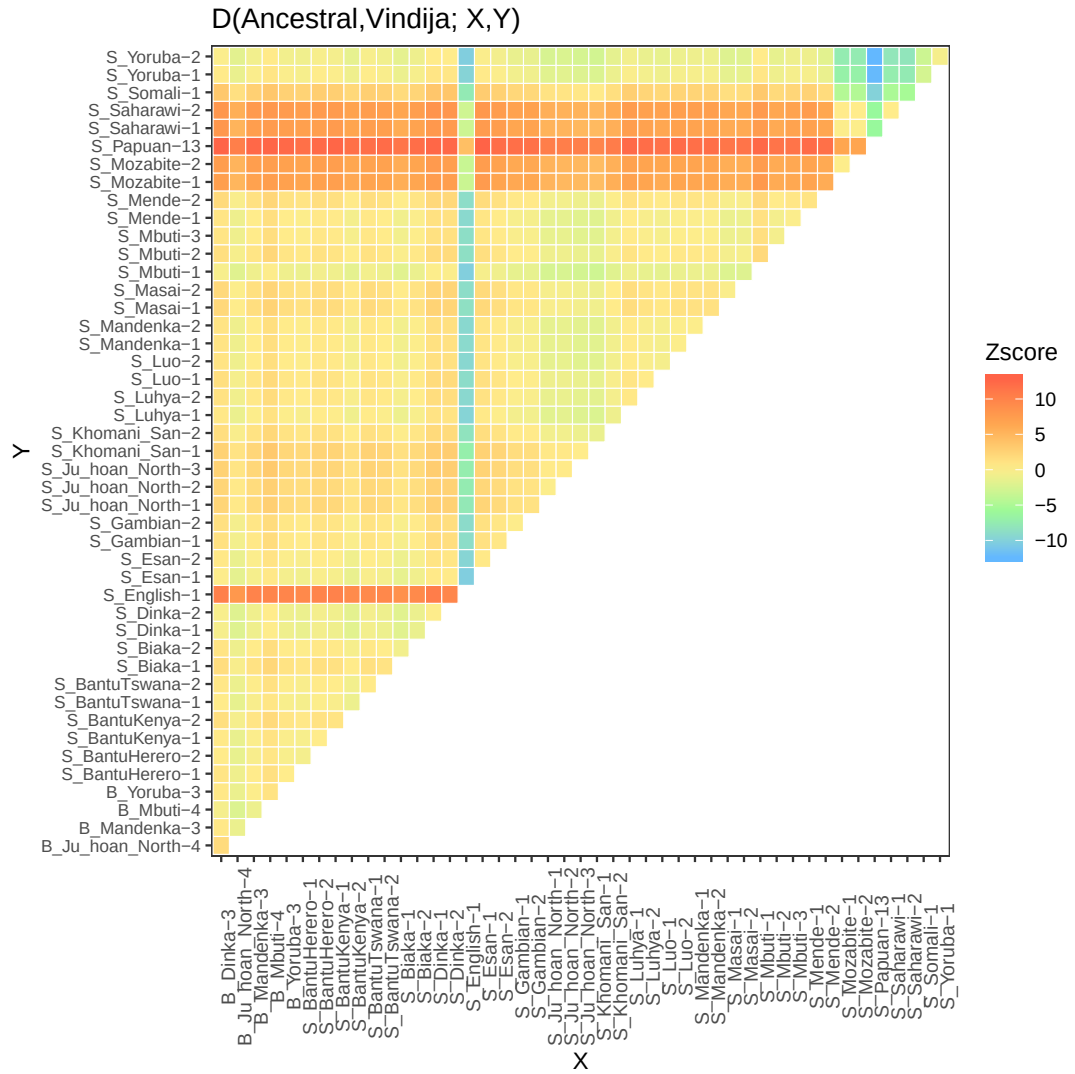


Fig. 5.16.: D -statistics comparing two individuals to Vindija Neanderthal with the reconstructed common ancestor (Ancestral) as the outgroup. The heatmap indicates whether individual on the Y-axis shares an excess of Neanderthal alleles (red) or an absence of them (blue).

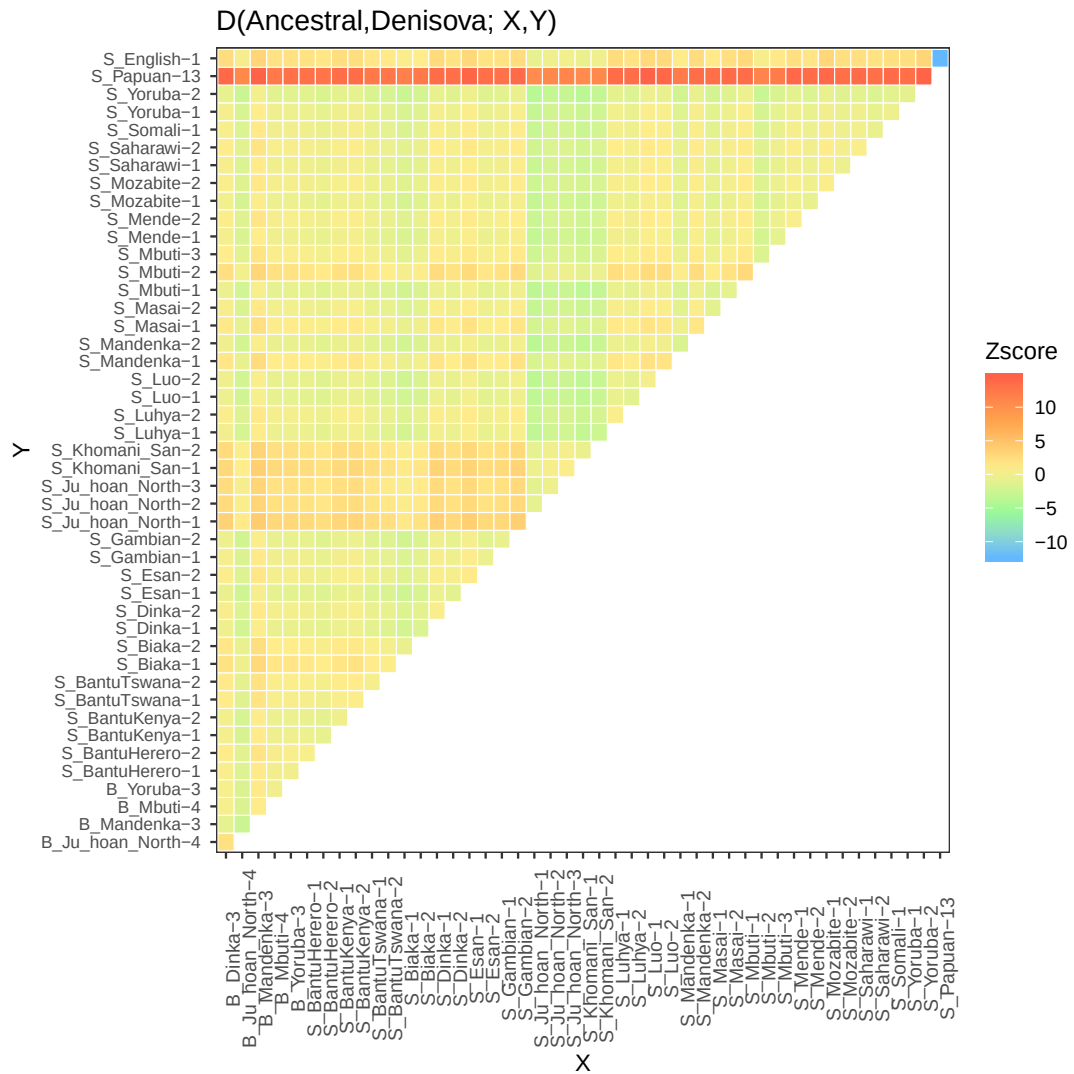


Fig. 5.17.: *D*-statistics comparing two individuals to Denisovans with Ancestral as the outgroup. The heatmap indicates whether individual on the Y-axis shares an excess of Denisovan alleles (red) or an absence of them (blue).

(Figure 5.17). To further explore these unexpected results I performed *D*-statistics comparing African and non-African individuals to both Denisovan and Neanderthal genomes (Figure 5.18).

English, Mozabite, and Saharawi are the three populations that are closest to Neanderthal, while Papuan resulted closer to Denisovans (Figure 5.18). These results are in line with an history of direct introgression of Neanderthal DNA into non-Africans, followed by secondary introgression in African populations due to human-human admixture. This is further supported when we correlate the *D*-statistics comparing African to non-African populations and those comparing African populations to Vindija Neanderthal (Figure S7.4) The closest affinity of Papuans to Denisovans clearly reflects the higher contribution of Denisovans than Neanderthal in this population. Although still

far below the mean (Z score < 0), the San, the Mbuti and a Mandenka individual share more derived alleles with Denisovans than other African populations despite the fact that a) Neanderthal and Denisovans should be equally related to African populations and b) no variation in European introgression is detected in these groups (Appendix Figure A.4). One possible explanation for these results is the fact that, given the larger effective population size experienced by some African populations in the past (Li & Durbin, 2010; Schiffels & Durbin, 2014; Gravel *et al.*, 2011; Relethford & Jorde, 1999), the excess allele sharing that those African populations display with Denisovans is potentially due to ILS—variation present in the ancestral population to humans, Neanderthal and Denisovans randomly sorted among the three groups. I therefore ran D -statistics on two subsets of alleles: 1) SNPs which are private to both Denisovans and Neanderthal (Figure 5.19) and 2) SNPs which are private to either Denisova or Neanderthal, respectively (Figure 5.20; Figure 5.21).

We observe that indeed Papuans, followed by English, share excess derived alleles with Denisovans when considering alleles which are also shared with Neanderthal (Figure 5.19). Given the higher proportion of archaic DNA hosted by Papuans, their affinity to Denisovans is expected to be higher than those of English genomes.

Notably this pattern is not observed for Denisovan private SNPs (Figure 5.20). In this case San, followed by Mbuti, share an excess of Denisovan private SNPs when compared to all other samples, including Papuans. Some RHG populations such as the Biaka and Bantu-speaking southern African populations such as the Bantu Tswana share about the same amount of private Denisovan SNPs as Papuans (Figure 5.20).

When looking at Vindija Neanderthal private SNPs (Figure 5.21), populations of non-African ancestry shared the most alleles with Neanderthals, followed by African populations with known Eurasian admixture (e.g. Mozabite and Saharawi). San and Mbuti are the other African populations that also display an excess of allele sharing with Vindija private SNPs, but less so than in the case of Denisovan private SNPs.

Papuans clearly share the most with both Denisovans and Neanderthal (Figures 5.15-5.21) through admixture with both archaic populations. The overall results indicate that African populations show different degrees of allele sharing with archaic Eurasian samples, mostly reflecting Eurasian contribution to African groups. In addition, Mbuti and San showed an expected excess of shared alleles with Denisovan, but not Neanderthal genomes. An excess of shared fragments between African and Denisovans has also been reported when looking at identity-by-descent (IBD) segments between Africans, Neanderthals, and Denisovans (Povysil & Hochreiter, 2016).

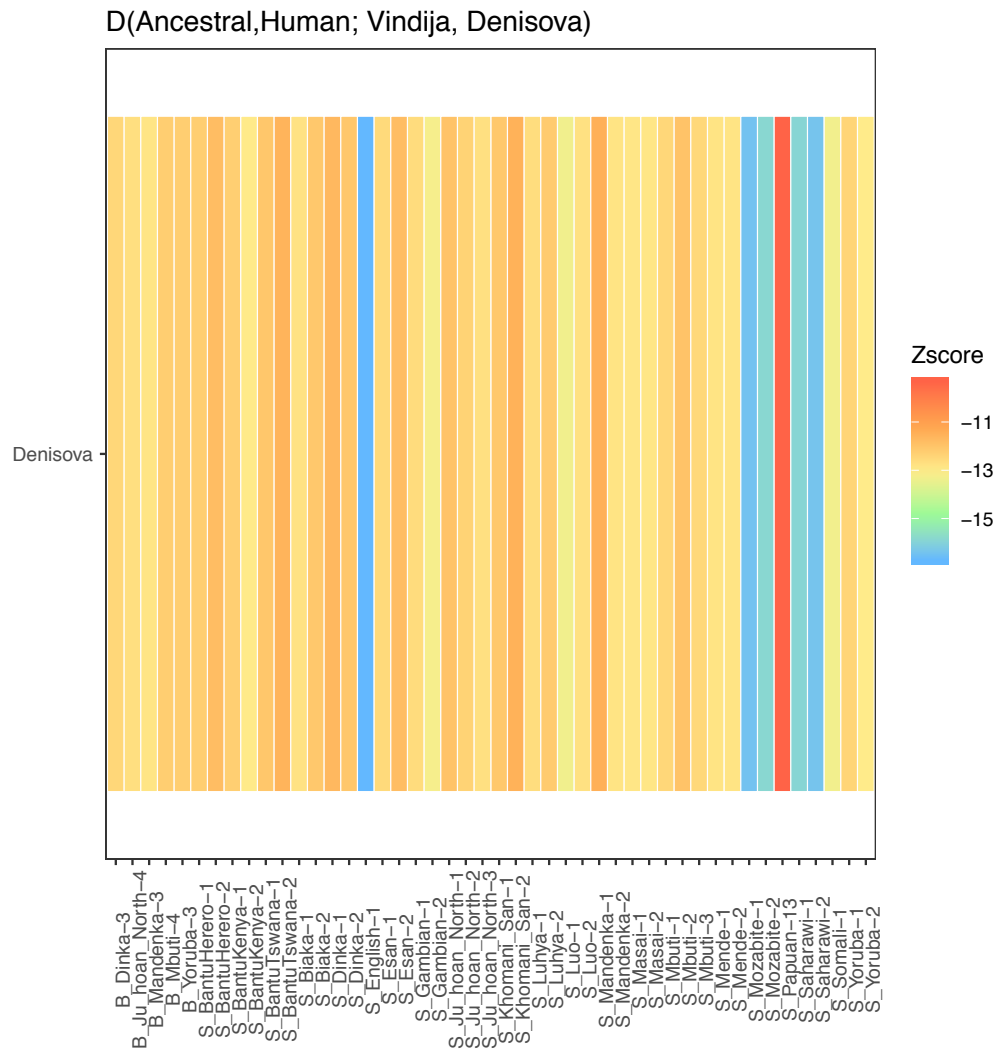


Fig. 5.18.: *D*-statistics comparing Neanderthal and Denisova to a modern human population with Ancestral as the outgroup. The heatmap indicates which human populations share an excess of alleles with Denisova (red) or Neanderthal (blue).

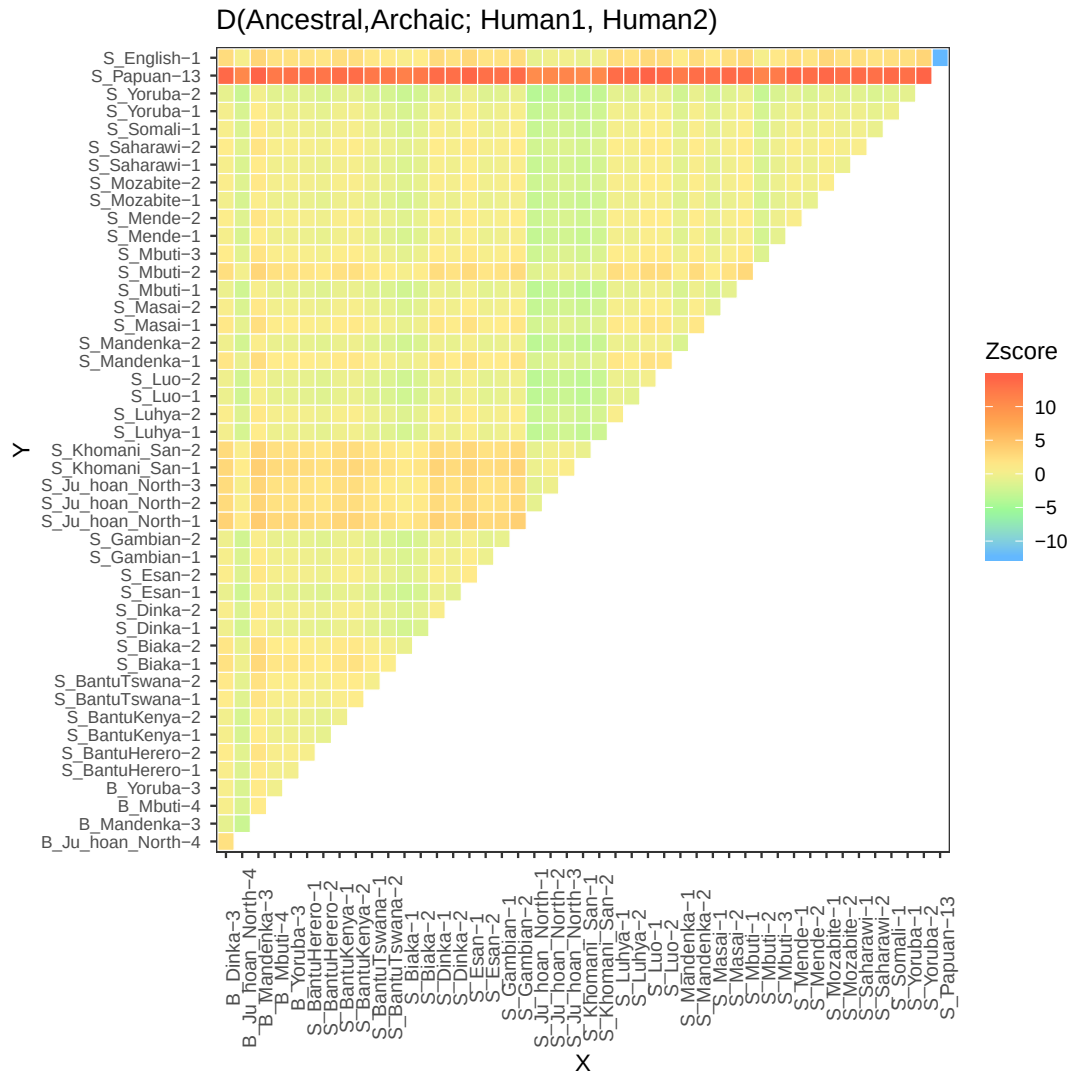


Fig. 5.19.: *D*-statistics for a pair of individuals to observe excess allele sharing to an archaic while the Ancestral remains as the outgroup. Only derived mutations shared exclusively by both Denisova and Neanderthals were used here. If the value is positive (red) the individual on the Y-axis is closer to an archaic than the individual on the X-axis.

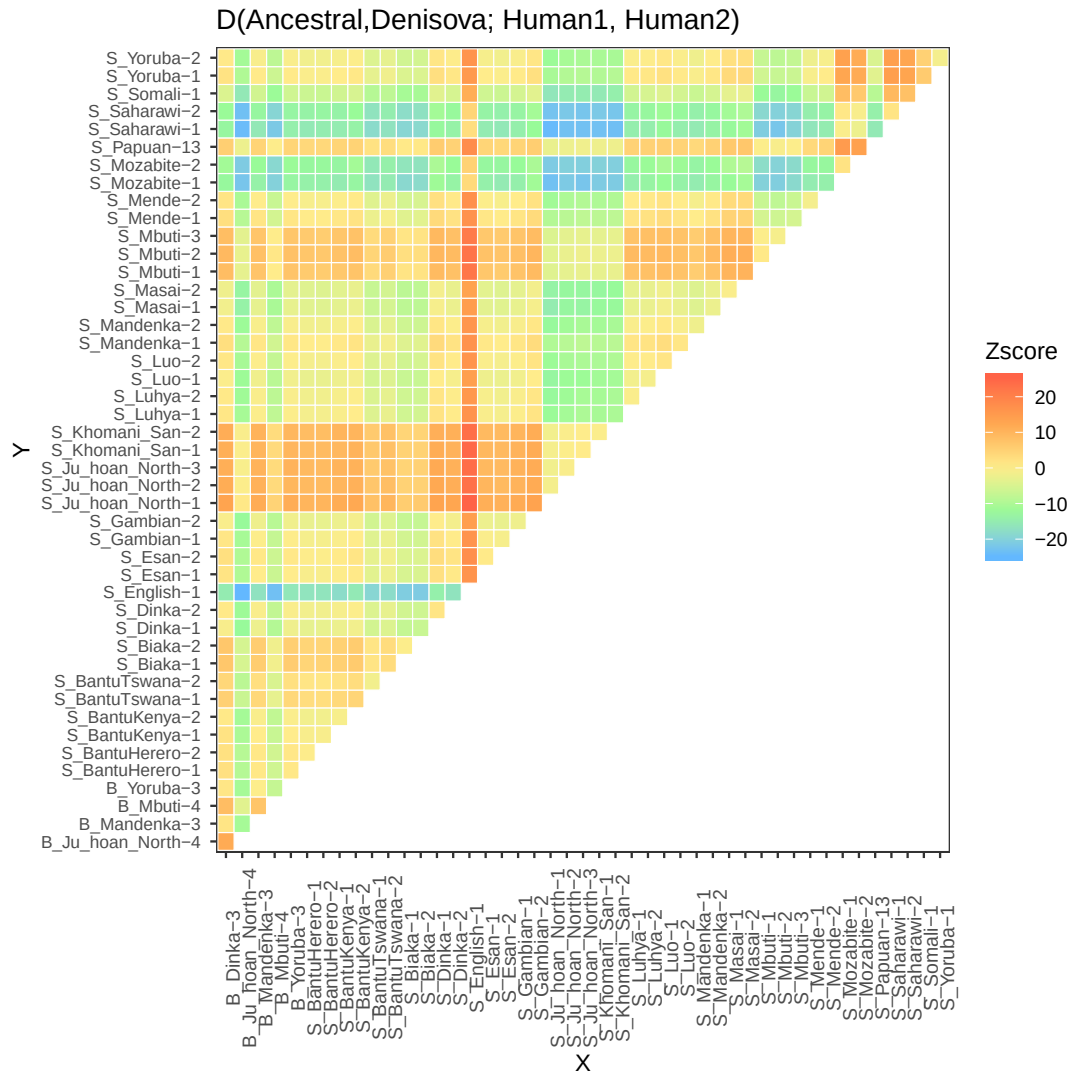


Fig. 5.20.: *D*-statistics for a pair of individuals to observe excess allele sharing with Denisovans while the Ancestral remains as the outgroup. Only positions where Denisovans had privately derived mutations were used. If the value is positive (red) the individual on the Y-axis is closer to Denisovans than the individual on the X-axis.

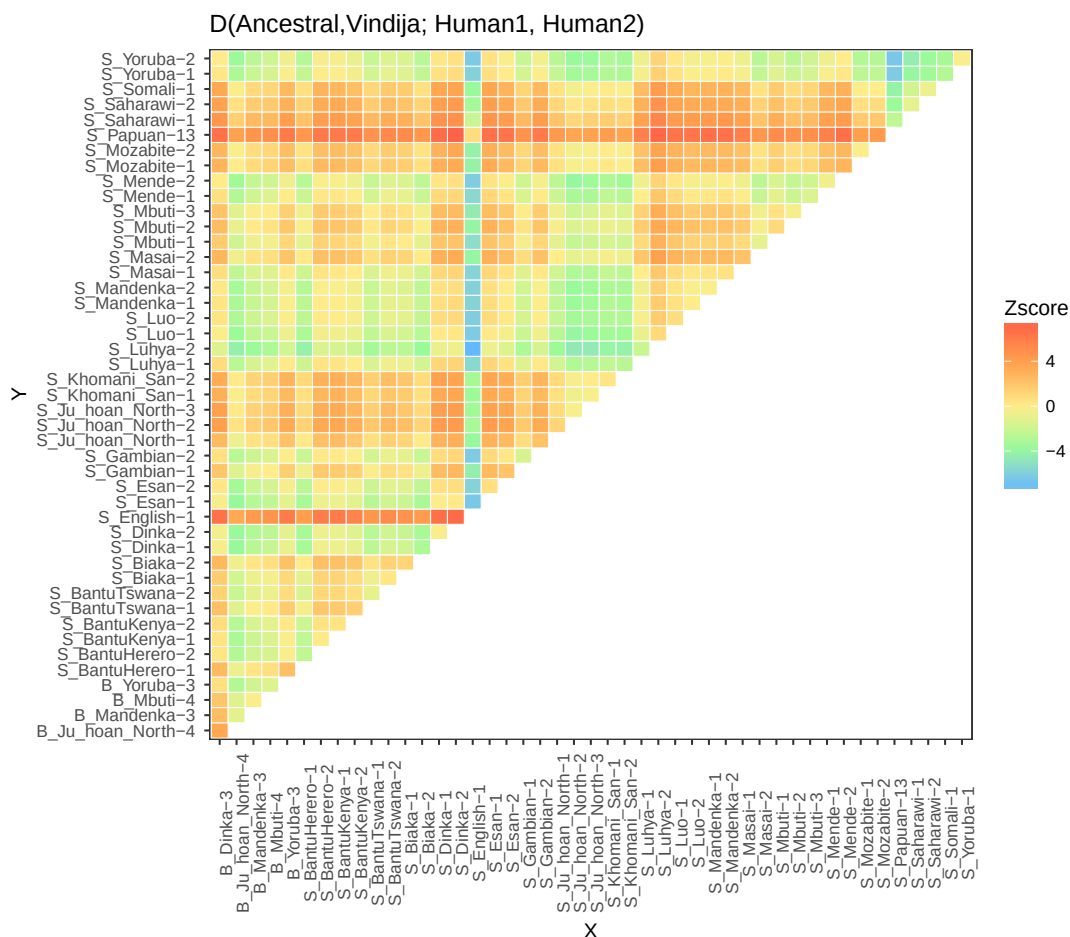


Fig. 5.21.: *D*-statistics for a pair of individuals to observe excess allele sharing with Vindija while the Ancestral remains as the outgroup. Only positions where Vindija had privately derived mutations were used. If the value is positive (red) the individual on the Y-axis is closer to Vindija than the individual on the X-axis.

5.8 Discussion

In the age of big data only a handful of studies have searched for evidence of archaic introgression specifically in Africa (Wall *et al.*, 2009; Hammer *et al.*, 2011; Lachance *et al.*, 2012; Hsieh *et al.*, 2016a; Durvasula & Sankararaman, 2018; Ragsdale & Gravel, 2018). All of these studies have focused, at most, on three African populations: KhoeSan, central African RHG, and West Africans (i.e. Yoruba, Esan, Mende, Mandenka). Here I have described, for the first time, the investigation of archaic sequences in high-coverage genomes of 44 African individuals representing 19 populations. Furthermore, I have shown for the first time results directly comparing confirmed signatures of introgression in Papuans to the candidate regions detected in African populations, by controlling for confounding factors such as genomic recombination and sample size. While previous studies have claimed to have found archaic admixture in some of the same populations we included (Hammer *et al.*, 2011; Lachance *et al.*, 2012; Hsieh *et al.*, 2016a), none have shown compelling evidence without using model-based inference approaches (Hsieh *et al.*, 2016a) or characterised in detail the properties of detected haplotypes (Lachance *et al.*, 2012).

The results from Chapter 4 supported the indication by Vernot and Akey (Vernot & Akey, 2015) that S^* values are dependent on what demographic model is considered, the variation in recombination rate across the genome in different populations, and the number of variants present in a region. In light of this, it must be considered that S^* scores alone are not comparable when looking at populations that have passed through different demographic histories, and as such cannot be used in isolation as evidence that archaic admixture may have taken place.

I have consistently observed that fragments identified as introgressed in African populations have larger TMRCA's but not larger median Sprime/ S^* scores than in Papuans (Figure 5.6; Figure 5.7; Figure 5.9). Additionally, as reported by others (Lachance *et al.*, 2012), I have confirmed that the top-scoring regions have TMRCA's older than the rest of the genome. Moreover, I have shown that the correlation between S^* scores and TMRCA's, while significant in simulated data (Figure 4.10), explains only a minor fraction of the total variation and does not appear to be significantly correlated with Sprime in real data (Figure 5.5). Such results could be interpreted as supporting archaic introgression in Africans from a source having a TMRCA with humans older than the TMRCA with Neanderthal and Denisovans.

However, as shown in Chapter 4, older TMRCA's for top-scoring fragments are identified even when no introgression has occurred and cannot be used as the sole evidence in support of archaic gene-flow as previous studies have done (Lachance *et al.*, 2012). In the simulations reported in Chapter 4, I have shown that in populations experiencing

archaic introgression followed by bottlenecks, as for the Papuans, the degree of reduction in the distribution of introgressed fragments is such that it would appear smaller when compared to populations with no introgression and no bottleneck (Figure 4.14; Appendix Figure A.2). Coupled with the fact that changes in growth rate are expected to increase the upper shift in the distribution of top-scoring fragments of non-introgressed populations, it appears that the major differences in Sprime/ S^* distributions I observed between African and Papuan populations could be reconciled just by demographic dynamics and are therefore compatible with both archaic contribution and demographic history scenarios without admixture.

It is also important to note that haplotypes detected in African populations were exceptionally long (Figure 5.15), larger than those previously identified in Papuans (Browning *et al.*, 2018). While relatively long fragments have often been used as evidence for archaic introgression when comparing LD patterns in non-Africans (i.e. from a few tens of thousands to a few hundreds of thousands of base pairs) (Sankararaman *et al.*, 2014; Sankararaman *et al.*, 2016; Fu *et al.*, 2014), the extremely long fragments detected by Sprime alone would imply admixture events with an archaic having occurred only a few generations ago, similar to the large fragments detected in the ancient Eurasian sample who had a Neanderthal ancestor between 4 to 6 generations earlier (Fu *et al.*, 2015). Using LD-based methods to detect introgression may lead to unexpectedly long haplotypes as a consequence of the high levels of variation within African populations. Because LD-based methods often require a filtering step where shared human variation (i.e. shared human-specific SNPs) is controlled for, this step is less thorough when using Eurasians as an outgroup for African populations. In fact non-African populations have only a subset of African variation and therefore variants uniquely present in Africa due to the demographic history of populations of the continent would not be accounted for. Despite using African-ancestry specific recombination maps (The HapMap Consortium *et al.*, 2010; Hinch *et al.*, 2011), it is interesting to note that most of the fragments detected in Africans are shared between populations (Figure 5.11) with few unique haplotypes, although some that are notably old (Figure 5.11, Figure 5.12). A possible explanation for this observation could be that archaic admixture took place very early in human history, possibly before or around the time the ancestors of the San diverged from the other African lineages approximately 300,000 years ago (Skoglund *et al.*, 2017). Alternatively, gene flow might have occurred in one or a subset of populations and then spread to the other groups following contacts among African groups. Another possibility for observing these top scoring fragments across African populations, as previous studies have supported, western Africans having ancestry from a basal African lineage that contributed differently to the Mende and to the Yoruba (Skoglund *et al.*, 2017) and therefore suggesting a basal African lineage that is not homogenous, potentially due to archaic introgression.

However the presence of these long haplotypes is not compatible with an extremely old contribution from an archaic source so far in the past and points to some methodological limitations in the detection of such regions, as mentioned above.

The *D*-statistics (ABBA-BABA) provided an independent way to test for variation in affinities of African and non-African to Eurasian archaic genomes. When performed on African individuals of the SGDP, results showed that, with the exception of Papuans, the San and the central African Pygmies deviated the most towards Denisovans (Figure 5.17), sharing an excess of derived alleles. Populations that have been most affected by Eurasian admixture showed an excess sharing of derived alleles with Vindija Neanderthal as would be expected (Figure 5.16; Figure 5.21; Figure S7.3). However a striking find was that the San and the Pygmies showed a stronger affinity with Denisovans than with Vindija Neanderthal (Figure 5.18). Moreover, I showed that this affinity is not shared with alleles shared only among Denisova and Neanderthal (Figure 5.19) but is shared with private derived alleles to Denisovans (Figure 5.20). There is some affinity between the San and central African Pygmies with Vindija when focusing only on Neanderthal private derived alleles (Figure 5.21) but not as strongly as with Denisovans. An earlier study (Povysil & Hochreiter, 2016) has shown that western and eastern African populations share a large proportion of identity-by-descent segments with both Denisovans and Neanderthals, which they interpret as ancestors of humans and ancient hominins having interbred within Africa before the Out-of-Africa migration and may have been common across the continent. However the admixture scenario is not the only one compatible with these observations. Another possible interpretation of these results which cannot be ruled out is ILS. A demographic model where Neanderthals experienced more drift than Denisovans, as suggested by differences in heterozygosity between Neanderthal and Denisovan genomes (Prufer *et al.*, 2014; Prufer *et al.*, 2017; Reich *et al.*, 2010), might explain why ancient shared polymorphisms might have been preserved more in Denisovans than Neanderthals. A larger effective population size for Khoesan and central African Pygmies, as suggested by Schlebusch *et al.* (2012) and Lopez *et al.* (2018), might similarly explain a larger presence of ancient shared polymorphisms in these two populations. The combination of these two scenarios would explain the variation in affinities for Denisovans across African populations as well as older TMRCAs for top scoring regions detected in these groups (Model depicted in Figure 5.22).

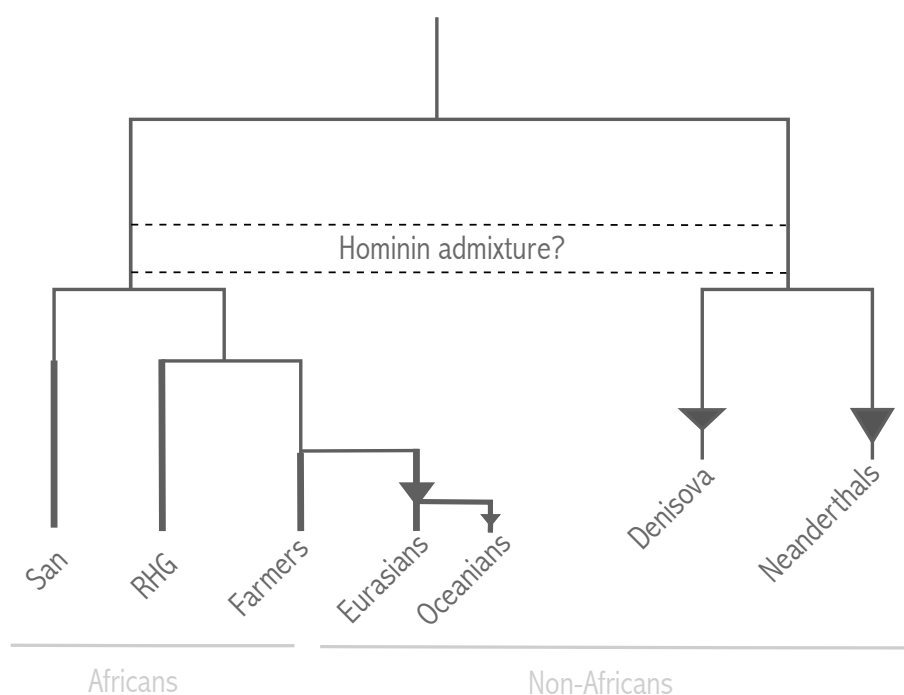


Fig. 5.22.: Proposed model. Schematic depicting combined demographic histories of both modern humans and Eurasian archaics which may best explain the strong presence of ancient shared polymorphisms in African populations with largest effective population size. Allele sharing with Denisovans more than Neanderthal for some African populations may be explained by the stronger bottleneck in Neanderthal (Prufer *et al.*, 2014; Prufer *et al.*, 2017). Triangles indicate the occurrence of bottlenecks. Dotted lines suggest potential gene flow at the level of the basal lineages.

In conclusion, my results can neither reject nor accept a scenario of archaic admixture in Africa. However, collectively seeing my results in light of previous studies (Skoglund *et al.*, 2017; Povysil & Hochreiter, 2016; Durvasula & Sankararaman, 2018; Speidel *et al.*, 2019) it is compatible with a model where admixture may have taken place among the ancestors of AMH or possibly at the level of the basal lineages.

Conclusion

” Typically, whereas Africa is emphasised in those earlier parts of world prehistory that are necessarily common to all human beings, it is excluded from consideration once the magic moment is reached at which hominins (in more recent syntheses, anatomically modern humans) expanded into Eurasia. All too often, the result is a history that confers universal validity and value on the past of Euro-American societies alone.

— **Lawrence Barham & Peter Mitchell**
The First Africans

The aim of this thesis was to investigate the presence of foreign DNA, from both archaic admixture and ancient retroviral infection, in our own species, *Homo sapiens*. Here I discuss the relevance of my results for the understanding of evolutionary virology and human evolutionary genetics.

In the first part of this dissertation I presented a method to detect mobile elements, specifically used to target HK2 integrations in high throughput sequencing data (HTS) (Chapter 2). I showed what the difficulties are in searching for mobile elements in HTS and why making a software flexible in accepting different formats and library insert sizes is of essence in order to benefit from high coverage genomes which have become more accessible in recent years. STEAK was developed to resolve the aforementioned challenges and has shown to be specific in finding HERVs, non-LTR retrotransposons, and exogenous retroviruses (based on simulated data). However some of the limitations of STEAK are precisely due to the nature of many HTS datasets. Short-reads make it difficult to determine whether the integration found, in the case of HK2, is a solo LTR or a full provirus. Distinguishing between the two often requires wet-lab validation. Wildschutte *et al.* (2016) use a k-mer based method to reconstruct some of the proviral genes which can benefit in silico characterisation of similar integrations and which in future studies would be worth implementing.

In Chapter 3, I went on to implement STEAK on 70 high coverage WGS from a cancer cohort (TCGA). These results in conjunction with data from a previous study (Wildschutte *et al.*, 2016) that looked at polymorphic HK2 integrations worldwide (N=2,453), allowed me to estimate the most recent integration rate for HK2 in the past 5 million years. This rate of about 1 integration for every 9,091 births along with the observed global frequencies of H2 integrations challenge previous studies (Marchi *et al.*, 2013) in pushing the date of HK2 activity cessation to more than 260,000 years ago. The youngest intact provirus found in the Mbuti Pygmies is shared with Denisovans and suggests that cessation of activity may have stopped prior to the emergence of anatomically modern humans (AMH) (Schlebusch *et al.*, 2012; Tattersall & Schwartz, 2009).

In order to pinpoint a more exact date of cessation for HK2 germline activity, we would need a model which accounts for host recombination, HK2 dynamics as both a virus and a mobile element, selection pressures on both the host and HK2, and finally the demographic history of an ancestral population. Developing this type of model would make it possible to appropriately answer this question, more so than the estimates used here and in previous studies (Marchi *et al.*, 2013; Belshaw *et al.*, 2005).

Why is it important to know when HK2 ceased germline activity in the distant past? Given the successful trajectory of HK2 in the last 40 million years, it would be interesting to know whether the cessation of HK2 is related to the emergence of AMH or if HK2 posed a selection pressure for early *Homo* populations. Given the number of solo LTRs compared to full proviruses found in the human genome (about 1:40 ratio) (Subramanian *et al.*, 2011), HK2 must have posed a pathogenic burden, albeit with low virulence, having infected the germline for so long. More intriguing is the possibility of HK2 exaptation. One study found HK2 viral particles to be present during early human development in embryogenesis (Grow *et al.*, 2015), possibly providing an immunoprotective effect for human embryos against different classes of viruses which are sensitive to IFITM1-type restriction—a protein that prevents infection before a virus manages to traverse the lipid bilayer of a cell (Grow *et al.*, 2015).

African populations retain some of the youngest HK2 integrations but have been, with the exception of a few studies (Wildschutte *et al.*, 2016; Jha *et al.*, 2011), largely ignored. Compellingly, the highest frequencies of rare HK2 integrations are found in populations with West African ancestry and in the Mbuti Pygmies (Wildschutte *et al.*, 2016). Studies have shown that during the dispersal through the rainforest, western Bantu-speaking groups encountered local populations of rainforest hunter-gatherers with whom they admixed 800 ya and in this gene flow they acquired beneficial alleles in the *HLA* gene region (Patin *et al.*, 2017). Although by this time HK2 was no longer active in the germline, it may be possible that these integrations were obtained through admixture between farmers and hunter-gatherers. An interesting way of seeing if integrations

may have been acquired through admixture and propagated across populations via host recombination is by scanning the genome for local excesses of different ancestries, observing admixture proportion in different chromosomes (Busby *et al.*, 2016). This method would also be useful when observing novel integrations to see when they have integrated into the genome.

Even though HK2 integrations have stopped replicating in the germline through means of reinfection and retrotransposition, understanding their drift in the population can allow us to see if they may be under negative or positive selection - particularly full proviruses with intact ORFs which, in theory, would still have the capacity to infect the host. Considering the different subsistence strategies found in Africa, it is compelling to see if putative HK2 pathogenicity may be related to post-reproductive age morbidity (e.g. cancer) (Hirschl *et al.*, 2007; Boller *et al.*, 1993; Dewannieux *et al.*, 2006; Grow *et al.*, 2015; Lower *et al.*, 1993) with a possible advantage during earlier stages of human life (e.g. embryogenesis) (Grow *et al.*, 2015). Alternatively, HK2 may serve as an immunological adaptation depending on the pathogenic environment the host finds itself in, similar to the relationship of sickle cell anemia and malaria. To explore these sort of questions, Africa is the ideal place of study for HK2.

The second half of this thesis focused on exploring current state-of-the-art methods to search for archaic introgression in modern human genomes. Some have referred to this as excavating for archaic genetic fossils in present-day humans (Vernot *et al.*, 2016). In particular I focused on how the LD-based method, S^* , behaves under different demographic models. While this method useful when knowing the demographic history of a population and/or when archaic hominin genomes are available, it can pose challenges for results interpretation when implemented on populations where the demographic history is unclear and there is no archaic reference available. In Chapter 4, I highlighted this by simulating different demographic models where by using S^* and TMRCA, result interpretation could be misleading. It has been previously noted that S^* scores are sensitive to demography. I showed that even in the absence of admixture, S^* can still provide high scores. This is especially true in certain growth scenarios, such as exponential growth. Consequently, for populations with complicated demographic histories both recently and in the deep past, S^* may give confounding results.

However, it is also important to also note some of the limitations of this work. When studying S^* behaviour in simulations I did not characterise regions that were truly introgressed in terms of other properties such as site frequency spectrum (SFS) or number of variants shared across a detected fragment as done in other studies (Vernot & Akey, 2015; Hsieh *et al.*, 2016a). Despite these limitations, I show that S^* should best be used with caution, especially in model-based approaches where the demographic history is unclear as in the case of African populations (Hsieh *et al.*, 2016a). In light of the results I show demonstrating that S^* correlates with older haplotypes even in

the absence of admixture, these concerns are augmented when considering studies that have assumed an older detected haplotype as equating to introgression for African populations (Lachance *et al.*, 2012). TMRCA estimates while useful can be misleading when exploring the possibility of introgression without understanding the demographic history of a population. These sort of studies might benefit from methods such as the recently developed Relate (Speidel *et al.*, 2019) which has been shown to have higher accuracy in estimating TMRCA and also calculating variable historical population sizes.

Building on my understanding of S^* , in Chapter 5 I explored African data from the Simon Genome Diversity Project (SGDP) with S^* , Sprime, and by estimating TMRCA of putatively introgressed fragments. Unlike S^* , Sprime seemed to show no significant correlation with TMRCA, but the score seemed to be more telling of true positives. This might be because Sprime takes into account the levels of recombination in the genome (Browning *et al.*, 2018). It has been found that introgression tends to be more common in regions of high recombination where there is linkage to fewer putative targets of selection (Schumer *et al.*, 2018). I accounted for differences in recombination by using two different genetic maps for African-ancestry. Overall, this chapter highlighted the limitations of not having an archaic reference to validate detected haplotypes against even if I compared with a population that has been confirmed to have archaic introgression (i.e. Papuans). It is clear that using non-African populations as an outgroup for African target populations is not sufficient to filter out human-derived alleles for performing LD-based methods.

In absence of an archaic reference, I used D-statistics to make sense of the patterns observed with Sprime. Particularly I compared African populations to Eurasian archaic genomes (Neanderthal and Denisovan). The rationale behind this is that in the absence of archaic introgression, two African individuals should be closer to each other than to a Eurasian archaic hominins. However if we could observe an African population that shares more derived alleles with an archaic, this could either be incomplete lineage sorting or archaic admixture. I find that African populations share more alleles with Denisovans than they do with Neanderthal (Vindija). An earlier study that looked at identity-by-descent (IBD) in African populations had found that both West and East Africans share a significant amount of IBD segments with both Eurasian hominins (Povysil & Hochreiter, 2016). My results suggest that the populations that share the most with Neanderthal are those that have been influenced by Eurasian back-migration whereas populations like the San share the most with Denisovans, especially Denisovan private derived alleles. It is not possible to disqualify incomplete lineage sorting for the patterns observed with both LD-based methods and D-statistics at this time. However a parsimonious explanation for the results observed can be best explained in Figure 5.22 where the respective bottlenecks in Denisovans and Neanderthal are influencing the signatures I observed. That is to say that the larger Neanderthal bottleneck removed

many of the alleles that African populations may have shared with this archaic hominin. Denisovans, on the other hand, have suffered from a smaller bottleneck potentially retaining a diversity that is from a ghost ancestral population in Africa.

A recent preprint using a tree-based approach found that Africans shared less than 20% of alleles with Neanderthals and Denisovans. Notably, this study focused on African populations from the 1000 Genomes Project which is primarily made up of West Africans, particularly Yoruba, and some east Africans (Luhya), which we also find to have little allele sharing with both archaic hominins. Using their method, Relate, Speidel *et al.* (2019) find that the Yoruba have contribution from an archaic hominin that is not closely related to Neanderthal. It would be interesting for this method to also be used on other African populations like the San and central African Pygmies in order to assess if the genetic contributions come from multiple different hominin groups although, for this sort of algorithm, more genomes than the amount analysed here would be better. Remarkably, even with a small sample size, my results also show that Sprime detected haplotypes unique to West Africans tend to be older than both shared and other unique haplotypes. Nevertheless, these results were not shown to be statistically significant. Skoglund *et al.* (2017) show that some western African populations have received contribution from a more deeply diverging group than the San, specifically, the Mende of Sierra Leone. They display a stronger contribution from this group than the Yoruba. Skoglund *et al.* (2017) refer to this group as a basal human lineage which separated before other sub-Saharan Africans.

I find that the most parsimonious model that explains my results is one where admixture took place in Africa among the ancestors of AMH, most likely at the level of the African basal lineage (Figure 5.22). In contrast to what other studies have suggested (Hammer *et al.*, 2011; Lachance *et al.*, 2012; Hsieh *et al.*, 2016a), I find no compelling support for recent archaic introgression. However an ascertainment of whether recent archaic introgression took place in Africa can only be answered with the sequencing of an African archaic hominin or a better understanding of the population structure of Africa prior the landmark migrations across the continent. The latter would also require more aDNA from ancient African human remains.

6.1 Future Work

Some of the questions that have emerged from my work during the DPhil are: what impact did both archaic admixture and ancient viral infections have for modern humans and why does it matter? Why have we retained this genetic material, or why has it not undergone purifying selection, and what purpose does it serve? How has it shaped

the genomic landscape of health and disease for populations today? Here I outline potential future work to answer some of these questions.

6.1.1 The Good: Adaptive introgression and viral exaptation

When searching for archaic introgression in African populations, both Lachance *et al.* (2012) and Hsieh *et al.* (2016a) found pronounced depletion of archaic sequence in genic regions suggesting that introgressed loci in hunter-gatherer populations are neutrally evolving remnants. However, a subsequent study (Xu *et al.*, 2017) has found that archaic introgression in Africa contributes to the variation in the human salivary gene *MUC7* consequently affecting the composition of the oral microbiome in modern Africans today. This study also suggested that copy number variation in *MUC7* has rapidly evolved under adaptive forces potentially shaped by pathogenic pressures among primates (Xu *et al.*, 2016). Furthermore, Durvasula & Sankararaman (2018) have detected a number of regions with putative archaic ancestry in Yoruba which are in genic regions and that may be under positive selection.

A number of studies have looked at how genetic ancestry and Neanderthal admixture have shaped immune responses to bacterial and viral pathogens (Quach *et al.*, 2016; Nedelec *et al.*, 2016). In particular it has been shown that admixture with Neanderthals has introduced regulatory variants into European genomes affecting responses to viral challenges, particularly RNA viruses (Quach *et al.*, 2016; Enard & Petrov, 2018). Performing similar investigations with putative archaic ancestry in African populations would also be of interest to better understand retention of archaic ancestry in relation to the health of those particular populations and the putative selection pressures behind it.

In a similar vein an interesting hypothesis could be that HK2 was part of a poison-antidote model for early forms of *Homo* given its expression of viral proteins leading to the activation of IFITM1-mediated restriction which protects germ cells from RNA viruses like influenza and West Nile virus (Grow *et al.*, 2015). One way to test for this is to search for associations between variants of HK2 (i.e. insertional polymorphisms, allelic variations, solo versus full-length) and human gene expression levels between different populations in order to see if these integrations may have regulatory effects related to cell type-specific immune responses (Wang *et al.*, 2017).

6.1.2 The Bad: Deleterious and pathogenic burdens

Existing research looking at African hunter-gatherer and farmer populations has explored how deleterious genetic variation among human populations is affected by

changes in population size and gene flow (Lopez *et al.*, 2018). Specifically hunter-gatherers are efficient in purifying selection despite having a recent population collapse (Simons *et al.*, 2014; Lopez *et al.*, 2018). Long-term selection against putative archaic introgression in Africa has not been looked at although recent studies have shown that non-African populations demonstrate selection against introgression in regulatory regions more than in protein-coding regions (Petr *et al.*, 2019). But more genomes (e.g. 1000 Genomes Project) will be required to map the genome-wide distribution of these regions and test for significance.

Comparatively, HK2 preferentially integrates into regulatory and gene-rich regions where there are high chances of disrupting normal transcription (Brady *et al.*, 2009). Exploring the drives of selection for or against full-length proviruses would be possible using whole-genome simulations to investigate a wide range of parameters for selection and HK2 replication in early humans. This would allow an understanding of HK2 depleted regions and solo LTR rich regions and if there are evident population-specific patterns.

6.1.3 The Ugly: Beyond SNPs and non-Africans

It is vital to point out that SNPs are the privileged biomarker of choice for the vast majority of tools that explore human history in high-throughput sequencing to answer questions about evolution and variation. While SNPs are undoubtedly useful genetic markers, structural variants (SVs) such as insertions, deletions, duplications, and inversions make up most of the variation between individuals (Sudmant *et al.*, 2015). Those with African-ancestry have been shown to exhibit significant difference in SV profile from other populations which is consistent with what has been observed with SNPs (Simons *et al.*, 2014; Sudmant *et al.*, 2015). However, the complexity of SVs present on both the individual and population level can often go unappreciated first due to reference-bias (Sherman *et al.*, 2019; Sudmant *et al.*, 2015) but also because SVs are not commonly considered when developing tools that measure genetic variation for non-clinical purposes (Santander *et al.*, 2017). This reference bias can make it challenging to observe what other regions of the genome are shared between certain populations and archaics, whether because of ancestral structure or introgression (Gardner *et al.*, 2017; Günther & Nettelblad, 2018; Sherman *et al.*, 2019). Sherman *et al.* (2019) report that by assembling a genome created from 910 individuals of African descent, they find an additional ~10% more DNA compared to the latest version of the human reference, GRCh38. This further supports the concern of the inherent bias when using a reference genome that is not fully representative of all human populations. Accounting for the additional DNA that African populations hold would be useful for evaluating archaic admixture in Africa. SNPs which are in comparison to a reference with less variation cannot be used to account for the large insertional segments that may be archaically

introgressed because they may be represented by only a few SNPs and therefore bias results obtained for African populations.

To conclude, this thesis shows that in order to answer the outstanding questions related to archaic introgression in Africa and the impact of HK2 in human evolutionary history, we must account for underrepresented populations and underrepresented parts of the genome.

Appendices

A.1 Supplementary Material for Chapter 2

Title: STEAK: a specific tool for transposable elements and retrovirus detection in high-throughput sequencing data.

Authors

Cindy G. Santander^{1,*}, Philippe Gambron², Emanuele Marchi³, Timokratis Karamitros¹, Aris Katzourakis¹, and Gkikas Magiorkinis^{1,4*}

Affiliations

1. Department of Zoology, University of Oxford, Oxfordshire, United Kingdom
2. Science and Technology Facilities Council, Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Didcot, Oxfordshire, United Kingdom
3. Nuffield Department of Medicine, University of Oxford, Oxfordshire, United Kingdom
4. Department of Hygiene, Epidemiology and Medical Statistics, Medical School, National and Kapodistrian University of Athens, Greece

* Correspondence to (G.M.) gkikasmag@gmail.com and (C.S.) cindy.santander@zoo.ox.ac.uk

This PDF file includes the following supplementary material:

- Supplementary Notes 1, 2, 3, and 4
- Supplementary Tables 1, 2, 3 and 4
- Supplementary Figures 1 and 2

Note 1

Benchmarking software for comparative evaluation

STEAK was written with the intention of looking for LTR transposons such as HERVs. The algorithm designed was meant to detect and visualise TE insertions as well as retroviral integrations. Our decision to use *RetroSeq* (Keane et al. 2013), *Tangram* (Wu et al. 2014) and *VirusSeq* (Chen et al. 2013) rose from how widespread their usage is (Table S3), what features they represent, and how appropriate they were in detecting HERV-K. *RetroSeq* is written in Perl and makes use of paired-end reads to determine novel TE integrations.

Tangram is a multithreaded program written in C++ and was designed to look for TE in both PE and SE libraries. Unfortunately it is no longer maintained and has difficulties handling BAM files not aligned with a specific version of *MOSAIK*.

VirusSeq is a pipeline that makes use of *MOSAIK* (Lee et al. 2014) (C++) and scripts written in Perl to determine virus presence and detect integration sites. However it was unable to recover HERV-K (HML-2) integrations (Note 3).

Whilst benchmarking for presence/absence with *t-lex2* (Fiston-Lavier et al. 2015) we discovered that it is unable to handle high-coverage genomes. We ran on ARC facilities *tlex2* pipeline on NA12878 using 256 GB of memory, 16 processes after replacing *blat* with *pblat* in the *tlex2* pipeline with several runs that exceeded and timed-out after 120 hours. Worryingly, *tlex2* does not produce errors or warnings when it runs out of memory and also produces differing batches of intermediate files depending on the memory availability suggesting that it was not meant to handle copious amounts of HTS data.

ViralFusionSeq (VFS) (Li et al. 2013) is a competitive virus detection pipeline with its implementation in Perl. On the simulation of chromosome 1 it detected all simulated HIV integrations created with their own simulator. However, VFS like *tlex2* is unable to handle high coverage HTS and is mainly meant for RNA-Seq data. Similarly VFS timed out on NA12878 and did not complete searching for HERV-K in 120 hours.

The literature for *t-lex2* and VFS suggest that these pipelines have high sensitivity and specificity but their implementation and file handling are best suited for low coverage genomes.

Other relevant software include *Mobster* (Thung et al. 2014) and *Jitterbug* (Henaff et al. 2015). *Mobster* and *Jitterbug* are algorithms already similar to *RetroSeq* in that their main mission is to sensitively detect *de novo* TE insertions. Other pipelines like *TEA* (Lee et al. 2012), *TraFiC* (Tubio et al. 2014), and *TranspoSeq* (Helman et al. 2014) were not deemed appropriate because they tend to focus on tumour/normal pair which we do not analyse in our study.

Supplementary Table 1. Relevant software considered for comparative benchmarking.

Software	Year Published ^a	Year Released ^b	Number of Citations [*]
RetroSeq	2013	2013	41
VirusSeq	2013	2013	38
VFS	2013	2013	24
Tangram	2014	2014	12
Mobster	2014	2013 ^c	9
T-lex2	2015	2012 ^d	9
Jitterbug	2015	2015 ^e	3
MELT	2015	2016 ^f	Not available

^a Year in which corresponding manuscript was published

^b Year that software was initially released and made available to public use

^c Retrieved from documentation on Sourceforge (2014).

^d Retrieved from documentation on T-lex web page (2011).

^e Retrieved from documentation on GitHub (2015).

^f Retrieved from documentation on MELT web page (2016).

^{*}Citation record of corresponding manuscripts as provided by “Citation Network” in *Web of Science Core Collection*. (Reuters)

Supplementary Table 2. The Cancer Genome Atlas 70 samples analysed in this study

Sample	Disease	Centre	Assembly	Coverage
TCGA-DK-A1AG-10A-01D-A13W-08	Bladder Urothelial Carcinoma	Broad Institute	HG19 Broad Variant	39×
TCGA-HW-7486-10A-01D-2024-08	Brain Lower Grade Glioma	Broad Institute	HG19 Broad Variant	45×
TCGA-A2-A0D2-10A-01D-A128-09	Breast invasive carcinoma	Washington University School of Medicine	GRCh37-lite	39×
TCGA-A2-A04T-10A-01D-A128-09	Breast invasive carcinoma	Washington University School of Medicine	GRCh37-lite	39×
TCGA-EK-A2RM-10A-01D-A18J-09	Cervical squamous cell carcinoma & Endocervical adenocarcinoma	Washington University School of Medicine	GRCh37-lite	39×
TCGA-06-0145-10A-01D-0507-08	Glioblastoma multiforme	Broad Institute	HG19 Broad Variant	50×
TCGA-DQ-5625-10A-01D-A32X-10	Head and Neck squamous cell carcinoma	Baylor College of Medicine	GRCh37-lite	50×
TCGA-KM-8441-10A-01D-2311-10	Kidney Chromophobe	Baylor College of Medicine	GRCh37-lite	50×
TCGA-DV-5566-10A-01D-2104-10	Kidney renal clear cell carcinoma	Baylor College of Medicine	GRCh37-lite	45×
TCGA-MH-A562-10A-01D-A26P-10	Kidney renal papillary cell carcinoma	Baylor College of Medicine	GRCh37-lite	50×
TCGA-FF-8062-10A-01D-2210-10	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	Baylor College of Medicine	GRCh37-lite	50×
TCGA-FF-8046-10A-01D-2210-10	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	Baylor College of Medicine	GRCh37-lite	50×
TCGA-13-0725-10B-01D-0446-08	Ovarian serous cystadenocarcinoma	Broad Institute	HG19 Broad Variant	50×
TCGA-13-0751-10A-01D-0446-08	Ovarian serous cystadenocarcinoma	Broad Institute	HG19 Broad Variant	50×
TCGA-HC-7233-10A-01D-2115-08	Prostate adenocarcinoma	Broad Institute	HG19 Broad Variant	39×
TCGA-G9-6336-10A-01D-1786-08	Prostate adenocarcinoma	Broad Institute	HG19 Broad Variant	39×
TCGA-EE-A2MI-	Skin Cutaneous	Broad Institute	HG19	39×

10A-01D-A199-08	Melanoma		Broad Variant	
TCGA-CD-5802-10A-01D-1600-08	Stomach adenocarcinoma	Broad Institute	HG19 Broad Variant	45×
TCGA-E8-A416-10A-01D-A23K-08	Thyroid carcinoma	Broad Institute	HG19 Broad Variant	50×
TCGA-AP-A053-10A-01D-A013-09	Uterine Corpus Endometrioid Carcinoma	Washington University School of Medicine	GRCh37-lite	39×
TCGA-A7-A0CE-11A-21D-A12L-09	Breast invasive carcinoma	Washington University School of Medicine	GRCh37-lite	50×
TCGA-A1-A0SM-10A-02D-A099-09	Breast invasive carcinoma	Washington University School of Medicine	GRCh37-lite	50×
TCGA-CH-5750-10A-01D-1576-08	Prostate adenocarcinoma	Broad Institute	HG19 Broad Variant	50×
TCGA-CH-5789-10A-01D-1576-08	Prostate adenocarcinoma	Broad Institute	HG19 Broad Variant	50×
TCGA-D3-A3MO-10A-01D-A21A-08	Skin Cutaneous Melanoma	Broad Institute	HG19 Broad Variant	50×
TCGA-BH-A0E0-11A-13D-A128-09	Breast invasive carcinoma	Washington University School of Medicine	GRCh37-lite	> 50×
TCGA-CH-5763-11A-01D-1576-08	Prostate adenocarcinoma	Broad Institute	HG19 Broad Variant	> 50×
TCGA-EE-A2M5-10A-01D-A199-08	Skin Cutaneous Melanoma	Broad Institute	HG19 Broad Variant	50×
TCGA-CH-5788-10A-01D-1576-08	Prostate adenocarcinoma	Broad Institute	HG19 Broad Variant	50×
TCGA-A6-2681-10A-01D-2188-10	Colon adenocarcinoma	Baylor College of Medicine	GRCh37-lite	> 50×
TCGA-D3-A1Q5-10A-01D-A198-08	Skin Cutaneous Melanoma	Broad Institute	HG19 Broad Variant	50×
TCGA-06-0125-10A-01D-1490-08	Glioblastoma multiforme	Broad Institute	HG19 Broad Variant	> 50×
TCGA-A6-2683-11A-01D-1554-10	Colon adenocarcinoma	Baylor College of Medicine	GRCh37-lite	> 50×
TCGA-FF-8041-10A-01D-2210-10	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	Baylor College of Medicine	GRCh37-lite	> 50×

TCGA-FF-8042-10A-01D-2210-10	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	Baylor College of Medicine	GRCh37- lite	> 50×
TCGA-FF-8043-10A-01D-2210-10	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	Baylor College of Medicine	GRCh37- lite	> 50×
TCGA-FF-8061-10A-01D-2210-10	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	Baylor College of Medicine	GRCh37- lite	> 50×
TCGA-X2-A95T-10A-01D-A37F-09	Sarcoma	Washington University School of Medicine	GRCh37- lite	> 50×
TCGA-AP-A0LE-10A-01D-A10B-09	Uterine Corpus Endometrioid Carcinoma	Washington University School of Medicine	GRCh37- lite	> 50×
TCGA-AP-A0L9-10A-01D-A066-09	Uterine Corpus Endometrioid Carcinoma	Washington University School of Medicine	GRCh37- lite	> 50×
TCGA-AX-A1CI-11A-11D-A135-09	Uterine Corpus Endometrioid Carcinoma	Washington University School of Medicine	GRCh37- lite	> 50×
TCGA-L5-A4OR-11A-11D-A27G-09	Esophageal carcinoma	Washington University School of Medicine	GRCh37- lite	> 50×
TCGA-06-0185-10B-01D-0507-08	Glioblastoma multiforme	Broad Institute	HG19 Broad Variant	> 50×
TCGA-06-0648-10A-01D-0507-08	Glioblastoma multiforme	Broad Institute	HG19 Broad Variant	> 50×
TCGA-G3-A25T-10A-01D-A16V-10	Liver hepatocellular carcinoma	Baylor College of Medicine	GRCh37- lite	> 50×
TCGA-IG-A3YB-10A-01D-A247-09	Esophageal carcinoma	Washington University School of Medicine	GRCh37- lite	> 50×
TCGA-A6-3807-10A-01D-1459-10	Colon adenocarcinoma	Baylor College of Medicine	GRCh37- lite	> 50×
TCGA-LN-A49M-10A-01D-A27G-09	Esophageal carcinoma	Washington University School of Medicine	GRCh37- lite	> 50×
TCGA-06-0188-10B-01D-0373-08	Glioblastoma multiforme	Broad Institute	HG19 Broad Variant	> 50×
TCGA-BT-A20T-11A-11D-A14W-08	Bladder Urothelial Carcinoma	Broad Institute	HG19 Broad Variant	> 50×
TCGA-BT-A20P-11A-11D-A14W-08	Bladder Urothelial Carcinoma	Broad Institute	HG19 Broad Variant	> 50×
TCGA-BT-A20Q-11A-11D-A14W-	Bladder Urothelial Carcinoma	Broad Institute	HG19 Broad	> 50×

08			Variant	
TCGA-BR-4255-11A-01D-A290-08	Stomach adenocarcinoma	Broad Institute	HG19 Broad Variant	> 50×
TCGA-CG-5730-11A-01D-1600-08	Stomach adenocarcinoma	Broad Institute	HG19 Broad Variant	> 50×
TCGA-EL-A3MY-11A-12D-A21A-08	Thyroid carcinoma	Broad Institute	HG19 Broad Variant	> 50×
TCGA-10-0937-11A-01D-A324-10	Ovarian serous cystadenocarcinoma	Baylor College of Medicine	GRCh37-lite	> 50×
TCGA-69-7763-10A-01D-A46I-10	Lung adenocarcinoma	Baylor College of Medicine	GRCh37-lite	> 50×
TCGA-CA-6718-10A-01D-1835-10	Colon adenocarcinoma	Baylor College of Medicine	GRCh37-lite	> 50×
TCGA-KL-8332-11A-01D-2310-10	Kidney Chromophobe	Baylor College of Medicine	GRCh37-lite	> 50×
TCGA-KM-8438-10A-01D-2311-10	Kidney Chromophobe	Baylor College of Medicine	GRCh37-lite	> 50×
TCGA-KO-8406-11A-01D-2311-10	Kidney Chromophobe	Baylor College of Medicine	GRCh37-lite	> 50×
TCGA-NJ-A4YQ-10A-01D-A46J-10	Lung adenocarcinoma	Baylor College of Medicine	GRCh37-lite	> 50×
TCGA-KM-8639-10A-01D-2397-10	Kidney Chromophobe	Baylor College of Medicine	GRCh37-lite	> 50×
TCGA-KM-8477-10A-01D-2311-10	Kidney Chromophobe	Baylor College of Medicine	GRCh37-lite	> 50×
TCGA-KL-8343-11A-01D-2310-10	Kidney Chromophobe	Baylor College of Medicine	GRCh37-lite	> 50×
TCGA-KL-8327-11A-01D-2310-10	Kidney Chromophobe	Baylor College of Medicine	GRCh37-lite	> 50×
TCGA-FV-A31I-10A-01D-A22F-10	Liver hepatocellular carcinoma	Baylor College of Medicine	GRCh37-lite	> 50×
TCGA-CC-5261-10A-01D-A12Z-10	Liver hepatocellular carcinoma	Baylor College of Medicine	GRCh37-lite	> 50×
TCGA-CA-6717-10A-01D-1835-10	Colon adenocarcinoma	Baylor College of Medicine	GRCh37-lite	> 50×
TCGA-B3-3926-11A-01D-1253-10	Kidney renal papillary cell carcinoma	Baylor College of Medicine	GRCh37-lite	> 50×

Supplementary Table 3. HK2 integrations searched for in samples.

Chromosome	Position		Repetitive element in vicinity	Reference or Non-reference	Literature	Status
1	1345186	1346153	MER74B	Reference	Subramanian et al. (2011)	
1	52471916	52472883	LTR8	Reference	Subramanian et al. (2011)	
1	66890497	66891457	HERV9-int	Reference	Subramanian et al. (2011)	
1	75842771	75849143	L1PA2	Reference	Subramanian et al. (2011)	Polymorphic
1	79792629	79792630	AluSz	Non-Reference	Wildschutte et al. (2016)	Polymorphic
1	93742338	93743306		Reference	Subramanian et al. (2011)	
1	106015874	106015881		Non-Reference	Marchi et al. (2014)	Polymorphic
1	111802591	111802598	MLT1A1	Non-Reference	Marchi et al. (2014)	Polymorphic (Denisova)
1	155596457	155605636	HERV17-int	Reference	Subramanian et al. (2011)	
1	156149013	156149981	AluSx1	Reference	Subramanian et al. (2011)	
1	223578303	223578310	L1MDa	Non-Reference	Marchi et al. (2014)	Polymorphic
1	224527532	224528500	L1M4	Reference	Subramanian et al. (2011)	
2	27682845	27683813		Reference	Subramanian et al. (2011)	
2	30836337	30837305	L2a	Reference	Subramanian et al. (2011)	
2	37452460	37453428	MIRb	Reference	Subramanian et al. (2011)	
2	130719538	130722209		Reference	Subramanian et al. (2011)	
2	207901556	207902517	MIRc	Reference	Subramanian et al. (2011)	
2	229228357	229229325	L1MD	Reference	Subramanian et al. (2011)	
2	231708353	231709311		Reference	Subramanian et al. (2011)	
2	232441494	232442454		Reference	Subramanian et al. (2011)	
3	14132684	14133652		Reference	Subramanian et al. (2011)	Polymorphic
3	47301023	47301991		Reference	Subramanian et al. (2011)	
3	50557322	50558279		Reference	Subramanian et al. (2011)	
3	54011540	54012508		Reference	Subramanian et al. (2011)	
3	75794210	75794245		Non-Reference	Lee et al. (2012)	Polymorphic
3	94943488	94943489	L1PA10	Non-Reference	Wildschutte et al. (2016)	Polymorphic
3	112743479	112752282	MLT2C1	Reference	Subramanian et al. (2011)	Polymorphic
3	125609302	125618416	AluYa8	Reference	Subramanian et al. (2011)	
3	129360117	129360905		Reference	Subramanian et al. (2011)	
3	175623337	175624305	L1PA5	Reference	Subramanian et al. (2011)	
3	185280336	185289515		Reference	Subramanian et al. (2011)	
3	186611008	186611975	MLT1B	Reference	Subramanian et al. (2011)	
3	195654395	195655363	L1PB4	Reference	Subramanian et al. (2011)	
3	197845799	197845879	PRIMA41-int	Non-Reference	Lee et al. (2012)	Polymorphic
4	145520	146488	AluSq	Reference	Subramanian et al. (2011)	
4	9603239	9603245	HERVS71-int	Non-Reference	Marchi et al. (2014)	Polymorphic
4	9981605	9981606	L2b	Non-Reference	Wildschutte et al. (2016)	Polymorphic
4	63806595	63807563	LTR81B	Reference	Subramanian et al. (2011)	

4	72994940	72995908	L2a	Reference	Subramanian et al. (2011)	
4	120263688	120264654	AluSx	Reference	Subramanian et al. (2011)	
4	157225913	157226870		Reference	Subramanian et al. (2011)	
4	165826665	165827635		Reference	Subramanian et al. (2011)	
4	190966833	190966919	LTR43-int	Non-Reference	Lee et al. (2012)	Polymorphic
5	1596091	1597059		Reference	Subramanian et al. (2011)	
5	4537604	4537605	LTR1C	Non-Reference	Wildschutte et al. (2016)	Polymorphic
5	8937853	8938820	L1PA13	Reference	Subramanian et al. (2011)	
5	35176486	35177455	L1MB7	Reference	Subramanian et al. (2011)	
5	44730588	44731556	L1PA8	Reference	Subramanian et al. (2011)	
5	54866907	54867867		Reference	Subramanian et al. (2011)	
5	64388439	64388446	L1M6	Non-Reference	Marchi et al. (2014)	Polymorphic (Neanderthal)
5	74901657	74902617	L1MEd	Reference	Subramanian et al. (2011)	
5	80442265	80442272		Non-Reference	Marchi et al. (2014)	Polymorphic (Denisova & Neanderthal)
5	116156834	116157794	AluSx1	Reference	Subramanian et al. (2011)	
5	156084717	156093896	AluSx	Reference	Subramanian et al. (2011)	
5	169423878	169424838	AluYh9	Reference	Subramanian et al. (2011)	
5	178940963	178941931	L1ME3A	Reference	Subramanian et al. (2011)	
6	16004793	16004926		Non-Reference	Wildschutte et al. (2016)	Polymorphic
6	27742238	27743207		Reference	Subramanian et al. (2011)	
6	32643383	32643537	L1PA10	Non-Reference		Polymorphic
6	32648035	32648041	L1PA10	Non-Reference	Marchi et al. (2014)	Polymorphic
6	33777736	33778696		Reference	Subramanian et al. (2011)	
6	34687236	34688205		Reference	Subramanian et al. (2011)	
6	52787477	52788438	HERVE-int	Reference	Subramanian et al. (2011)	Polymorphic
6	78427019	78436083		Reference	Subramanian et al. (2011)	Polymorphic
6	79568504	79569472	AluSx	Reference	Subramanian et al. (2011)	
6	89091306	89092274		Reference	Subramanian et al. (2011)	
6	93883083	93884048	MLT1B	Reference	Subramanian et al. (2011)	
6	111576299	111577258		Reference	Subramanian et al. (2011)	
6	134980247	134981215		Reference	Subramanian et al. (2011)	
6	151776344	151777320		Reference	Subramanian et al. (2011)	
6	161270898	161270905		Non-Reference	Marchi et al. (2014)	Polymorphic (Denisova)
7	4622057	4640031		Reference	Subramanian et al. (2011)	Polymorphic
7	4630561	4640031	AluSx	Reference	Subramanian et al. (2011)	Polymorphic
7	16237346	16238314	MER67B	Reference	Subramanian et al. (2011)	
7	23079474	23080442	L1PA2	Reference	Subramanian et al. (2011)	
7	104388369	104393266	L1MC2	Reference	Subramanian et al. (2011)	
7	123420690	123421658		Reference	Subramanian et al. (2011)	
7	124861223	124862191		Reference	Subramanian et al. (2011)	
7	125808213	125809181		Reference	Subramanian et al. (2011)	

7	158029482	158030450		Reference	Subramanian et al. (2011)	
7	158773312	158773459		Non-Reference	Wildschutte et al. (2016)	Polymorphic
8	7355397	7364859	MIR	Reference	Subramanian et al. (2011)	Polymorphic
8	18651458	18652426		Reference	Subramanian et al. (2011)	Polymorphic
8	37050885	37051853		Reference	Subramanian et al. (2011)	
8	43594681	43595649		Reference	Subramanian et al. (2011)	
8	48087655	48088615	AluSx	Reference	Subramanian et al. (2011)	
8	54944240	54945208	L1MB5	Reference	Subramanian et al. (2011)	
8	58112161	58113123		Reference	Subramanian et al. (2011)	
8	91696259	91697227		Reference	Subramanian et al. (2011)	
8	140472149	140475236		Reference	Subramanian et al. (2011)	
8	144914158	144915127	HAL1	Reference	Subramanian et al. (2011)	
8	146021278	146022246		Reference	Subramanian et al. (2011)	
8	146086169	146086170		Non-Reference	Wildschutte et al. (2016)	Polymorphic
9	17445358	17446326		Reference	Subramanian et al. (2011)	
9	31632876	31633844		Reference	Subramanian et al. (2011)	
9	68277342	68278302	AluSx	Reference	Subramanian et al. (2011)	
9	72378052	72379020		Reference	Subramanian et al. (2011)	
9	111356892	111357852	MIR	Reference	Subramanian et al. (2011)	
9	124191383	124192351		Reference	Subramanian et al. (2011)	
9	132205208	132205208	MLT1C	Non-Reference	Lee et al. (2012)	Polymorphic
9	134232105	134233063		Reference	Subramanian et al. (2011)	
9	136957557	136958526	AluJb	Reference	Subramanian et al. (2011)	
10	27182399	27183380	MER34-int	Reference	Subramanian et al. (2011)	Polymorphic
10	42526055	42527021		Reference	Subramanian et al. (2011)	
10	43832650	43833618		Reference	Subramanian et al. (2011)	
10	70285428	70286393		Reference	Subramanian et al. (2011)	
10	101016044	101016228	MSTD	Non-Reference		Polymorphic
10	101580569	101587716		Reference	Subramanian et al. (2011)	Polymorphic
10	134444012	134444013	LTR46	Non-Reference	Wildschutte et al. (2016)	Polymorphic
11	10412855	10413823	L2b	Reference	Subramanian et al. (2011)	
11	24467639	24468598	MLT1F1	Reference	Subramanian et al. (2011)	
11	60449889	60449889	L1M4	Non-Reference	Marchi et al. (2014)	Polymorphic (Denisova)
11	61422484	61423448		Reference	Subramanian et al. (2011)	
11	61962131	61963100	MLT1D	Reference	Subramanian et al. (2011)	
11	63295442	63296410	MER21B	Reference	Subramanian et al. (2011)	
11	63297785	63298753		Reference	Subramanian et al. (2011)	
11	67370824	67371784	LTR26	Reference	Subramanian et al. (2011)	
11	67635434	67636394	L1PA4	Reference	Subramanian et al. (2011)	
11	71875417	71876385	MER61-int	Reference	Subramanian et al. (2011)	
11	101565794	101575259		Reference	Subramanian et al. (2011)	Polymorphic

12	6995023	6995985		Reference	Subramanian et al. (2011)	
12	9753060	9754028		Reference	Subramanian et al. (2011)	
12	30084986	30085954		Reference	Subramanian et al. (2011)	
12	32252444	32253412	L1MB8	Reference	Subramanian et al. (2011)	
12	44313656	44313662	L1MB1	Non-Reference	Marchi et al. (2014)	Polymorphic (Neanderthal)
12	51848070	51849038	AluJb	Reference	Subramanian et al. (2011)	
12	55727214	55728183		Reference	Subramanian et al. (2011)	Polymorphic
12	56794144	56795109	HERVL-int	Reference	Subramanian et al. (2011)	
12	58721242	58730698	L2b	Reference	Subramanian et al. (2011)	Polymorphic
12	105872336	105873305		Reference	Subramanian et al. (2011)	
12	111007843	111009325		Reference	Subramanian et al. (2011)	
12	118543788	118544748	AluY	Reference	Subramanian et al. (2011)	
12	124066476	124066483	AluSx1	Non-Reference	Marchi et al. (2014)	Polymorphic
13	20174357	20175325		Reference	Subramanian et al. (2011)	
13	50173157	50174126		Reference	Subramanian et al. (2011)	
13	90743182	90743189		Non-Reference	Marchi et al. (2014)	Polymorphic (Neanderthal)
14	55491267	55492221	L2c	Reference	Subramanian et al. (2011)	
14	65445305	65446273	LTR107_Mam	Reference	Subramanian et al. (2011)	
15	28430044	28430186		Non-Reference	Lee et al. (2012); Wildschutte et al. (2016)	Polymorphic
15	63374593	63374600	AluJr	Non-Reference	Marchi et al. (2014)	Polymorphic
15	65518879	65519847		Reference	Subramanian et al. (2011)	
15	89083782	89084750	HERVS71-int	Reference	Subramanian et al. (2011)	
15	102400179	102401147	LTR64	Reference	Subramanian et al. (2011)	
16	5804388	5805356		Reference	Subramanian et al. (2011)	
16	8228762	8229730		Reference	Subramanian et al. (2011)	
16	23610804	23611764	L2c	Reference	Subramanian et al. (2011)	
16	47898369	47899337		Reference	Subramanian et al. (2011)	
16	74833301	74834258	L1MB3	Reference	Subramanian et al. (2011)	
17	4977974	4978941		Reference	Subramanian et al. (2011)	
17	29027058	29028026	AluSx	Reference	Subramanian et al. (2011)	
17	52278934	52279902	L1ME4b	Reference	Subramanian et al. (2011)	
17	66600132	66601092	L2b	Reference	Subramanian et al. (2011)	
17	78525007	78525976	L2a	Reference	Subramanian et al. (2011)	
18	2000814	2001782	HAL1	Reference	Subramanian et al. (2011)	
18	4917277	4918246	THE1B-int	Reference	Subramanian et al. (2011)	
19	21841536	21841542		Non-Reference	Marchi et al. (2014)	Polymorphic (Denisova)
19	22414303	22414381	L1M3	Non-Reference	Lee et al. (2012)	Polymorphic
19	22457244	22457245	AluSq	Non-Reference	Wildschutte et al. (2016)	Polymorphic (Denisova)
19	28128498	28137361	AluYa5	Reference	Subramanian et al. (2011)	
19	29855781	29855787		Non-Reference	Marchi et al. (2014)	Polymorphic (Denisova)
19	36738045	36738826	AluSx	Reference	Subramanian et al. (2011)	

19	37822292	37823274	AluSq2	Reference	Subramanian et al. (2011)	
19	38120851	38121813	L1ME3A	Reference	Subramanian et al. (2011)	
19	38356816	38357784	AluJb	Reference	Subramanian et al. (2011)	
19	45097417	45098385		Reference	Subramanian et al. (2011)	
19	52546384	52547353		Reference	Subramanian et al. (2011)	
19	57996939	57996940		Non-Reference	Wildschutte et al. (2016)	Polymorphic (Neanderthal)
20	12402386	12402392		Non-Reference	Lee et al. (2012) & Marchi et al. (2014)	Polymorphic (Denisova)
20	33849926	33850894	L2c	Reference	Subramanian et al. (2011)	
20	40599535	40600503	L2a	Reference	Subramanian et al. (2011)	
21	19933916	19941962		Reference	Subramanian et al. (2011)	
21	44567239	44568199		Reference	Subramanian et al. (2011)	
22	18926187	18935307	AluJb	Reference	Subramanian et al. (2011)	
22	23852639	23852640	MLT1C	Non-Reference	Wildschutte et al. (2016)	Polymorphic
X	90205613	90206582	LTR17	Reference	Subramanian et al. (2011)	
X	93606603	93606604	L1MD1	Non-Reference	Wildschutte et al. (2016)	Polymorphic (Denisova)
X	124913492	124914460	L1PB	Reference	Subramanian et al. (2011)	
X	144802358	144803318		Reference	Subramanian et al. (2011)	
Y	6616929	6617896	L1PA10	Reference	Subramanian et al. (2011)	
Y	15215736	15216703	L1PB4	Reference	Subramanian et al. (2011)	
Y	25039185	25040153		Reference	Subramanian et al. (2011)	
Y	26672934	26673902		Reference	Subramanian et al. (2011)	
Y	27288494	27289462		Reference	Subramanian et al. (2011)	

Greyed out cells were not included in analysis to not introduce gender bias into evaluation of polymorphisms.

Supplementary Table 4. HK2 LTRs manually inserted into hg19 chromosome 20.

Chromosome	Position
20	126699
20	162699
20	402099
20	410543
20	413551
20	512092
20	544052
20	572945
20	588355
20	639693

Note 2

Technical considerations: Repetitive regions and pseudogenes

In addition to using *STEAK* for marking presence/absence of unfixed HK2 integrations, we were also able to recover an integration within the germline of a lung adenocarcinoma patient (Figure S2) that was not picked up by the other pipelines we benchmarked. This integration is particularly difficult to recover because its exact genomic location remains obscure. Originally in our analyses, reads belonging to this integration had mapped discordantly to different chromosomes. After processing with *STEAK*, a one-sided cluster, representing host reads flanking only one end of the LTR, was recovered in chromosome 3 in an exon of pseudogene FAM86DP (Figure 5A).

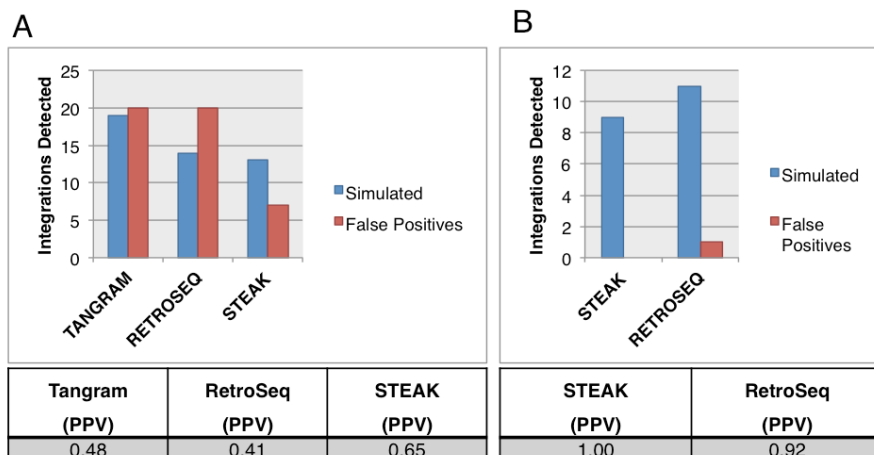
Eukaryotic elongation factor2 lysine methyltransferases (FAM86) is a family of genes and pseudogenes, which are dispersed throughout the human genome. Because of the various FAM86 copies, the cluster of reads prior to trimming had originally been mapped to different locations. By using the breakpoint interval observed with the trimmed reads produced by *STEAK*, we were able to create a contig of the unknown left flank of the HK2 integration and recover the original site. We produced a junction of the opposite side of the LTR with a predicted 4 bp target site duplication (TSD) and the rest of the exon. By remapping all reads against this contig, we were able to extract the reads that belonged to the other side of the HK2 integration and observed a 5 bp TSD of “CTTGG” (Figure S2B). While we were able to reconstruct the pre-integration site, we were still unable to confidently assign this integration a genomic location. Using BLAT we compared the pre-integration sequence against the hg19 genome. The top results with the highest percent identities ($\leq 96.9\%$) consisted of FAM86 copies. While this almost 4% divergence between our pre-integration site and a particular FAM86 gene or pseudogene leaves the precise genomic location of our novel integration unidentified, we can confirm that this particular integration is embedded in a FAM86-related gene or pseudogene.

Note 3: Pilot performance on simulated genomes

We had previously evaluated *STEAK*'s performance by testing two simulated paired-end NGS datasets. The first dataset consists of reads made from human chromosome 20 with artificial AluY insertions at 5X coverage originally created by *Tangram*. The second dataset is also derived from chromosome 20 but with artificial HK2 integrations and at 50X coverage. This second dataset was created to evaluate the performance in a high coverage genome and to benchmark against *VirusSeq*. However, on our simulated dataset we noted that no HK2 integrations were picked up on *VirusSeq*. As a test to confirm *VirusSeq* was truly not finding HK2 and that it was not a troubleshooting error, we created another simulated genome at 50X coverage with the same HK2 insertions with an additional HBV insertion. After running *VirusSeq* on this simulated genome, we observed that *VirusSeq* was only able to recognise the HBV integration but none of the HK2 integrations. This suggests that *VirusSeq* works best in finding viruses that are not found at all in the host reference.

Tangram was the most sensitive of the three programs performing on the 5X coverage genome. However, *STEAK* maintained 65% positive predictive value (PPV which is the probability that a suggested integration is true), whereas *RetroSeq* and *Tangram* achieved PPVs below 50% (Supplementary Figure 2A). We evaluated specificity prior to filtering for known insertions since both *RetroSeq* and *Tangram* are software that aim to search for non-reference putative insertions. We ran *STEAK* with parameters to search for non-reference insertions only. PPVs were calculated by comparing the number of true positive calls to the total number of positive calls (including false positives) when marking the simulated genomes.

On the 50X coverage dataset, *STEAK* gave no false positives compared to *RetroSeq* (Supplementary Figure 2B). However, two of the insertions within other repetitive elements, such as LINEs and SINEs, could not be recovered by *STEAK* whereas *RetroSeq* was able to do so (Supplementary Figure 2B). *VirusSeq* could not recover any HK2 integrations.



Supplementary Figure 2. Comparative performance of *Tangram*, *RetroSeq*, and *STEAK* on simulated data. (A&B) Blue indicates detection of simulated integrations and red indicated the number of false positive detected by each respective program. (A) Bar graph shows integrations detected in 5x coverage simulated genome with 20 artificial Alu insertions. The table below gives the positive predictive value (PPV) for evaluation of specificity. (B) Integrations detected in a 50x coverage simulated genome with 11 artificial HK2 integrations. The tables below A & B give the PPV for each program tested.

Note 4: Commands used for analyses and benchmarking

Tangram

-Simulation data script provided by Tangram works with their files from their repository (e.g. 5X coverage ALU simulation).

-Below are the commands used on real WGS data:

```
tangram_bam -i NA12878.platinum.bwa_merge.bam -r HK113LTR.fasta -o NA12878_tangram.bam
tangram_index -ref hs37d5.fasta -sp HK113LTR.fasta -out hs37d5_tangram.ref
tangram_scan -in tangram_scan.list -dir ./Tangram/
tangram_detect -lb ./Tangram/output_tangram/lib_table.dat //
-ht./Tangram/output_tangram/hist.dat -in tangram_bam.list -rg 20 //
-ref hs37d5_tangram.ref
```

##With MOSAIK aligner BAM:

```
MosaikBuild -fr hs37d5.fa.gz -oa hs37d5.bin
MosaikBuild -q NA12878_PE.fastq -out tangram_NA12878_PE.bin -st illumina
MosaikJump -ia hs37d5.bin -out hs37d5.jmp -hs 15
MosaikBuild -fr HK113LTR.fasta -oa HK113LTR.bin
MosaikJump -ia HK113LTR.bin -out HK113LTR.jmp -hs 15
MosaikAligner -ia hs37d5.bin -j hs37d5.jmp -sref HK113LTR -srefn 50 -annpe
./MOSAIK/src/networkFile/2.1.26.pe.100.0065.ann -annse ./MOSAIK/src/networkFile/2.1.26.se.100.005.ann -mmp
0.15 -mhp 200 -ls 476 -act 20 -bw 29 -p 8 -in tangram_NA12878_PE.bin -out tangram_NA12878_PE.bam
```

(Both end up in a Segmentation Fault when put through tangram_detect. Tangram was installed locally as well as a module on ARC. Both installations led to the same error.)

VirusSeq for 50x simulation with HK2

```
gzip COAD.1.fastq
gzip COAD.2.fastq
```

```
MosaikBuild -q $DATA/benchmark/COAD/COAD.1.fastq.gz -q2 $DATA/benchmark/COAD/COAD.2.fastq.gz -out
$DATA/benchmark/COAD/COAD_Virus.bin -st illumina
```

```
MosaikAligner -in $DATA/benchmark/COAD/COAD_Virus.bin -ia $DATA/benchmark/virusseq/jumpdb/hg19.fa.bin -out
$DATA/benchmark/COAD/COAD.bin.aligned -hs 15 -mmp 0.1 -mmal -minp 0.5 -act 25 -mhp 100 -m unique -j
$DATA/benchmark/virusseq/jumpdb/hg19.JumpDb -p 16 -km -pm -rur $DATA/benchmark/COAD/COAD_unalg.fq
```

```
MosaikJump -ia hg19Virus.fa.bin -out hg19Virus.JumpDb -hs 15 -mhp 100 -km
```

```
MosaikJump -ia hg19.fa.bin -out hg19.JumpDb -hs 15 -mhp 100 -km
```

```
MosaikJump -ia $DATA/benchmark/virusseq/jumpdb/hg19Virus.bin -out
$DATA/benchmark/virusseq/jumpdb/hg19Virus.jmp -hs 15 -mhp 100 -km
```

##for gib virus reference genome

```
MosaikBuild -fr $DATA/tools/VirusSeq/gibVirus.fa -oa $DATA/tools/VirusSeq/gibVirus.fa.bin -st illumina -
assignQual 40
```

```
MosaikJump -ia $DATA/tools/VirusSeq/gibVirus.fa.bin -out $DATA/tools/VirusSeq/gibVirus.JumpDb -hs 15 -mhp 100
-km
```

##version-0.89 and for combined hg19Virus reference genome

```
MosaikBuild -fr $DATA/benchmark/virusseq/jumpdb/hg19Virus.fa -oa
$DATA/benchmark/virusseq/jumpdb/hg19Virus.fa.bin -st illumina -assignQual 40
```

```
MosaikJump -ia $DATA/benchmark/virusseq/jumpdb/hg19Virus.fa.bin -out
$DATA/benchmark/virusseq/jumpdb/hg19Virus.JumpDb -hs 15 -mhp 100 -km
```

##for hg19 reference genome

```
MosaikBuild -fr $DATA/references/hg19/hg19.fa -oa hg19.fa.bin -st illumina -assignQual 40
```

```
MosaikJump -ia hg19.fa.bin -out hg19.JumpDb -hs 15 -mhp 100 -km
```

```
MosaikAligner -in $DATA/benchmark/simulated/750K/sim_Virus.bin -ia MosaikBuild -q
$DATA/benchmark/simulated/750K/HBV_HK2_sim_1.fq -out $DATA/benchmark/simulated/750K/sim50x_1.bin -st illumina
```

```
MosaikBuild -q $DATA/benchmark/simulated/750K/HBV_HK2_sim_2.fq -out
$DATA/benchmark/simulated/750K/sim50x_2.bin -st illumina
```

```
MosaikAligner -in $DATA/benchmark/simulated/750K/sim50x_1.bin -ia
$DATA/benchmark/virusseq/jumpdb/hg19Virus.fa.bin -out $DATA/benchmark/simulated/750K/sim50x_1.bin.aligned -hs
15 -mmp 0.1 -mmal -minp 0.5 -act 20 -mhp 100 -m unique -j $DATA/benchmark/virusseq/jumpdb/hg19Virus.JumpDb -p
16 -km -pm
```

```
MosaikAligner -in $DATA/benchmark/simulated/750K/sim50x_2.bin -ia
$DATA/benchmark/virusseq/jumpdb/hg19Virus.fa.bin -out $DATA/benchmark/simulated/750K/sim50x_2.bin.aligned -hs
15 -mmp 0.1 -mmal -minp 0.5 -act 20 -mhp 100 -m unique -j $DATA/benchmark/virusseq/jumpdb/hg19Virus.JumpDb -p
16 -km -pm
```

```

MosaikSort -in $DATA/benchmark/simulated/750K/sim50x_1.bin.sorted -out
$DATA/benchmark/simulated/750K/sim50x_1.bin.sorted

MosaikSort -in $DATA/benchmark/simulated/750K/sim50x_2.bin.sorted -out
$DATA/benchmark/simulated/750K/sim50x_2.bin.sorted

MosaikMerge -in $DATA/benchmark/simulated/750K/sim50x_1.bin.sorted -in
$DATA/benchmark/simulated/750K/sim50x_2.bin.sorted -out $DATA/benchmark/simulated/750K/sim50x_SE.sorted

MosaikText -in $DATA/benchmark/simulated/750K/sim50x_SE.sorted -axt
$DATA/benchmark/simulated/750K/sim50x_SE.sorted.axt

MosaikAligner -in $DATA/benchmark/simulated/750K/sim50x.bin -ia $DATA/benchmark/virusseq/jumpdb/hg19.fa.bin -
out $DATA/benchmark/simulated/750K/sim50x.bin.sorted -hs 15 -mmp 0.1 -mmal -minp 0.5 -act 25 -mhp 100 -m
unique -j $DATA/benchmark/virusseq/jumpdb/hg19.JumpDb -p 16 -km -pm -rur
$DATA/benchmark/simulated/750K/sim_unalg.fq

MosaikBuild -q $DATA/benchmark/simulated/750K/sim_unalg.fq -out $DATA/benchmark/simulated/750K/sim_Virus.bin -
st illumina

MosaikAligner -in $DATA/benchmark/simulated/750K/sim_Virus.bin -ia
$DATA/benchmark/virusseq/jumpdb/gibVirus.bin -out $DATA/benchmark/simulated/750K/sim_Virus.bin.sorted -hs 15
-mmp 0.10 -act 20 -mhp 100 -m all -j $DATA/benchmark/virusseq/jumpdb/gibVirus.jmp -p 16 -km -pm

MosaikSort -in $DATA/benchmark/simulated/750K/sim_Virus.bin.sorted -out
$DATA/benchmark/simulated/750K/sim_Virus.bin.sorted

MosaikAssembler -in $DATA/benchmark/simulated/750K/sim_Virus.bin.sorted -ia
$DATA/benchmark/virusseq/jumpdb/gibVirus.bin -out $DATA/benchmark/simulated/750K/sim_Virus.bin.assembled -f
ace > $DATA/benchmark/simulated/750K/sim_viruslog.txt

perl $DATA/benchmark/virusseq/scripts/VirusSeq_Detection.pl $DATA/benchmark/simulated/750K/sim_viruslog.txt
1000 $DATA/benchmark/simulated/750K/sim_virusName.txt

Spanner --scan --infile $DATA/benchmark/simulated/750K/sim50x.bin.viral --outdir
$DATA/benchmark/simulated/750K/SV_gDNA

Spanner --build --infile $DATA/benchmark/simulated/750K/sim50x.bin.viral --outdir
$DATA/benchmark/simulated/750K/SV_gDNA -f $DATA/benchmark/simulated/750K/SV_gDNA/MSK.stats -a
$DATA/benchmark/virusseq/jumpdb/Spanner_anchor_hg19Virus.txt -t

perl $DATA/benchmark/simulated/750K/SV_gDNA/Spanner_cross_converter.pl
$DATA/benchmark/virusseq/jumpdb/hg19Virus_refGene_RIS.txt $DATA/benchmark/simulated/750K/sim50x_SE.sorted.axt
$DATA/benchmark/simulated/750K/sim50x_crossReads.txt

RetroSeq - Commands & parameters used for WGS and TE

retroseq.pl -discover -bam $(NAME).bam -id 90 -len 20 -eref HK2_retroseq.list -align -output ${OUT}_retroseq

retroseq.pl -call -bam $(NAME).bam -input ${OUT}_retroseq -ref ${REF} -output ${OUT}_retroseq.vcf

MELT - Commands & parameters

java -jar ./MELTv2.0.2/MELT.jar BuildTransposonZIP HK113LTR.fa HK2_Hs_reference.bed HK2 10

java -jar ./MELTv2.0.2/MELT.jar Single -nocleanup -l
$DATA/benchmark/melt/${SAMPLE}/${SAMPLE}_h37d5_merged.sorted.bam -c 30 -h $DATA/benchmark/melt/hs37d5.fa -n
./MELTv2.0.2/add_bed_files/1KGP_Hg19/hg19.genes.bed -w $DATA/benchmark/melt/${SAMPLE}/ -t
$DATA/benchmark/melt/ERVK_MELT.zip

java -jar ./MELTv2.0.2/MELT.jar Deletion-Genotype -b $DATA/benchmark/melt/HK2_Hs_reference.bed -l
$DATA/benchmark/melt/${SAMPLE}/${SAMPLE}_h37d5_merged.sorted.bam -h $DATA/benchmark/melt/hs37d5.fa -w
$DATA/benchmark/melt/${SAMPLE}

####For target enrichment####
java -jar ./MELTv2.0.2/MELT.jar Single -r 300 -e 1200 -nocleanup -l
$DATA/benchmark/melt/${SAMPLE}/${SAMPLE}_h37d5_merged.sorted.bam -c 30 -h $DATA/benchmark/melt/hs37d5.fa -n
./MELTv2.0.2/add_bed_files/1KGP_Hg19/hg19.genes.bed -w $DATA/benchmark/melt/${SAMPLE}/ -t
$DATA/benchmark/melt/ERVK_MELT.zip

java -jar ./MELTv2.0.2/MELT.jar Deletion-Genotype -e 1200 -b $DATA/benchmark/melt/HK2_Hs_reference.bed -l
$DATA/benchmark/melt/${SAMPLE}/${SAMPLE}_h37d5_merged.sorted.bam -h $DATA/benchmark/melt/hs37d5.fa -w
$DATA/benchmark/melt/${SAMPLE}

Tlex2 - Over 5 days and did not complete requires more than 256 GB of RAM for high coverage genomes
tlex-open-v2.pl -s human -id 90 -limp 20 -lima 20 -processes 16 -noclean -paired yes -T ./ERVK_id.list -M
./ERVK_unique_strand.bed -G ./hs37d5.fa -R ./pedigree/NA12878/

VFS - Over 5 days and did not complete. At day 5, CAP3 phase began.
viral.fusion.pl --config vfs.conf --insertSIZE NA NA12891 ERR194160_1.fastq.gz ERR194160_2.fastq.gz

STEAK - parameters used for both WGS and TE
steak --input NA12878.PE.bam --paired --TE-reference HK113LTR.fa --alignment-quality 1.0 --match-quality 0.9 -
-aligned --transposon-length 20 --host-length 20

```


Citations

- Chen YX, Yao H, Thompson EJ, Tannir NM, Weinstein JN, Su XP. 2013. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics* **29**: 266-267.
- Fiston-Lavier AS, Barron MG, Petrov DA, Gonzalez J. 2015. T-lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. *Nucleic Acids Research* **43**.
- GitHub. 2015. Jitterbug. Available from: <https://github.com/elzbth/jitterbug> [Accessed 4 January 2017]
- Helman E, Lawrence MS, Stewart C, Sougnez C, Getz G, Meyerson M. 2014. Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Res* **24**: 1053-1063.
- Henaff E, Zapata L, Casacuberta JM, Ossowski S. 2015. Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution. *Bmc Genomics* **16**.
- Keane TM, Wong K, Adams DJ. 2013. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics* **29**: 389-390.
- Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, Lohr JG, Harris CC, Ding L, Wilson RK et al. 2012. Landscape of Somatic Retrotransposition in Human Cancers. *Science* **337**: 967-971.
- Lee WP, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. 2014. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS One* **9**: e90581.
- Li JW, Wan R, Yu CS, Co NN, Wong N, Chan TF. 2013. ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. *Bioinformatics* **29**: 649-651.
- MELT. 2016. University of Maryland. Available from: <http://melt.igs.umaryland.edu/index.php> [Accessed 4 January 2017]
- Reuters T. Citation Network. Web of Science. Available from: <http://apps.webofknowledge.com> [Accessed 4 January 2017]
- SourceForge. 2014. Mobster. Available from: <https://sourceforge.net/projects/mobster/> [Accessed 4 January 2017]
- Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MHY et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75-+.
- t-lex. 2011. Stanford University. Available from: <http://petrov.stanford.edu/cgi-bin/Tlex.html> [Accessed 4 January 2017].
- Thung DT, de Ligt J, Vissers LEM, Steehouwer M, Kroon M, de Vries P, Slagboom EP, Ye K, Veltman JA, Hehir-Kwa JY. 2014. Mobster: accurate detection of mobile element insertions in next generation sequencing data. *Genome Biology* **15**.
- Tubio JM, Li Y, Ju YS, Martincorena I, Cooke SL, Tojo M, Gundem G, Pipinikas CP, Zamora J, Raine K et al. 2014. Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**: 1251343.
- Wu J, Lee W-P, Ward A, Walker JA, Konkel MK, Batzer MA, Marth GT. 2014. Tangram: a comprehensive toolbox for mobile element insertion detection. *BMC Genomics* **15**: 795.

A.2 Appendix: Chapter 3

Tab. A.1.: The Cancer Genome Atlas 70 patients analysed in this study.

Sample	Disease	Centre	Assembly	Coverage	Self-reported Race
TCGA-DK-A1AG-10A-01D-A13W-08	Bladder Urothelial Carcinoma	Broad Institute	HG19 Broad Variant	39×	White
TCGA-HW-7486-10A-01D-2024-08	Brain Lower Grade Glioma	Broad Institute	HG19 Broad Variant	45×	White
TCGA-A2-A0D2-10A-01D-A128-09	Breast invasive carcinoma	Washington University School of Medicine	GRCh37-lite	39×	White
TCGA-A2-A04T-10A-01D-A128-09	Breast invasive carcinoma	Washington University School of Medicine	GRCh37-lite	39×	White
TCGA-EK-A2RM-10A-01D-A18J-09	Cervical squamous cell carcinoma	Washington University School of Medicine	GRCh37-lite	39×	White
TCGA-EK-A2RM-10A-01D-A18J-09	Endocervical adenocarcinoma	Washington University School of Medicine	GRCh37-lite	39×	White
TCGA-06-0145-10A-01D-0507-08	Glioblastoma multiforme	Broad Institute	HG19 Broad Variant	50×	White
TCGA-DQ-5625-10A-01D-A32X-10	Head and Neck squamous cell carcinoma	Baylor College of Medicine	GRCh37-lite	50×	White
TCGA-KM-8441-10A-01D-2311-10	Kidney Chromophobe	Baylor College of Medicine	GRCh37-lite	50×	White
TCGA-DV-5566-10A-01D-2104-10	Kidney renal clear cell carcinoma	Baylor College of Medicine	GRCh37-lite	45×	White
TCGA-MH-A562-10A-01D-A26P-10	Kidney renal papillary cell carcinoma	Baylor College of Medicine	GRCh37-lite	50×	African American
TCGA-FF-8062-10A-01D-2210-10	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	Baylor College of Medicine	GRCh37-lite	50×	Asian
TCGA-FF-8046-10A-01D-2210-10	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	Baylor College of Medicine	GRCh37-lite	50×	Asian
TCGA-13-0725-10B-01D-0446-08	Ovarian serous cystadenocarcinoma	Broad Institute	HG19 Broad Variant	50×	White
TCGA-13-0751-10A-01D-0446-08	Ovarian serous cystadenocarcinoma	Broad Institute	HG19 Broad Variant	50×	White
TCGA-HC-7233-10A-01D-2115-08	Prostate adenocarcinoma	Broad Institute	HG19 Broad Variant	39×	White
TCGA-G9-6336-10A-01D-1786-08	Prostate adenocarcinoma	Broad Institute	HG19 Broad Variant	39×	White
TCGA-EE-A2MI-10A-01D-A199-08	Skin Cutaneous Melanoma	Broad Institute	HG19 Broad Variant	39×	White
TCGA-CD-5802-10A-01D-1600-08	Stomach adenocarcinoma	Broad Institute	HG19 Broad Variant	45×	Asian
TCGA-E8-A416-10A-01D-A23K-08	Thyroid carcinoma	Broad Institute	HG19 Broad Variant	50×	Asian
TCGA-APA053-10A-01D-A013-09	Uterine Corpus Endometrioid Carcinoma	Washington University School of Medicine	GRCh37-lite	39×	White
TCGA-A7-A0CE-11A-21D-A12L-09	Breast invasive carcinoma	Washington University School of Medicine	GRCh37-lite	50×	White
TCGA-A1-A0SM-10A-02D-A099-09	Breast invasive carcinoma	Washington University School of Medicine	GRCh37-lite	50×	White
TCGA-CH-5750-10A-01D-1576-08	Prostate adenocarcinoma	Broad Institute	HG19 Broad Variant	50×	White
TCGA-CH-5789-10A-01D-1576-08	Prostate adenocarcinoma	Broad Institute	HG19 Broad Variant	50×	White
TCGA-D3-A3MO-10A-01D-A21A-08	Skin Cutaneous Melanoma	Broad Institute	HG19 Broad Variant	50×	White
TCGA-BH-A0E0-11A-13D-A128-09	Breast invasive carcinoma	Washington University School of Medicine	GRCh37-lite	50×	White
TCGA-CH-5763-11A-01D-1576-08	Prostate adenocarcinoma	Broad Institute	HG19 Broad Variant	50×	White
TCGA-EE-A2M5-10A-01D-A199-08	Skin Cutaneous Melanoma	Broad Institute	HG19 Broad Variant	50×	White
TCGA-CH-5788-10A-01D-1576-08	Prostate adenocarcinoma	Broad Institute	HG19 Broad Variant	50×	White
TCGA-A6-2681-10A-01D-2188-10	Colon adenocarcinoma	Baylor College of Medicine	GRCh37-lite	50×	White
TCGA-D3-A1Q5-10A-01D-A198-08	Skin Cutaneous Melanoma	Broad Institute	HG19 Broad Variant	50×	White
TCGA-06-0125-10A-01D-1490-08	Glioblastoma multiforme	Broad Institute	HG19 Broad Variant	50×	White
TCGA-A6-2683-11A-01D-1554-10	Colon adenocarcinoma	Baylor College of Medicine	GRCh37-lite	50×	White
TCGA-FF-8041-10A-01D-2210-10	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	Baylor College of Medicine	GRCh37-lite	50×	Asian
TCGA-FF-8042-10A-01D-2210-10	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	Baylor College of Medicine	GRCh37-lite	50×	Asian
TCGA-FF-8043-10A-01D-2210-10	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	Baylor College of Medicine	GRCh37-lite	50×	Asian
TCGA-FF-8061-10A-01D-2210-10	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	Baylor College of Medicine	GRCh37-lite	50×	Asian
TCGA-X2-A95T-10A-01D-A37F-09	Sarcoma	Washington University School of Medicine	GRCh37-lite	50×	White
TCGA-APA0LE-10A-01D-A10B-09	Uterine Corpus Endometrioid Carcinoma	Washington University School of Medicine	GRCh37-lite	50×	White
TCGA-AP-A0L9-10A-01D-A066-09	Uterine Corpus Endometrioid Carcinoma	Washington University School of Medicine	GRCh37-lite	50×	White
TCGA-AX-A1CI-11A-11D-A135-09	Uterine Corpus Endometrioid Carcinoma	Washington University School of Medicine	GRCh37-lite	50×	White
TCGA-L5-A4OR-11A-11D-A27G-09	Esophageal carcinoma	Washington University School of Medicine	GRCh37-lite	50×	White
TCGA-06-0185-10B-01D-0507-08	Glioblastoma multiforme	Broad Institute	HG19 Broad Variant	50×	White
TCGA-06-0648-10A-01D-0507-08	Glioblastoma multiforme	Broad Institute	HG19 Broad Variant	50×	White
TCGA-G3-A25T-10A-01D-A16V-10	Liver hepatocellular carcinoma	Baylor College of Medicine	GRCh37-lite	50×	White
TCGA-IG-A3YB-10A-01D-A247-09	Esophageal carcinoma	Washington University School of Medicine	GRCh37-lite	50×	White
TCGA-A6-3807-10A-01D-1459-10	Colon adenocarcinoma	Baylor College of Medicine	GRCh37-lite	50×	White
TCGA-LN-A49M-10A-01D-A27G-09	Esophageal carcinoma	Washington University School of Medicine	GRCh37-lite	50×	Asian
TCGA-06-0188-10B-01D-0373-08	Glioblastoma multiforme	Broad Institute	HG19 Broad Variant	50×	White
TCGA-BT-A20T-11A-11D-A14W-08	Bladder Urothelial Carcinoma	Broad Institute	HG19 Broad Variant	50×	White
TCGA-BT-A20P-11A-11D-A14W-08	Bladder Urothelial Carcinoma	Broad Institute	HG19 Broad Variant	50×	White
TCGA-BT-A20Q-11A-11D-A14W-08	Bladder Urothelial Carcinoma	Broad Institute	HG19 Broad Variant	50×	White
TCGA-BR-4255-11A-01D-A290-08	Stomach adenocarcinoma	Broad Institute	HG19 Broad Variant	50×	White
TCGA-CG-5730-11A-01D-1600-08	Stomach adenocarcinoma	Broad Institute	HG19 Broad Variant	50×	Not reported
TCGA-EL-A3MY-11A-12D-A21A-08	Thyroid carcinoma	Broad Institute	HG19 Broad Variant	50×	White
TCGA-10-0937-11A-01D-A324-10	Ovarian serous cystadenocarcinoma	Baylor College of Medicine	GRCh37-lite	50×	White
TCGA-69-7763-10A-01D-A461-10	Lung adenocarcinoma	Baylor College of Medicine	GRCh37-lite	50×	White
TCGA-CA-6718-10A-01D-1835-10	Colon adenocarcinoma	Baylor College of Medicine	GRCh37-lite	50×	Asian
TCGA-KL-8332-11A-01D-2310-10	Kidney Chromophobe	Baylor College of Medicine	GRCh37-lite	50×	White
TCGA-KM-8438-10A-01D-2311-10	Kidney Chromophobe	Baylor College of Medicine	GRCh37-lite	50×	White
TCGA-KO-8406-11A-01D-2311-10	Kidney Chromophobe	Baylor College of Medicine	GRCh37-lite	50×	White
TCGA-NJ-A4YQ-10A-01D-A46J-10	Lung adenocarcinoma	Baylor College of Medicine	GRCh37-lite	50×	White
TCGA-KM-8639-10A-01D-2397-10	Kidney Chromophobe	Baylor College of Medicine	GRCh37-lite	50×	White
TCGA-KM-8477-10A-01D-2311-10	Kidney Chromophobe	Baylor College of Medicine	GRCh37-lite	50×	White
TCGA-KL-8343-11A-01D-2310-10	Kidney Chromophobe	Baylor College of Medicine	GRCh37-lite	50×	White
TCGA-KL-8327-11A-01D-2310-10	Kidney Chromophobe	Baylor College of Medicine	GRCh37-lite	50×	White
TCGA-FV-A311-10A-01D-A22F-10	Liver hepatocellular carcinoma	Baylor College of Medicine	GRCh37-lite	50×	White
TCGA-CC-5261-10A-01D-A12Z-10	Liver hepatocellular carcinoma	Baylor College of Medicine	GRCh37-lite	50×	Asian
TCGA-CA-6717-10A-01D-1835-10	Colon adenocarcinoma	Baylor College of Medicine	GRCh37-lite	50×	Asian
TCGA-B3-3926-11A-01D-1253-10	Kidney renal papillary cell carcinoma	Baylor College of Medicine	GRCh37-lite	50×	White

Tab. A.2.: Polymorphic integrations found in 70 human genomes.

Chromosome	Coordinates (hg19)	Orientation to Reference	Flanking or Nearby Gene	Directionality to Gene & Intronic Location	Flanking Repetitive Elements
1	106015874–106015881	(-)			
1	111802591–111802598	(-)			MLT1A1
1	223578303–223578310	(+)			L1MDa
2	130719538–130722209	(-)			
3	148281477–148285396	(-)			
4	9603239–9603245	(-)			HERVS71-int
4	120263688–120264654	(-)			AluSx
5	8937853–8938820	(+)			L1PA13
5	64388439–64388446	(+)			L1M6
5	80442265–80442272	(-)	<i>RASGRF2</i>	Antisense; Ex 18/17	
6	16004859–16004863	(+)			
6*	32643459–32643464	(+)			L1PA10
6	32648035–32648041	(+)			L1PA10
6	161270898–161270905	(+)			
7	16237346–16238314	(+)	<i>ISPD</i>	Antisense; Ex 9/10	MER67B
7	158773385–158773390	(+)	<i>PTPRN2</i>	Antisense; Ex 3/4	
8	7355397–7364859	(-)			MIR
8	18651458–18652426	(+)	<i>PSD3</i>	Antisense; Ex 5/6	
8	37050885–37051853	(-)	<i>MIR1268A</i>	Antisense; Ex 1/2	
8	91696259–91697227	(+)			
11	60449889–60449889	(-)	<i>LINC00301</i>	Sense; Ex 6/7	L1M4
12	44313656–44313662	(+)	<i>TMEM117</i>	Sense; Ex 2/3	L1MB1
12	55727214–55728183	(+)	<i>OR6C3</i>	Sense	
12	118543788–118544748	(+)			AluY
12	124066476–124066483	(-)	<i>LOC101927415</i>	Antisense; Ex 1/2	AluSx1
13	90743182–90743189	(-)	<i>LINC00559</i>	Sense; Ex 3/4	
15	28430088	(+)	<i>HERC2</i>	Antisense; Ex 56/57	
15	63374593–63374600	(-)			AluJr
16	74833301–74834258	(-)			L1MB3
19	21841536–21841542	(-)			
19	29855781–29855787	(-)	<i>LOC284395</i>	Sense; Ex 9/10	
19	57996939–57996940	(-)	<i>ZNF419</i>	Antisense	

* Asterisk denotes novel integration found in this study.

A.3 Appendix: Chapter 4

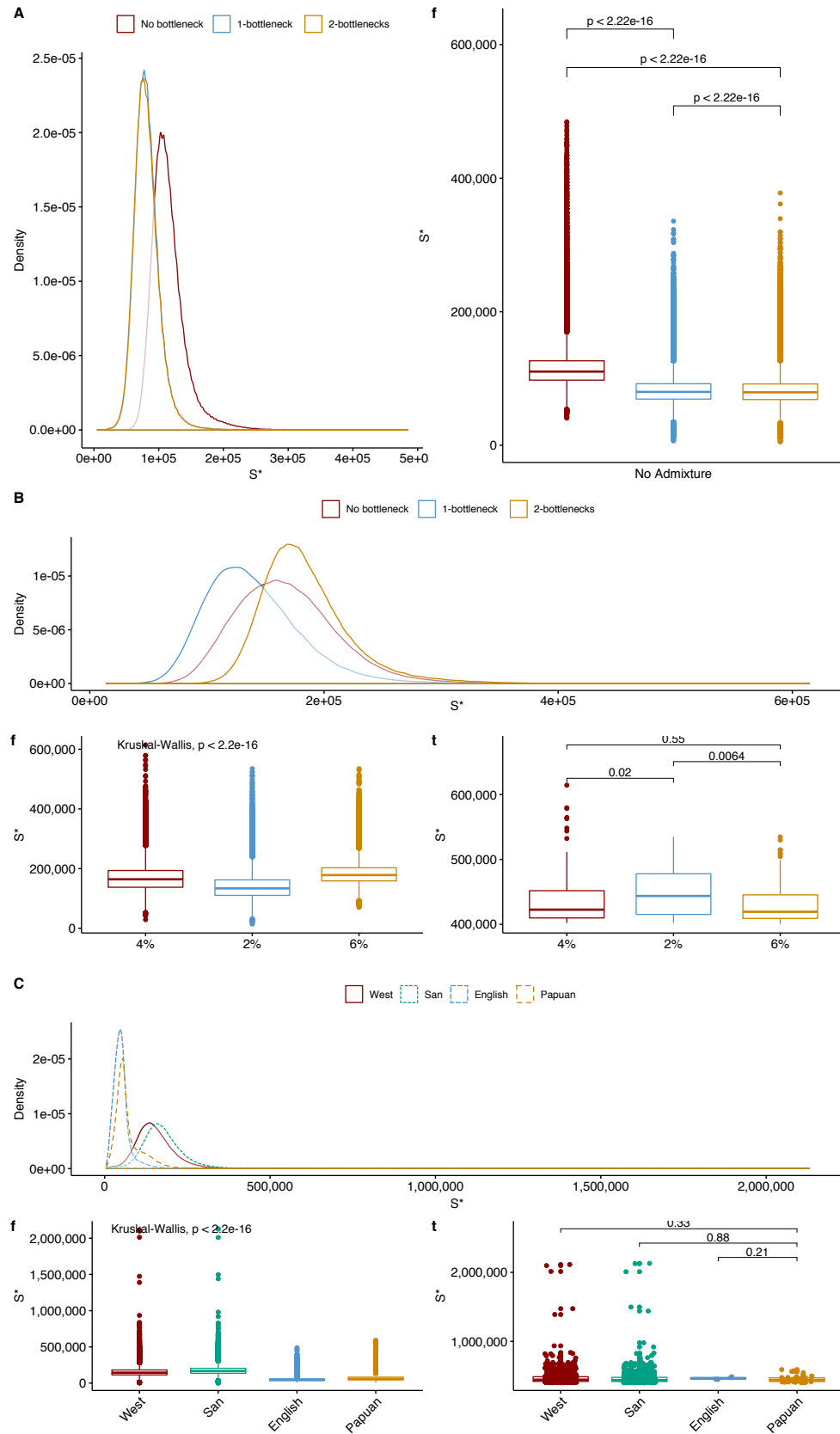


Fig. A.1.: Density plots and corresponding boxplots for S^* distributions of simulations (A) without admixture, (B) with admixture, and for (C) SGDP data. f: full distribution, t: top candidate distribution.

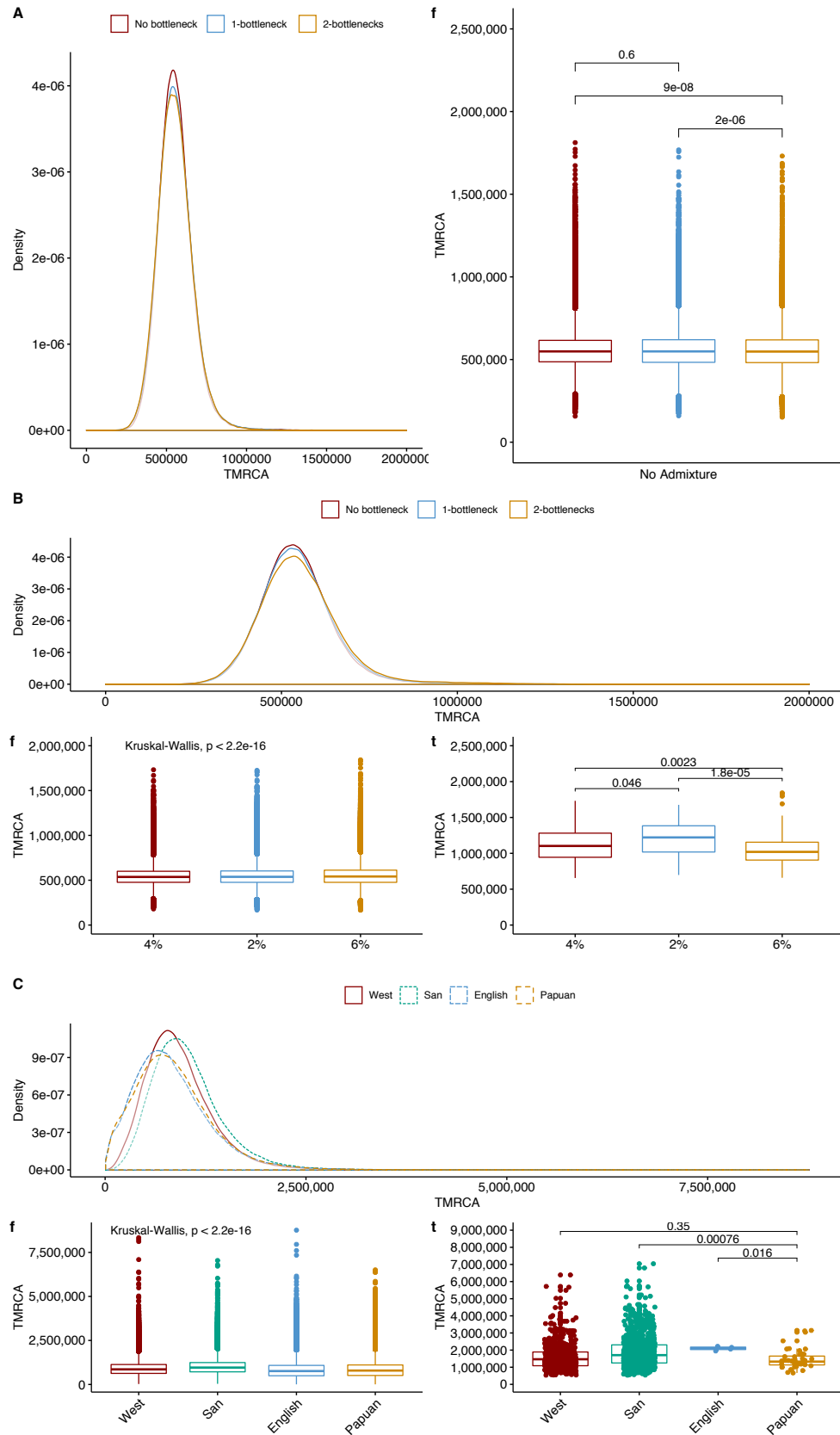


Fig. A.2.: Density plots and corresponding boxplots for TMRCA distributions of simulations (A) without admixture, (B) with admixture, and for (C) SGDP data. f: full S^* distribution, t: top candidate distribution.

A.4 Appendix: Chapter 5

Tab. A.3.: Subsets of African individuals from the Simons Genome Diversity Project per target grouping.

Group	Populations	Individuals
Pygmies	Mbuti	S_Mbuti-1
		S_Mbuti-2
		S_Mbuti-3
		B_Mbuti-4
	Biaka	S_Biaka-1
		S_Biaka-2
San	Khomani	S_Khomani_San-1
		S_Khomani_San-2
	Ju 'hoansi	S_Ju_hoan_North-1
		S_Ju_hoan_North-2
		S_Ju_hoan_North-3
		B_Ju_hoan_North-4
West	Yoruba	S_Yoruba-2
		B_Yoruba-3
	Gambian	S_Gambian-1
	Mandenka	S_Mandenka-2
	Mende	S_Mende-2
	Esan	S_Esan-1
East	Luhya	S_Luhya-1
	Masai	S_Masai-2
	Dinka	B_Dinka-3
	Luo	S_Luo-2
	BantuKenya	S_BantuKenya-2
	Somali	S_Somali-1
South	BantuHerero	S_BantuHerero-1
		S_BantuHerero-2
	BantuTswana	S_BantuTswana-1
		S_BantuTswana-2
North	Saharawi	S_Saharawi-1
		S_Saharawi-2
	Mozabite	S_Mozabite-1
		S_Mozabite-2

Tab. A.4.: Eurasian outgroup used when performing S^* and Sprime on African target populations.

Populations	Individuals
Basque	S_Basque-1
	S_Basque-2
Bergamo	S_Bergamo-1
	S_Bergamo-2
English	S_English-1
	S_English-2
Estonian	S_Estonian-2
	S_Estonian-1
Finnish	S_Finnish-2
	S_Finnish-1
	S_Finnish-3
French	S_French-1
	S_French-2
	B_French-3
Georgian	S_Georgian-2
	S_Georgian-1
Greek	S_Greek-1
	S_Greek-2
Japanese	S_Japanese-1
	S_Japanese-2
	S_Japanese-3
Korean	S_Korean-2
	S_Korean-1
Naxi	S_Naxi-2
	S_Naxi-3
	S_Naxi-1
Norwegian	S_Norwegian-1
Russian	S_Russian-1
	S_Russian-2
Sardinian	S_Sardinian-1
	S_Sardinian-2
	B_Sardinian-3
She	S_She-2
	S_She-1
Spanish	S_Spanish-1
	S_Spanish-2
Tuscan	S_Tuscan-2
	S_Tuscan-1
Xibo	S_Xibo-1
	S_Xibo-2
Yi	S_Yi-1
	S_Yi-2

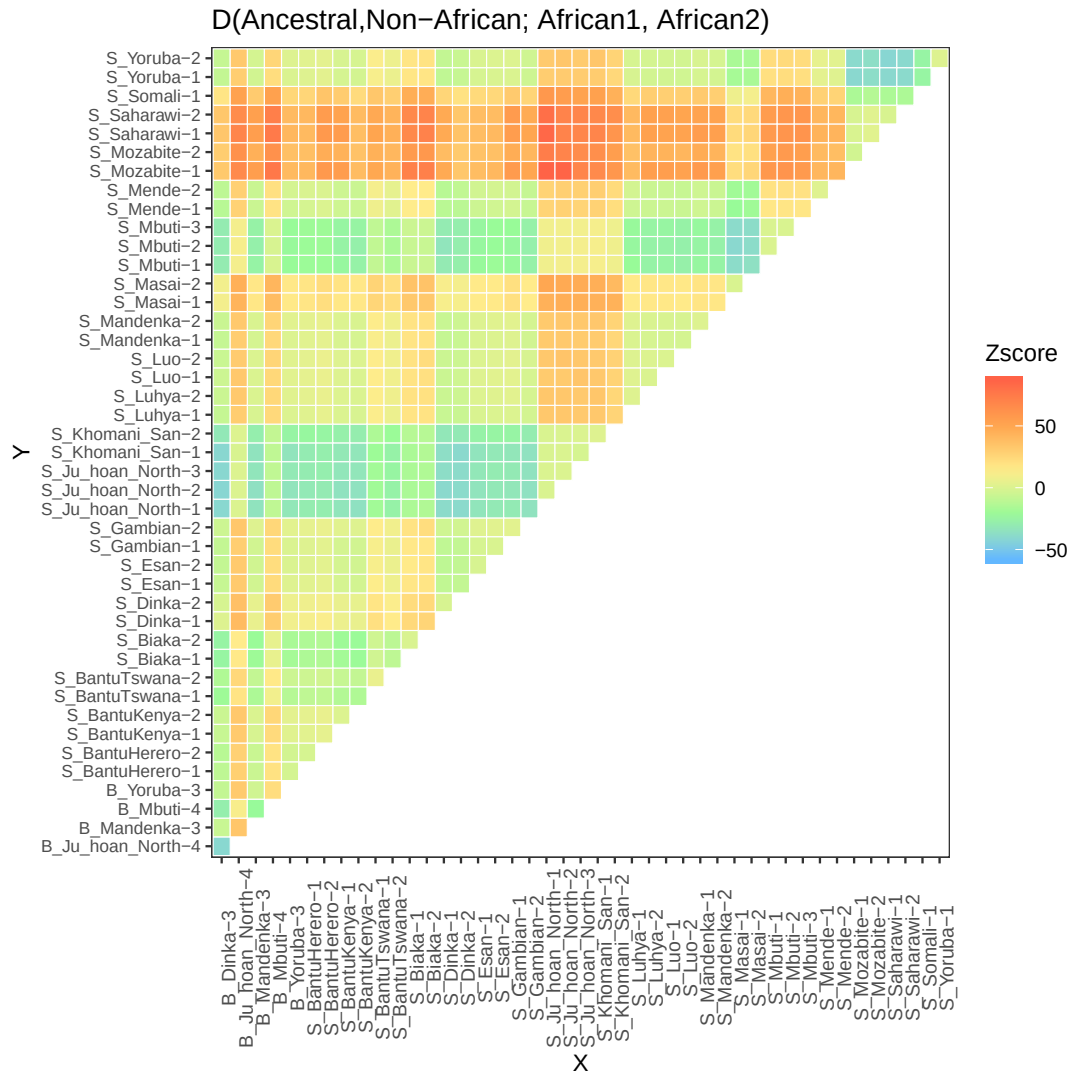


Fig. A.3.: *D*-statistics comparing two African individuals to a Non-African (English) with the reconstructed common ancestor (Ancestral) as the outgroup. The heatmap indicates whether individual on the Y shares an excess of alleles (red) or an absence of them (blue) with the English individual.

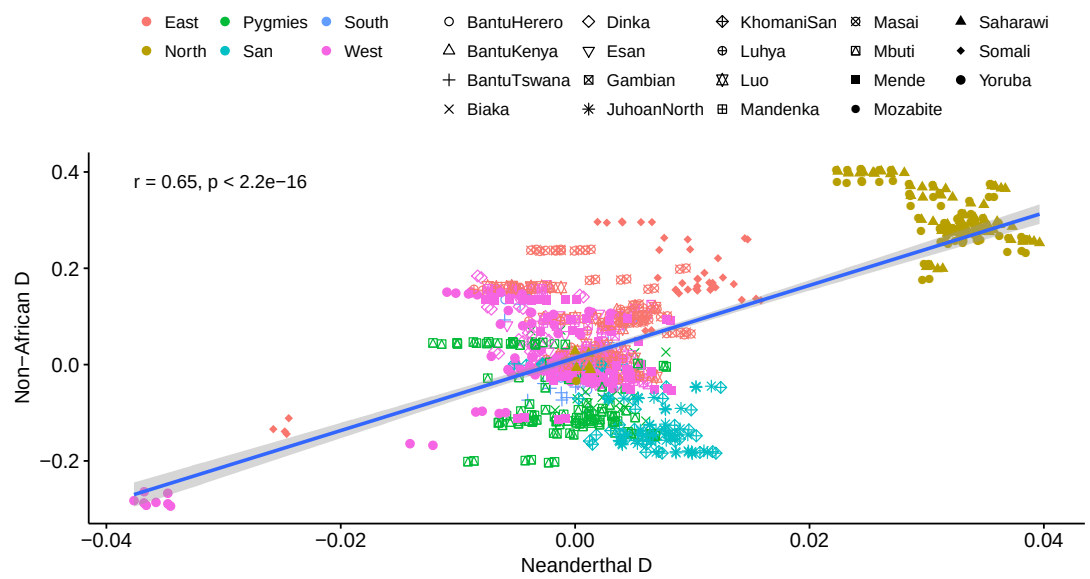


Fig. A.4.: Correlation of D -statistics comparing Non-African and Vindija shared ancestry across African populations. Each point is a pairing of $P3$ versus $P4$ where the represented population is $P4$ in the calculation of $D(\text{Ancestral}, P2, P3, P4)$

Bibliography

1. Agoni, L., Golden, A., Guha, C. & Lenz, J. Neandertal and Denisovan retroviruses. English. *Current Biology* **22**, R437–R438 (2012) (cit. on pp. 46, 48).
2. Allendorf, F. W. Genetic drift and the loss of alleles versus heterozygosity. *Zoo Biology* **5**, 181–190 (1986) (cit. on p. 77).
3. Antón, S. C. *Natural History of Homo erectus* 2003 (cit. on p. 11).
4. Armbruster, V., Sauter, M., Krautkraemer, E., *et al.* A Novel Gene from the Human Endogenous Retrovirus K Expressed in Transformed Cells. en. *Clinical Cancer Research* **8**, 1800–1807 (2002) (cit. on p. 5).
5. Arnaud, F., Varela, M., Spencer, T. E. & Palmarini, M. Endogenous retroviruses: Coevolution of endogenous betaretroviruses of sheep and their host. *Cellular and Molecular Life Sciences* (2008) (cit. on pp. 4, 45).
6. Ballouz, S., Dobin, A. & Gillis, J. A. Is it time to change the reference genome? *Genome Biology* **20**, 159 (2019) (cit. on p. 9).
7. Bannert, N. & Kurth, R. The Evolutionary Dynamics of Human Endogenous Retroviral Families. en. *Annual Review of Genomics and Human Genetics* **7**, 149–173 (2006) (cit. on pp. 5, 10, 44, 45, 48, 49).
8. Barbulescu, M., Turner, G., Seaman, M. I., *et al.* Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. *Current Biology* **9**, 861–S1 (1999) (cit. on pp. 6, 48).
9. Barbulescu, M., Turner, G., Su, M., *et al.* A HERV-K provirus in chimpanzees, bonobos and gorillas, but not humans. eng. *Current Biology* **11**, 779–783 (2001) (cit. on p. 6).
10. Barham, L., Llona, A & Stringer, C. Bone tools from Broken Hill (Kabwe) cave, Zambia, and their evolutionary significance. *Before Farming* (2014) (cit. on p. 11).
11. Behar, D. M., Villems, R., Soodyall, H., *et al.* The Dawn of Human Matrilineal Diversity. *The American Journal of Human Genetics* **82**, 1130–1140 (2008) (cit. on p. 69).
12. Belshaw, R., Dawson, A. L. A., Woolven-Allen, J, *et al.* Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): Implications for present-day activity. English. *Journal of Virology* **79**, 12507–12514 (2005) (cit. on pp. 48, 50, 133).
13. Belshaw, R., Pereira, V., Katzourakis, A., *et al.* Long-term reinfection of the human genome by endogenous retroviruses. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 4894–4899 (2004) (cit. on pp. 1, 3).

14. Benton, M. J. & Donoghue, P. C. J. Paleontological evidence to date the tree of life. *Molecular Biology and Evolution* **24**, 26–53 (2007) (cit. on p. 14).
15. Bentvelzen, P., Daams, J. H., Hageman, P & Calafat, J. Genetic transmission of viruses that incite mammary tumor in mice. *Proceedings of the National Academy of Sciences of the United States of America* **67**, 377–84 (1970) (cit. on pp. 4, 45).
16. Berg, I. L., Neumann, R., Sarbajna, S., *et al.* Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 12378–12383 (2011) (cit. on p. 96).
17. Berger, L. R. & Parkington, J. E. A new pleistocene hominid-bearing locality at Hoedjiespunt, South Africa. *American Journal of Physical Anthropology* (1995) (cit. on p. 91).
18. Berger, L. R., Hawks, J., de Ruiter, D. J., *et al.* Homo naledi, a new species of the genus Homo from the Dinaledi Chamber, South Africa. *eLife* (2015) (cit. on p. 91).
19. Berniell-Lee, G., Calafell, F., Bosch, E., *et al.* Genetic and demographic implications of the bantu expansion: Insights from human paternal lineages. *Molecular Biology and Evolution* (2009) (cit. on pp. 90, 114).
20. Besenbacher, S., Liu, S., Izarzugaza, J. M., *et al.* Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nature Communications* (2015) (cit. on p. 14).
21. Beyer, U., Moll-Rocek, J., Moll, U. M. & Dobbelstein, M. Endogenous retrovirus drives hitherto unknown proapoptotic p63 isoforms in the male germ line of humans and great apes. *Proceedings of the National Academy of Sciences of the United States of America* (2011) (cit. on p. 44).
22. Bhandari, S., Zhang, X., Cui, C., *et al.* Sherpas share genetic variations with Tibetans for high-altitude adaptation. *Molecular Genetics and Genomic Medicine* **5**, 76–84 (2017) (cit. on p. 14).
23. Bieda, K., Hoffmann, A. & Boller, K. Phenotypic heterogeneity of human endogenous retrovirus particles produced by teratocarcinoma cell lines. *Journal of General Virology* **82**, 591–596 (2001) (cit. on p. 5).
24. Blumenschine, R. J., Peters, C. R., Masao, F. T., *et al.* Late Pliocene homo and hominid land use from western Olduvai Gorge, Tanzania. *Science* (2003) (cit. on p. 11).
25. Boeke, J. D. & Stoye, J. P. eng. in *Retroviruses* (eds Coffin, J. M., Hughes, S. H. & Varmus, H. E.) (Cold Spring Harbor (NY), 1997) (cit. on pp. 4, 23, 49).
26. Boller, K., Konig, H., Sauter, M., *et al.* Evidence That Herv-K Is the Endogenous Retrovirus Sequence That Codes for the Human Teratocarcinoma-Derived Retrovirus Htdv. English. *Virology* **196**, 349–353 (1993) (cit. on pp. 45, 134).
27. Brady, T., Lee, Y. N., Ronen, K., *et al.* Integration target site selection by a resurrected human endogenous retrovirus. English. *Genes & Development* **23**, 633–642 (2009) (cit. on pp. 5, 138).
28. Bräuer, G. in *The origins of modern humans: A world survey of the fossil evidence* (eds Smith, F. H. & Spencer, F.) 327–410 (Liss, New York, 1984) (cit. on p. 13).
29. Bräuer, G. in *Continuity or Replacement: Controversies in Homo sapiens Evolution* (1992) (cit. on p. 13).

30. Bräuer, G. The origin of modern anatomy: By speciation or intraspecific evolution? *Evolutionary Anthropology* (2008) (cit. on p. 13).
31. Breton, G., Schlebusch, C. M., Lombard, M., *et al.* Lactase persistence alleles reveal partial east African ancestry of southern African Khoe pastoralists. *Current Biology* **24**, 852–858 (2014) (cit. on pp. 91, 120).
32. Brown, P., Sutikna, T., Morwood, M. J., *et al.* A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia. *Nature* **431**, 1055–1061 (2004) (cit. on p. 11).
33. Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S. & Akey, J. M. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell* **173** (2018) (cit. on pp. 1, 13, 17, 63, 64, 68, 71, 74, 92, 98, 100, 101, 103–105, 109, 129, 135).
34. Bryant, J. E., Holmes, E. C. & Barrett, A. D. T. Out of Africa: A Molecular Perspective on the Introduction of Yellow Fever Virus into the Americas. *PLoS Pathogens* **3**, e75 (2007) (cit. on p. 10).
35. Busby, G. B., Band, G., Si Le, Q., *et al.* Admixture into and within sub-Saharan Africa. *eLife* **5** (2016) (cit. on pp. 90, 134).
36. Büscher, K., Trefzer, U., Hofmann, M., *et al.* Expression of Human Endogenous Retrovirus K in Melanomas and Melanoma Cell Lines. *Cancer Research* **65**, 4172–4180 (2005) (cit. on pp. 5, 45).
37. Campana, M. G., Bower, M. A. & Crabtree, P. J. *Ancient DNA for the Archaeologist: The Future of African Research* tech. rep. 1 (2013), 21–37 (cit. on pp. 18, 88).
38. Cann, H., de Toma, C., Cazes, L., *et al.* A Human Genome Diversity Cell Line Panel. *Science* **296**, 261b–262 (2002) (cit. on p. 50).
39. Cann, R. L., Stoneking, M. & Wilson, A. C. *Mitochondrial DNA and human evolution* 1987 (cit. on pp. 13, 58).
40. Caulin, A. F. & Maley, C. C. *Peto's Paradox: Evolution's prescription for cancer prevention* 2011 (cit. on p. 45).
41. Chen, F., Welker, F., Shen, C.-C., *et al.* A late Middle Pleistocene Denisovan mandible from the Tibetan Plateau. *Nature* **569**, 409–412 (2019) (cit. on p. 11).
42. Chen, G. K., Marjoram, P. & Wall, J. D. Fast and flexible simulation of DNA sequence data. *Genome Research* (2009) (cit. on p. 69).
43. Chen, K., Wallis, J. W., McLellan, M. D., *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. English. *Nature Methods* **6**, 677–U76 (2009) (cit. on p. 23).
44. Chen, Y. X., Yao, H., Thompson, E. J., *et al.* VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. English. *Bioinformatics* **29**, 266–267 (2013) (cit. on pp. 24, 25).
45. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. English. *Science* **351**, 1083–1087 (2016) (cit. on p. 4).
46. Cohen, C. J., Lock, W. M. & Mager, D. L. Endogenous retroviral LTRs as promoters for human genes: A critical assessment. English. *Gene* **448**, 105–114 (2009) (cit. on p. 4).

47. Criscione, S. W., Zhang, Y., Thompson, W., Sedivy, J. M. & Neretti, N. Transcriptional landscape of repetitive elements in normal and cancer human cells. en. *BMC Genomics* **15**, 583 (2014) (cit. on pp. 7, 23, 41).
48. De Filippo, C., Bostoen, K., Stoneking, M. & Pakendorf, B. Bringing together linguistic and genetic evidence to test the Bantu expansion. *Proceedings of the Royal Society B: Biological Sciences* (2012) (cit. on pp. 89, 90).
49. De Manuel, M., Kuhlwilm, M., Frandsen, P., *et al.* Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* **354**, 477–481 (2016) (cit. on p. 86).
50. Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nature Methods*. arXiv: arXiv:1011.1669v3 (2012) (cit. on p. 97).
51. Denne, M., Sauter, M., Armbruster, V., *et al.* Physical and Functional Interactions of Human Endogenous Retrovirus Proteins Np9 and Rec with the Promyelocytic Leukemia Zinc Finger Protein. en. *Journal of Virology* **81**, 5607–5616 (2007) (cit. on p. 5).
52. Denner, J. & Tönjes, R. R. Infection Barriers to Successful Xenotransplantation Focusing on Porcine Endogenous Retroviruses. en. *Clinical Microbiology Reviews* **25**, 318–343 (2012) (cit. on p. 2).
53. Deschamps, M., Laval, G., Fagny, M., *et al.* Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. *American Journal of Human Genetics* **98**, 5–21 (2016) (cit. on p. 14).
54. Détroit, F., Mijares, A. S., Corny, J., *et al.* A new species of Homo from the Late Pleistocene of the Philippines. *Nature* **568**, 181–186 (2019) (cit. on p. 11).
55. Devière, T., Karavanić, I., Comeskey, D., *et al.* Direct dating of Neanderthal remains from the site of Vindija Cave and implications for the Middle to Upper Paleolithic transition. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 10606–10611 (2017) (cit. on p. 12).
56. Dewannieux, M & Heidmann, T. Endogenous retroviruses: acquisition, amplification and taming of genome invaders. English. *Current Opinion in Virology* **3**, 646–656 (2013) (cit. on p. 3).
57. Dewannieux, M, Harper, F, Richaud, A, *et al.* Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. English. *Genome Research* **16**, 1548–1556 (2006) (cit. on pp. 45, 134).
58. Djebali, S, Davis, C. A., Merkel, A, *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012) (cit. on p. 23).
59. Durvasula, A. & Sankararaman, S. Recovering signals of ghost archaic admixture in the genomes of present-day Africans. *bioRxiv*, 285734 (2018) (cit. on pp. 84, 87, 95, 96, 114, 115, 128, 131, 137).
60. Dusseldorp, G., Lombard, M. & Wurz, S. Pleistocene homo and the updated stone age sequence of South Africa. *South African Journal of Science* (2013) (cit. on p. 11).
61. Enard, D. & Petrov, D. A. Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. *Cell* (2018) (cit. on pp. 1, 137).

62. Eriksson, A & Manica, A. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 13956–13960 (2012) (cit. on pp. 13, 59, 86).
63. Evrony, G. D., Lee, E., Mehta, B. K., *et al.* Cell Lineage Analysis in Human Brain Using Endogenous Retroelements. English. *Neuron* **85**, 49–59 (2015) (cit. on p. 22).
64. Ewing, A. D. Transposable element detection from whole genome sequence data. *Mobile DNA* **6**, 24 (2015) (cit. on pp. 29, 33).
65. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics* **9** (ed Akey, J. M.) e1003905 (2013) (cit. on p. 71).
66. Ferring, R., Oms, O., Agusti, J., *et al.* Earliest human occupations at Dmanisi (Georgian Caucasus) dated to 1.85–1.78 Ma. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 10432–10436 (2011) (cit. on p. 11).
67. Fiston-Lavier, A.-S., Barrón, M. G., Petrov, D. A. & González, J. T-lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. *Nucleic Acids Research* **43**, e22–e22 (2015) (cit. on p. 24).
68. Fu, Q., Li, H., Moorjani, P., *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*. arXiv: NIHMS150003 (2014) (cit. on pp. 13, 15, 17, 129).
69. Fu, Q., Hajdinjak, M., Moldovan, O. T., *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015) (cit. on pp. 17, 90, 129).
70. Fu, Q., Posth, C., Hajdinjak, M., *et al.* The genetic history of Ice Age Europe. *Nature* **534**, 200–205 (2016) (cit. on p. 90).
71. Gallego Llorente, M., Jones, E. R., Eriksson, A., *et al.* Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. *Science* (2015) (cit. on pp. 21, 88, 91, 92, 105).
72. Galway-Witham, J. & Stringer, C. How did Homo sapiens evolve? *Science* **360**, 1296–1298 (2018) (cit. on pp. 14, 58, 73).
73. Gao, Z., Moorjani, P., Sasani, T. A., *et al.* Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 9491–9500 (2019) (cit. on p. 15).
74. Garcia-Montojo, M., Dominguez-Mozo, M., Arias-Leal, A., *et al.* The DNA Copy Number of Human Endogenous Retrovirus-W (MSRV-Type) Is Increased in Multiple Sclerosis Patients and Is Influenced by Gender and Disease Severity. *PLoS One* **8**, e53623 (2013) (cit. on p. 3).
75. Gardner, E. J., Lam, V. K., Harris, D. N., *et al.* The mobile element locator tool (MELT): Population-scale mobile element discovery and biology. *Genome Research* (2017) (cit. on pp. 24, 29, 138).
76. Gibbons, A. *A new view of the birth of homo sapiens* 2011 (cit. on p. 14).
77. Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. Recombination in the hemagglutinin gene of the 1918 "Spanish Flu". *Science* (2001) (cit. on p. 10).

78. Gnirke, A., Melnikov, A., Maguire, J., *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology* **27**, 182–189 (2009) (cit. on p. 26).
79. González-Santos, M. G., Montinaro, F., Oosthuizen, O., *et al.* Genome-Wide snp analysis of southern african populations provides new insights into the dispersal of bantu-Speaking groups. *Genome Biology and Evolution* (2015) (cit. on p. 89).
80. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17**, 333–351 (2016) (cit. on p. 22).
81. Gravel, S., Henn, B. M., Gutenkunst, R. N., *et al.* Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences of the United States of America* (2011) (cit. on p. 123).
82. Gravel, S. Population genetics models of local ancestry. *Genetics*. arXiv: 1202.4811 (2012) (cit. on pp. 66, 69, 89).
83. Graves, A. B. R. & J, B. Synthesis and processing of viral RNA. *en*, 205–262 (1997) (cit. on p. 2).
84. Green, R. E., Krause, J., Briggs, A. W., *et al.* A draft sequence of the neandertal genome. *Science*. arXiv: 15334406 (2010) (cit. on pp. 12, 13, 15, 58, 59, 67, 88, 96, 99).
85. Griffiths, R. C. & Marjoram, P. An ancestral recombination graph. *Progress in population genetics and human evolution (Minneapolis, MN, 1994)* (1997) (cit. on p. 93).
86. Grow, E. J., Flynn, R. A., Chavez, S. L., *et al.* Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. English. *Nature* **522**, 221–246 (2015) (cit. on pp. 4, 45, 133, 134, 137).
87. Grun, R., Brink, J. S., Spooner, N. A., *et al.* Direct dating of Florisbad hominid. *Nature* **382**, 500–501 (1996) (cit. on pp. 59, 91).
88. Günther, T. & Nettelblad, C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *bioRxiv*, 487983 (2018) (cit. on p. 138).
89. Hajdinjak, M., Fu, Q., Hübner, A., *et al.* Reconstructing the genetic history of late Neanderthals. *Nature* **555**, 652–656. arXiv: NIHMS150003 (2018) (cit. on pp. 12, 15).
90. Hammer, M. F., Woerner, A. E., Mendez, F. L., Watkins, J. C. & Wall, J. D. Genetic evidence for archaic admixture in Africa. *Proceedings of the National Academy of Sciences of the United States of America*. arXiv: arXiv:1408.1149 (2011) (cit. on pp. 17, 59, 60, 71, 76, 93, 94, 96, 114, 115, 128, 136).
91. Hanafusa, H, Hanafusa, T & Rubin, H. The defectiveness of Rous sarcoma virus. *Proceedings of the National Academy of Sciences of the United States of America* **49**, 572–580 (1963) (cit. on p. 45).
92. Harding, R. M. & McVean, G. *A structured ancestral population for the evolution of modern humans* 2004 (cit. on pp. 59, 89).
93. Harding, R. M., Fullerton, S. M., Griffiths, R. C., *et al.* Archaic African and Asian lineages in the genetic ancestry of modern humans. *American Journal of Human Genetics* **60**, 772–89 (1997) (cit. on p. 58).

94. Hartl, D. L. & Clark, A. G. *Principles of Population Genetics* 3rd, xiii, 542 p. (Sinauer Associates, Sunderland, MA, 1997) (cit. on pp. 50, 56, 57).
95. Harvati, K., Stringer, C., Grün, R., *et al.* The Later Stone Age Calvaria from Iwo Eleru, Nigeria: Morphology and Chronology. *PLoS One* **6** (ed Relethford, J. H.) e24024 (2011) (cit. on pp. 18, 89).
96. Havecker, E. R., Gao, X. & Voytas, D. F. The diversity of LTR retrotransposons. *Genome Biology* **5**, 225 (2004) (cit. on p. 44).
97. Hayward, A., Cornwallis, C. K. & Jern, P. Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. *Proceedings of the National Academy of Sciences of the United States of America* (2014) (cit. on p. 45).
98. Henn, B. M., Cavalli-Sforza, L. L. & Feldman, M. W. The great human expansion. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 17758–17764 (2012) (cit. on p. 46).
99. Henn, B. M., Gignoux, C. R., Jobin, M., *et al.* Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 5154–5162 (2011) (cit. on pp. 89, 91).
100. Henn, B. M., Steele, T. E. & Weaver, T. D. Clarifying distinct models of modern human origins in Africa. *Current Opinion in Genetics and Development* **53**, 148–156 (2018) (cit. on pp. 18, 59, 86, 89, 97).
101. Hercus, C. *Novocraft: short read alignment package*. 2009 (cit. on p. 27).
102. Higham, T., Ramsey, C. B., Karavanic, I., Smith, F. H. & Trinkaus, E. Revised direct radiocarbon dating of the Vindija G1 Upper Paleolithic Neandertals. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 553–557 (2006) (cit. on p. 12).
103. Hinch, A. G., Tandon, A., Patterson, N., *et al.* The landscape of recombination in African Americans. *Nature* **476**, 170–175 (2011) (cit. on pp. 96, 98, 116, 129).
104. Hirschl, S., Schanab, O., Seppel, H., *et al.* Sequence variability of retroviral particles derived from human melanoma cells: Melanoma-associated retrovirus. English. *Virus Research* **123**, 211–215 (2007) (cit. on pp. 45, 134).
105. Hobolth, A., Christensen, O. F., Mailund, T. & Schierup, M. H. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genetics* **3**, 294–304 (2007) (cit. on p. 15).
106. Hodgson, J. A. & Disotell, T. R. No evidence of a Neanderthal contribution to modern human diversity. *Genome Biology* **9**, 206 (2008) (cit. on p. 12).
107. Hohn, O., Hanke, K. & Bannert, N. HERV-K(HML-2), the Best Preserved Family of HERVs: Endogenization, Expression, and Implications in Health and Disease. *Frontiers in Oncology* **3**, 246 (2013) (cit. on pp. 6, 23, 45).
108. Hollfelder, N., Schlebusch, C. M., Günther, T., *et al.* Northeast African genomic variation shaped by the continuity of indigenous groups and Eurasian migrations. *PLoS Genetics* (2017) (cit. on p. 91).
109. Holliday, T. W., Gautney, J. R. & Friedl, L. Right for the wrong reasons: Reflections on modern human origins in the post-neanderthal genome era. *Current Anthropology* **55**, 696–724 (2014) (cit. on pp. 13, 14).

110. Holloway, J. R., Williams, Z. H., Freeman, M. M., Bulow, U. & Coffin, J. M. Gorillas have been infected with the HERV-K (HML-2) endogenous retrovirus much more recently than humans and chimpanzees. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 1337–1346 (2019) (cit. on pp. 56, 57).
111. Houldcroft, C. J. & Underdown, S. J. Neanderthal genomics suggests a pleistocene time frame for the first epidemiologic transition. *American Journal of Physical Anthropology* (2016) (cit. on p. 14).
112. Howells, W. W. Explaining modern man: Evolutionists Versus migrationists. *Journal of Human Evolution* (1976) (cit. on p. 13).
113. Hsieh, P., Woerner, A. E., Wall, J. D., *et al.* Corrigendum: Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies. *Genome Research* **26**, 717 1 (2016) (cit. on pp. 17, 59, 69–71, 74–76, 81, 86, 93, 94, 96, 97, 101, 114–116, 128, 134, 136, 137).
114. Hsieh, P., Veeramah, K. R., Lachance, J., *et al.* Whole-genome sequence analyses of Western Central African Pygmy hunter-gatherers reveal a complex demographic history and identify candidate genes under positive natural selection. *Genome Research* **26**, 279–290 (2016) (cit. on pp. 69, 70, 76, 86, 94, 101).
115. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012) (cit. on p. 26).
116. Hublin, J. J. The origin of Neandertals. *Proceedings of the National Academy of Sciences of the United States of America* (2009) (cit. on p. 12).
117. Hublin, J.-J. The last Neanderthal. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 10520–10522 (2017) (cit. on pp. 12, 18, 59).
118. Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002) (cit. on p. 50).
119. Hudson, R. R. The variance of coalescent time estimates from DNA sequences. *Journal of Molecular Evolution* (2007) (cit. on pp. 70, 99, 113).
120. Huerta-Sanchez, E., Jin, X., Asan, *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194–197 (2014) (cit. on p. 14).
121. Hughes, J. F. & Coffin, J. M. Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. *eng. Nature Genetics* **29**, 487–489 (2001) (cit. on pp. 6, 48).
122. Hughes, J. F. & Coffin, J. M. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: Implications for human and viral evolution. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 1668–1672 (2004) (cit. on pp. 49, 57).
123. Hughes, J. F. & Coffin, J. M. Human Endogenous Retroviral Elements as Indicators of Ectopic Recombination Events in the Primate Genome. *Genetics* **171**, 1183–1194 (2005) (cit. on p. 49).
124. Ingicco, T., Van Den Bergh, G. D., Jago-On, C., *et al.* Earliest known hominin activity in the Philippines by 709 thousand years ago. *Nature* **557**, 233–237 (2018) (cit. on p. 11).
125. Ingman, M., Kaessmann, H., Pääbo, S. & Gyllensten, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713 (2000) (cit. on p. 58).

126. Jaenisch, R. Germ line integration and Mendelian transmission of the exogenous Moloney leukemia virus. *Proceedings of the National Academy of Sciences of the United States of America* **73**, 1260 (1976) (cit. on p. 44).
127. Jha, A. R., Pillai, S. K., York, V. A., *et al.* Cross-Sectional Dating of Novel Haplotypes of HERV-K 113 and HERV-K 115 Indicate These Proviruses Originated in Africa before Homo sapiens. English. *Molecular Biology and Evolution* **26**, 2617–2626 (2009) (cit. on p. 47).
128. Jha, A. R., Nixon, D. F., Rosenberg, M. G., *et al.* Human endogenous retrovirus K106 (HERV-k106) was infectious after the emergence of anatomically modern humans. *PLoS ONE* **6**, e20234 (2011) (cit. on pp. 46, 50, 54, 56, 133).
129. Jobling, M. A. & Tyler-Smith, C. *The human Y chromosome: An evolutionary marker comes of age* 2003 (cit. on p. 58).
130. Johnson, W. E. & Coffin, J. M. Constructing primate phylogenies from ancient retrovirus sequences. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 10254–10260 (1999) (cit. on p. 47).
131. Johnson, W. E. Origins and evolutionary consequences of ancient endogenous retroviruses. *Nature Reviews Microbiology* (2019) (cit. on p. 45).
132. Kahyo, T., Yamada, H., Tao, H., Kurabe, N. & Sugimura, H. Insertionally polymorphic sites of human endogenous retrovirus-K (HML-2) with long target site duplications. *BMC Genomics* **18**, 487 (2017) (cit. on p. 31).
133. Karafet, T. M., Mendez, F. L., Meilerman, M. B., *et al.* New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Research* (2008) (cit. on p. 69).
134. Karamitros, T. & Magiorkinis, G. A novel method for the multiplexed target enrichment of MinION next generation sequencing libraries using PCR-generated baits. en. *Nucleic Acids Research*, gkv773 (2015) (cit. on p. 9).
135. Karamitros, T., Hurst, T., Marchi, E., *et al.* Human Endogenous Retrovirus-K HML-2 integration within RASGRF2 is associated with intravenous drug abuse and modulates transcription in a cell-line model. *Proceedings of the National Academy of Sciences of the United States of America*, 201811940 (2018) (cit. on p. 45).
136. Katzourakis, A., Magiorkinis, G., Lim, A. G., *et al.* Larger Mammalian Body Size Leads to Lower Retroviral Activity. en. *PLoS Pathogens* **10** (ed Maley, C.) e1004214 (2014) (cit. on p. 45).
137. Keane, T. M., Wong, K & Adams, D. J. RetroSeq: transposable element discovery from next-generation sequencing data. en. *Bioinformatics* **29**, 389–390 (2013) (cit. on pp. 9, 23–25, 29).
138. Kears, M., Moir, R., Wilson, A., *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012) (cit. on pp. 31, 42).
139. Kellis, M., Wold, B., Snyder, M. P., *et al.* Defining functional DNA elements in the human genome. en. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 6131–6138 (2014) (cit. on p. 7).

140. Kim, H. L., Ratan, A., Perry, G. H., *et al.* Khoisan hunter-gatherers have been the largest population throughout most of modern-human demographic history. *Nature Communications* **5**, 5692 (2014) (cit. on pp. 76, 90).
141. Kimbel, W. H., Johanson, D. C. & Rak, Y. Systematic assessment of a maxilla of Homo from Hadar, Ethiopia. *American Journal of Physical Anthropology* (1997) (cit. on p. 10).
142. Kingman, J. The coalescent. *Stochastic Processes and their Applications* **13**, 235–248 (1982) (cit. on p. 77).
143. Kong, A., Thorleifsson, G., Gudbjartsson, D. F., *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099–1103 (2010) (cit. on p. 96).
144. Krings, M., Stone, A., Schmitz, R. W., *et al.* Neandertal DNA sequences and the origin of modern humans. *Cell* (1997) (cit. on p. 12).
145. Kurth, R. & Bannert, N. *Retroviruses : molecular biology, genomics, and pathogenesis* 454 (Caister Academic Press, 2010) (cit. on pp. 1, 44).
146. Lachance, J, Vernot, B, Elbers, C. C., *et al.* Evolutionary History and Adaptation from High-Coverage Whole-Genome Sequences of Diverse African Hunter-Gatherers. English. *Cell* **150**, 457–469 (2012) (cit. on pp. 10, 17, 59, 70, 74–76, 80, 81, 86, 93, 94, 96, 97, 114, 128, 135–137).
147. Lahr, M. M. & Foley, R. A. Towards a theory of modern human origins: geography, demography, and diversity in recent human evolution. *American Journal of Physical Anthropology* **41**, 137–176 (1998) (cit. on pp. 59, 89).
148. Lamprecht, B., Walter, K., Kreher, S., *et al.* Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nature Medicine* **16**, 571–579 (2010) (cit. on pp. 6, 7).
149. Lander, E. S., Linton, L. M., Birren, B., *et al.* Initial sequencing and analysis of the human genome. English. *Nature* **409**, 860–921 (2001) (cit. on pp. 3, 22).
150. Lavachery, P. The holocene archaeological sequence of Shum Laka rock shelter (Grassfields, western Cameroon). *African Archaeological Review* (2001) (cit. on p. 18).
151. Lazaridis, I., Patterson, N., Mitnik, A., *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014) (cit. on p. 90).
152. Lee, E, Iskow, R, Yang, L, *et al.* Landscape of Somatic Retrotransposition in Human Cancers. en. *Science* **337**, 967–971 (2012) (cit. on pp. 5, 23, 30, 48).
153. Li, F., Nellåker, C., Yolken, R. H. & Karlsson, H. A systematic evaluation of expression of HERV-W elements; Influence of genomic context, viral structure and orientation. *BMC Genomics* (2011) (cit. on p. 3).
154. Li, H & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. English. *Bioinformatics* **26**, 589–595 (2010) (cit. on p. 123).
155. Li, H & Durbin, R. Inference of human population history from individual whole-genome sequences. English. *Nature* **475**, 493–U84 (2011) (cit. on p. 27).
156. Li, H., Handsaker, B., Wysoker, A., *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009) (cit. on pp. 28, 38).

157. Li, J. W., Wan, R., Yu, C. S., *et al.* ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. *Bioinformatics* **29**, 649–651 (2013) (cit. on pp. 24, 26, 98).
158. Li, S., Schlebusch, C. & Jakobsson, M. Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. *Proceedings of the Royal Society B: Biological Sciences* **281**, 20141448 (2014) (cit. on pp. 90, 114).
159. Li, W., Jin, Y., Prazak, L., Hammell, M & Dubnau, J. Transposable elements in TDP-43-mediated neurodegenerative disorders. *PLoS ONE* **7**, e44099 (2012) (cit. on p. 23).
160. Liang, M. & Nielsen, R. The lengths of admixture tracts. *Genetics* (2014) (cit. on p. 60).
161. Lin, A. T., Santander, C. G., Marchi, E., *et al.* Targeting endogenous retrovirus gene transcription in human cancers. *bioRxiv*, 449686 (2018) (cit. on p. 45).
162. Liu, L., Li, Y. H., Li, S. L., *et al.* Comparison of Next-Generation Sequencing Systems. English. *Journal of Biomedicine and Biotechnology* (2012) (cit. on p. 22).
163. Lopez, M., Kousathanas, A., Quach, H., *et al.* *The demographic history and mutational load of African hunter-gatherers and farmers* 2018 (cit. on pp. 56, 57, 76, 85, 90, 130, 138).
164. Lower, R., Lower, J & Frank, H. Human teratocarcinomas cultured in vitro produce unique retrovirus-like viruses. en. *Journal of General Virology* **65**, 887–898 (1984) (cit. on p. 45).
165. Lower, R., Boller, K., Hasenmaier, B, *et al.* Identification of Human Endogenous Retroviruses with Complex Messenger-Rna Expression and Particle Formation. English. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 4480–4484 (1993) (cit. on pp. 5, 45, 134).
166. Lunter, G & Goodson, M. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. English. *Genome Research* **21**, 936–939 (2011) (cit. on p. 27).
167. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 961–968 (2010) (cit. on p. 15).
168. Macfarlane, C. & Simmonds, P. Allelic variation of HERV-K(HML-2) endogenous retroviral elements in human populations. en. *Journal of Molecular Evolution* **59**, 642–656 (2004) (cit. on p. 45).
169. Maddison, D. R. African Origin of Human Mitochondrial DNA Reexamined. *Systematic Biology* **40**, 355–363 (1991) (cit. on p. 58).
170. Mager, D. L. & Freeman, J. D. HERV-H Endogenous Retroviruses: Presence in the New World Branch but Amplification in the Old World Primate Lineage. *Virology* **213**, 395–404 (1995) (cit. on p. 30).
171. Magin, C., Löwer, R. & Löwer, J. cORF and RcRE, the Rev/Rex and RRE/RxRE Homologues of the Human Endogenous Retrovirus Family HTDV/HERV-K. *Journal of Virology* **73**, 9496–9507 (1999) (cit. on p. 5).
172. Magiorkinis, G., Gifford, R. J., Katzourakis, A., De Ranter, J & Belshaw, R. Env-less endogenous retroviruses are genomic superspreaders. en. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 7385–7390 (2012) (cit. on p. 5).

173. Magiorkinis, G., Belshaw, R. & Katzourakis, A. 'There and back again': revisiting the pathophysiological roles of human endogenous retroviruses in the post-genomic era. en. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **368**, 20120504 (2013) (cit. on p. 56).
174. Magiorkinis, G., Blanco-Melo, D. & Belshaw, R. The decline of human endogenous retroviruses: extinction and survival. en. *Retrovirology* **12**, 8 (2015) (cit. on p. 45).
175. Maklakov, A. A. & Immler, S. The Expensive Germline and the Evolution of Ageing. *Current Biology* **26**, R577–R586 (2016) (cit. on p. 15).
176. Mallick, S., Li, H., Lipson, M., *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*. arXiv: NIHMS150003 (2016) (cit. on pp. 64, 68, 97, 100).
177. Mamanova, L., Coffey, A. J., Scott, C. E., *et al.* Target-enrichment strategies for next-generation sequencing. *Nature Methods* **7**, 111–118 (2010) (cit. on p. 26).
178. Marchi, E., Kanapin, A., Byott, M., Magiorkinis, G. & Belshaw, R. Neanderthal and Denisovan retroviruses in modern humans. *Current Biology* **23**, R994–R995 (2013) (cit. on pp. 6, 9, 10, 23, 30, 31, 38, 42, 46, 48–50, 52, 54, 56, 57, 133).
179. Marks, S. J., Montinaro, F., Levy, H., *et al.* Static and Moving Frontiers: The Genetic Landscape of Southern African Bantu-Speaking Populations. *Molecular Biology and Evolution* (2015) (cit. on p. 89).
180. Martin, S. H., Davey, J. W. & Jiggins, C. D. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Molecular Biology and Evolution* **32**, 244–257 (2015) (cit. on p. 13).
181. Matsumura, S. & Forster, P. Generation time and effective population size in Polar Eskimos. *Proceedings of the Royal Society B: Biological Sciences* **275**, 1501–1508 (2008) (cit. on pp. 14, 47, 71).
182. Mayr, E. in *Species Concepts and Phylogenetic Theory: A Debate* (2000) (cit. on p. 13).
183. Mazet, O., Rodríguez, W., Grusea, S., Boitard, S. & Chikhi, L. On the importance of being structured: Instantaneous coalescence rates and human evolution-lessons for ancestral population size inference? *Heredity* (2016) (cit. on pp. 89, 91).
184. McDougall, I., Brown, F. H. & Fleagle, J. G. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* (2005) (cit. on p. 18).
185. Medstrand, P., Landry, J. R. & Mager, D. L. Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. eng. *The Journal of Biological Chemistry* **276**, 1896–1903 (2001) (cit. on p. 5).
186. Medstrand, P., van de Lagemaat, L. N. & Mager, D. L. Retroelement distributions in the human genome: Variations associated with age and proximity to genes. English. *Genome Research* **12**, 1483–1495 (2002) (cit. on p. 52).
187. Meisler, M. H. & Ting, C. N. The remarkable evolutionary history of the human amylase genes. *Critical Reviews in Oral Biology and Medicine* (1993) (cit. on p. 4).
188. Meyer, M., Arsuaga, J. L., de Filippo, C., *et al.* Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature* **531**, 504–507 (2016) (cit. on p. 11).

189. Meyer, M., Fu, Q., Aximu-Petri, A., *et al.* A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature* **505**, 403–406 (2014) (cit. on p. 11).
190. Mi, S., Lee, X., Li, X.-p., *et al.* Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *en. Nature* **403**, 785–789 (2000) (cit. on pp. 4, 44).
191. Mills, S., Lunt, D. H. & Gómez, A. Global isolation by distance despite strong regional phylogeography in a small metazoan. *en. BMC Evolutionary Biology* **7**, 225 (2007) (cit. on p. 22).
192. Montesion, M., Bhardwaj, N., Williams, Z. H., Kuperwasser, C. & Coffin, J. M. Mechanisms of HERV-K (HML-2) Transcription during Human Mammary Epithelial Cell Transformation. *Journal of Virology* **92**, e01258–17 (2018) (cit. on pp. 4, 45).
193. Montinaro, F. & Capelli, C. The evolutionary history of Southern Africa. *Current Opinion in Genetics and Development* (2018) (cit. on pp. 92, 106).
194. Moorjani, P., Patterson, N., Hirschhorn, J. N., *et al.* The history of african gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genetics* (2011) (cit. on p. 80).
195. Muster, T., Waltenberger, A., Grassauer, A., *et al.* An endogenous retrovirus derived from human melanoma cells. English. *Cancer Research* **63**, 8735–8741 (2003) (cit. on p. 45).
196. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. Genetics: A fine-scale map of recombination rates and hotspots across the human genome. *Science* (2005) (cit. on p. 76).
197. Nakagawa, K. & Harrison, L. C. The Potential Roles of Endogenous Retroviruses in Autoimmunity. *Immunological Reviews* (2006) (cit. on p. 3).
198. Narasimhan, V. M., Rahbari, R., Scally, A., *et al.* Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nature Communications* **8**, 303 (2017) (cit. on p. 96).
199. Nedelec, Y., Sanz, J., Baharian, G., *et al.* Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. *Cell* **167** (2016) (cit. on p. 137).
200. Nielsen, R., Akey, J. M., Jakobsson, M., *et al.* Tracing the peopling of the world through genomics. *Nature*. arXiv: 15334406 (2017) (cit. on pp. 16, 88).
201. Oricchio, E., Sciamanna, I., Beraldi, R., *et al.* Distinct roles for LINE-1 and HERV-K retroelements in cell proliferation, differentiation and tumor progression. *Oncogene* **26**, 4226–4233 (2007) (cit. on p. 5).
202. Ostertag, E. M., Goodier, J. L., Zhang, Y. & Kazazian, H. H. SVA Elements Are Nonautonomous Retrotransposons that Cause Disease in Humans. *The American Journal of Human Genetics* **73**, 1444–1451 (2003) (cit. on p. 22).
203. Pagani, L., Kivisild, T., Tarekegn, A., *et al.* Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. *American Journal of Human Genetics* **91**, 83–96 (2012) (cit. on pp. 91, 92, 105, 120).
204. Patin, E., Siddle, K. J., Laval, G., *et al.* The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. *Nature Communications* (2014) (cit. on pp. 56, 90).

205. Patin, E., Lopez, M., Grollemund, R., *et al.* Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science* **356**, 543–546 (2017) (cit. on pp. 90, 133).
206. Patterson, N., Moorjani, P., Luo, Y., *et al.* Ancient admixture in human history. *Genetics*. arXiv: 15334406 (2012) (cit. on p. 59).
207. Patzke, S., Lindeskog, M., Munthe, E & Aasheim, H. C. Characterization of a novel human endogenous retrovirus, HERV-H/F, expressed in human leukemia cell lines. English. *Virology* **303**, 164–173 (2002) (cit. on p. 45).
208. Perry, G. H., Dominy, N. J., Claw, K. G., *et al.* Diet and the evolution of human amylase gene copy number variation. *Nature Genetics* (2007) (cit. on p. 4).
209. Petr, M., Pääbo, S., Kelso, J. & Vernot, B. Limits of long-term selection against Neandertal introgression. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 1639–1644 (2019) (cit. on p. 138).
210. Pickrell, J. K., Patterson, N., Loh, P. R., *et al.* Ancient west Eurasian ancestry in southern and eastern Africa. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 2632–2637 (2014) (cit. on pp. 80, 91, 120).
211. Pickrell, J. K., Patterson, N., Barbieri, C., *et al.* The genetic prehistory of southern Africa. *Nature Communications* **3**, 1143 (2012) (cit. on p. 90).
212. Pimenoff, V. N., De Oliveira, C. M., Bravo, I. G., Mendes De Oliveira, C. & Bravo, I. G. Transmission between archaic and modern human ancestors during the evolution of the oncogenic human papillomavirus 16. *Molecular Biology and Evolution* (2017) (cit. on pp. 1, 14).
213. Plagnol, V. & Wall, J. D. Possible Ancestral Structure in Human Populations. *PLoS Genetics* **2**, e105 (2006) (cit. on pp. 13, 17, 47, 60, 62, 69–71, 86, 93, 94, 114, 115).
214. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016) (cit. on p. 88).
215. Porat, N., Chazan, M., Grün, R., *et al.* New radiometric ages for the Fauresmith industry from Kathu Pan, southern Africa: Implications for the Earlier to Middle Stone Age transition. *Journal of Archaeological Science* (2010) (cit. on pp. 59, 91).
216. Povysil, G. & Hochreiter, S. IBD sharing between Africans, Neandertals, and Denisovans. *Genome Biology and Evolution* (2016) (cit. on pp. 89, 93, 123, 130, 131, 135).
217. Prufer, K., Racimo, F., Patterson, N., *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014) (cit. on pp. 12, 60, 64, 66, 84, 96, 130, 131).
218. Prufer, K., de Filippo, C., Grote, S., *et al.* A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655–658 (2017) (cit. on pp. 12, 130, 131).
219. Qin, P. & Stoneking, M. Denisovan ancestry in East Eurasian and Native American populations. *Molecular Biology and Evolution* (2015) (cit. on p. 64).
220. Quach, H., Rotival, M., Pothlichet, J., *et al.* Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. *Cell* **167**, 643–656 e17 (2016) (cit. on pp. 14, 137).

221. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *eng. Bioinformatics* **26**, 841–842 (2010) (cit. on p. 30).
222. Quintana-Murci, L. *Human Immunology through the Lens of Evolutionary Genetics* 2019 (cit. on p. 14).
223. Racimo, F., Sankararaman, S., Nielsen, R. & Huerta-Sanchez, E. Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics* **16**, 359–371 (2015) (cit. on pp. 10, 14, 17).
224. Raghavan, M., DeGiorgio, M., Albrechtsen, A., *et al.* The genetic prehistory of the New World Arctic. *Science* **345**, 1255832 (2014) (cit. on p. 90).
225. Ragsdale, A. P. & Gravel, S. Models of archaic admixture and recent history from two-locus statistics. *bioRxiv*, 489401 (2018) (cit. on pp. 95, 96, 128).
226. Rands, C. M., Meader, S., Ponting, C. P. & Lunter, G. 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *PLoS Genetics* **10** (ed Schierup, M. H.) e1004525 (2014) (cit. on pp. 7, 36).
227. Reich, D., Green, R. E., Kircher, M., *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010) (cit. on pp. 11–13, 46, 58, 59, 67, 89, 96, 100, 109, 120, 130).
228. Reich, D., Patterson, N., Kircher, M., *et al.* Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *American Journal of Human Genetics* (2011) (cit. on pp. 12, 13, 58).
229. Relethford, J. H. & Jorde, L. B. Genetic evidence for larger African population size during recent human evolution. *American Journal of Physical Anthropology* **108**, 251–260 (1999) (cit. on p. 123).
230. Reuter, J. A., Spacek, D. V. & Snyder, M. P. High-Throughput Sequencing Technologies. *Molecular Cell* **58**, 586–597 (2015) (cit. on pp. 7, 9).
231. Ribet, D., Harper, F., Dupressoir, A., *et al.* An infectious progenitor for the murine IAP retrotransposon: Emergence of an intracellular genetic parasite from an ancient retrovirus. *Genome Research* **18**, 597–609 (2008) (cit. on p. 44).
232. Rifkin, R. F., Potgieter, M., Ramond, J. B. & Cowan, D. A. Ancient oncogenesis, infection and human evolution. *Evolutionary Applications* **10**, 949–964 (2017) (cit. on p. 45).
233. Rightmire, G. P. Out of Africa: modern human origins special feature: middle and later Pleistocene hominins in Africa and Southwest Asia. *Proceedings of the National Academy of Sciences of the United States of America* (2009) (cit. on pp. 11, 18, 59).
234. Rishishwar, L., Marino-Ramirez, L. & Jordan, I. K. Benchmarking computational tools for polymorphic transposable element detection. *Brief Bioinform* **18**, 908–918 (2017) (cit. on p. 41).
235. Rishishwar, L., Wang, L., Clayton, E. A., *et al.* Population and clinical genetics of human transposable elements in the (post) genomic era. *Mobile Genetic Elements* **7**, 1–20 (2017) (cit. on p. 52).
236. Roach, J. C., Glusman, G., Smit, A. F., *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010) (cit. on p. 15).

237. Roebroeks, W. & Villa, P. On the earliest evidence for habitual use of fire in Europe. *Proceedings of the National Academy of Sciences of the United States of America* (2011) (cit. on p. 11).
238. Romanish, M. T., Cohen, C. J. & Mager, D. L. Potential mechanisms of endogenous retroviral-mediated genomic instability in human cancer. en. *Seminars in Cancer Biology* **20**, 246–253 (2010) (cit. on p. 5).
239. Rowe, H. M. & Trono, D. Dynamic control of endogenous retroviruses during development. eng. *Virology* **411**, 273–287 (2011) (cit. on p. 4).
240. Ruprecht, K., Mayer, J., Sauter, M., Roemer, K. & Mueller-Lantzsch, N. *Endogenous retroviruses: Endogenous retroviruses and cancer* 2008 (cit. on p. 3).
241. Sankararaman, S., Mallick, S., Dannemann, M., *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354–357 (2014) (cit. on pp. 17, 62, 67, 84, 96, 116, 129).
242. Sankararaman, S., Patterson, N., Li, H., Pääbo, S. & Reich, D. The Date of Interbreeding between Neandertals and Modern Humans. *PLoS Genetics* **8** (ed Akey, J. M.) e1002947 (2012) (cit. on pp. 58, 67).
243. Sankararaman, S., Mallick, S., Patterson, N. & Reich, D. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Current Biology* **26**, 1241–1247 (2016) (cit. on pp. 1, 10, 12, 58, 64, 68, 84, 96, 116, 129).
244. Santander, C. G., Gambrón, P., Marchi, E., *et al.* STEAK: A specific tool for transposable elements and retrovirus detection in high-throughput sequencing data. *Virus Evolution* **3**, 1–12 (2017) (cit. on pp. 10, 20, 138).
245. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics* **13**, 745–753 (2012) (cit. on pp. 14, 69–71, 99).
246. Scerri, E. The Stone Age Archaeology of West Africa. *Oxford Research Encyclopedia of African History* **1** (2017) (cit. on p. 18).
247. Scerri, E. M., Drake, N. A., Jennings, R. & Groucutt, H. S. Earliest evidence for the structure of *Homo sapiens* populations in Africa. *Quaternary Science Reviews* (2014) (cit. on p. 89).
248. Scerri, E. M., Thomas, M. G., Manica, A., *et al.* Did Our Species Evolve in Subdivided Populations across Africa, and Why Does It Matter? *Trends in Ecology and Evolution* **33**, 582–594 (2018) (cit. on pp. 59, 71, 89).
249. Schiffels, S & Durbin, R. Inferring human population size and separation history from multiple genome sequences. English. *Nature Genetics* **46**, 919–925 (2014) (cit. on p. 123).
250. Schlebusch, C. M., Skoglund, P., Sjödin, P., *et al.* Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science* **338**, 374–379 (2012) (cit. on pp. 49, 56, 88–92, 105, 120, 130, 133).
251. Schlebusch, C. M. & Jakobsson, M. Tales of Human Migration, Admixture, and Selection in Africa. *Annual Review of Genomics and Human Genetics* **19**, 405–428 (2018) (cit. on pp. 18, 19, 21, 69, 88).
252. Schlebusch, C. M., Malmström, H., Günther, T., *et al.* Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science* **358**, 652–655. arXiv: science.aao6266(2017) [10.1126] (2017) (cit. on pp. 89, 90).

253. Schmitt, K., Reichrath, J., Roesch, A., Meese, E. & Mayer, J. Transcriptional profiling of human endogenous retrovirus group HERV-K(HML-2) loci in melanoma. *Genome Biology and Evolution* (2013) (cit. on p. 5).
254. Schmitt, K., Heyne, K., Roemer, K., Meese, E. & Mayer, J. HERV-K(HML-2) rec and np9 transcripts not restricted to disease but present in many normal human tissues. *Mobile DNA* **6**, 4 (2015) (cit. on p. 45).
255. Schmitz, R. W., Serre, D., Bonani, G., *et al.* The Neandertal type site revisited: Interdisciplinary investigations of skeletal remains from the Neander Valley, Germany. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 13342–13347 (2002) (cit. on p. 12).
256. Schumer, M., Xu, C., Powell, D. L., *et al.* Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science* (2018) (cit. on p. 135).
257. Seguin-Orlando, A., Korneliussen, T. S., Sikora, M., *et al.* Genomic structure in Europeans dating back at least 36,200 years. *Science* (2014) (cit. on pp. 10, 64, 66, 67, 84).
258. Sherman, R. M., Forman, J., Antonescu, V., *et al.* *Assembly of a pan-genome from deep sequencing of 910 humans of African descent* 2019 (cit. on pp. 9, 88, 138).
259. Shukla, R., Upton, K. R., Munoz-Lopez, M., *et al.* Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell* **153**, 101–111 (2013) (cit. on p. 23).
260. Simmonds, P., Aiewsakun, P. & Katzourakis, A. Prisoners of war — host adaptation and its constraints on virus evolution. *Nature Reviews Microbiology*, 1 (2018) (cit. on p. 1).
261. Simola, D. F. & Kim, J. Sniper: improved SNP discovery by multiply mapping deep sequenced reads. *Genome Biology* **12**, R55 (2011) (cit. on p. 23).
262. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nature Genetics* **46**, 220–224. arXiv: 1305.2061 (2014) (cit. on p. 138).
263. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019) (cit. on p. 88).
264. Skoglund, P. & Jakobsson, M. Archaic human ancestry in East Asia. *Proceedings of the National Academy of Sciences of the United States of America* (2011) (cit. on p. 64).
265. Skoglund, P., Thompson, J. C., Prendergast, M. E., *et al.* Reconstructing Prehistoric African Population Structure. *Cell* **171**, 59–71 e21 (2017) (cit. on pp. 19, 21, 59, 88–93, 105, 120, 129, 131, 136).
266. Skoglund, P., Mallick, S., Bortolini, M. C., *et al.* Genetic evidence for two founding populations of the Americas. *Nature* **525**, 104 (2015) (cit. on p. 99).
267. Skoglund, P., Posth, C., Sirak, K., *et al.* Genomic insights into the peopling of the Southwest Pacific. *Nature* (2016) (cit. on p. 12).
268. Skov, L. *Personal communication: archaic admixture in Africa using HMM-based detection software.* 2018 (cit. on p. 119).
269. Skov, L., Hui, R., Shchur, V., *et al.* Detecting archaic introgression using an unadmixed outgroup. *PLoS Genetics* **14** (ed Racimo, F.) e1007641 (2018) (cit. on pp. 68, 84, 92, 119).
270. Slatkin, M. & Racimo, F. Ancient DNA and human history. *Proceedings of the National Academy of Sciences of the United States of America* (2016) (cit. on p. 88).

271. Slon, V., Mafessoni, F., Vernot, B., *et al.* The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature* (2018) (cit. on p. 12).
272. Smith, F. H. in *The origins of modern humans: A world survey of the fossil evidence* (eds Smith, F. H. & Spencer, F.) 137–209 (Liss, New York, 1984) (cit. on p. 13).
273. Smith, F. H. in *Origins of anatomically modern humans*. (eds Nitecki, M. & Nitecki, D.) 227–249 (Plenum Press, New York, 1994) (cit. on p. 13).
274. Smith, F. H. in *A Companion to Biological Anthropology* (2010) (cit. on p. 13).
275. Solyom, S. & Kazazian, H. H. Mobile elements in the human genome: implications for disease. *Genome Medicine* **4**, 12 (2012) (cit. on pp. 22, 23).
276. Speidel, L., Forest, M., Shi, S. & Myers, S. A method for genome-wide genealogy estimation for thousands of samples. *bioRxiv*, 550558 (2019) (cit. on pp. 131, 135, 136).
277. Storey, J. D., Tibshirani, R. & Green, P. P. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100** (2003) (cit. on p. 87).
278. Stoye, J. P. & Coffin, J. M. eng. in *Molecular biology of tumor viruses* (ed Weiss, R.) 2nd editio. 10C,2, 357–404 (Cold Spring Harbor Laboratory, Cold Spring Harbor, 1985) (cit. on pp. 2, 5).
279. Stoye, J. P. Studies of endogenous retroviruses reveal a continuing evolutionary saga. *Nature Reviews Microbiology* (2012) (cit. on pp. 3, 45).
280. Stringer, C. B & Andrews, P. Genetic and fossil evidence for the origin of modern humans. *Science* (1988) (cit. on pp. 13, 50, 54, 58).
281. Stringer, C. Modern human origins-distinguishing the models. *African Archaeological Review* **18**, 67–75 (2001) (cit. on pp. 13, 14).
282. Stringer, C. Modern human origins: Progress and prospects. *Philosophical Transactions of the Royal Society B: Biological Sciences* (2002) (cit. on p. 89).
283. Stringer, C. The status of *Homo heidelbergensis* (Schoetensack 1908). *Evolutionary Anthropology* **21**, 101–107 (2012) (cit. on p. 11).
284. Stringer, C. The origin and evolution of *homo sapiens*. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371** (2016) (cit. on p. 59).
285. Subramanian, R. P., Wildschutte, J. H., Russo, C. & Coffin, J. M. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology* **8** (2011) (cit. on pp. 6, 8, 30, 45, 47–52, 133).
286. Sudmant, P. H., Rausch, T., Gardner, E. J., *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015) (cit. on pp. 46, 52, 138).
287. Sutikna, T., Tocheri, M. W., Morwood, M. J., *et al.* Revised stratigraphy and chronology for *Homo floresiensis* at Liang Bua in Indonesia. *Nature* **532**, 366–369 (2016) (cit. on p. 11).
288. Tajima, F. Infinite-allele model and infinite-site model in population genetics. *Journal of Genetics* **75**, 27–31 (1996) (cit. on p. 15).
289. Tarailo-Graovac, M & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **Chapter 4**, Unit 4 10 (2009) (cit. on p. 30).

290. Tarlinton, R. E., Meers, J. & Young, P. R. Retroviral invasion of the koala genome. *Nature* (2006) (cit. on pp. 4, 44, 45).
291. Tattersall, I. & Schwartz, J. H. Evolution of the Genus Homo. *Annual Review of Earth and Planetary Sciences* (2009) (cit. on pp. 56, 133).
292. Taylor, S. A. & Larson, E. L. Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nature Ecology and Evolution* **3**, 170–177 (2019) (cit. on p. 13).
293. Teleki, G., Hunt, E. E. & Pfifferling, J. H. Demographic observations (1963-1973) on the chimpanzees of Gombe National Park, Tanzania. *Journal of Human Evolution* (1976) (cit. on pp. 47, 71).
294. Templeton, A. R. Human origins and analysis of mitochondrial DNA sequences. *Science* **255**, 737–739 (1992) (cit. on p. 58).
295. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74. arXiv: 15334406 (2015) (cit. on pp. 50, 100).
296. The Chimpanzee Sequencing and Analysis Consortium, Mikkelsen, T. S., Hillier, L. W., *et al.* Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* (2005) (cit. on p. 49).
297. The HapMap Consortium, Altshuler, D, Gibbs, R. a., *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* (2010) (cit. on pp. 76, 96, 98, 116, 129).
298. The UK10K Consortium, Walter, K., Min, J. L., *et al.* The UK10K project identifies rare variants in health and disease. *Nature*. arXiv: arXiv:1411.3180v1 (2015) (cit. on p. 64).
299. Theunert, C & Slatkin, M. Distinguishing recent admixture from ancestral population structure. *Genome Biology and Evolution* (2017) (cit. on pp. 59, 80, 85, 86).
300. Thomson, R, Pritchard, J. K., Shen, P, Oefner, P. J. & Feldman, M. W. Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 7360–7365 (2000) (cit. on p. 70).
301. Tischler, G. & Leonard, S. biobambam: tools for read pair collation based algorithms on BAM files. *Source Code for Biology and Medicine* **9**, 13 (2014) (cit. on p. 28).
302. Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., *et al.* The genetic structure and history of Africans and African Americans. *Science* **324**, 1035–1044 (2009) (cit. on pp. 19, 46, 52, 89, 90, 92, 106, 114).
303. Tomlins, S. A., Laxman, B, Dhanasekaran, S. M., *et al.* Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. English. *Nature* **448**, 595–U9 (2007) (cit. on p. 7).
304. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* **13**, 36–46 (2011) (cit. on pp. 7–9, 22).
305. Triska, P., Soares, P., Patin, E., *et al.* Extensive admixture and selective pressure across the sahel belt. *Genome Biology and Evolution* (2015) (cit. on p. 92).

306. Tryon, C. A., Crevecoeur, I., Faith, J. T., *et al.* Late Pleistocene age and archaeological context for the hominin calvaria from GvJm-22 (Lukenya Hill, Kenya). *Proceedings of the National Academy of Sciences of the United States of America* **112**, 2682–2687 (2015) (cit. on pp. 18, 89).
307. Tung, J., Charpentier, M. J., Garfield, D. A., Altmann, J & Alberts, S. C. Genetic evidence reveals temporal change in hybridization patterns in a wild baboon population. *Molecular Ecology* **17**, 1998–2011 (2008) (cit. on p. 13).
308. Turner, G., Barbulescu, M., Su, M., *et al.* Insertional polymorphisms of full-length endogenous retroviruses in humans. *eng. Current Biology* **11**, 1531–1535 (2001) (cit. on pp. 6, 48).
309. Van Regenmortel, M. H. V., Mayo, M. A., Fauquet, C. M. & Maniloff, J. Virus nomenclature: consensus versus chaos. English. *Archives of Virology* **145**, 2227–2232 (2000) (cit. on p. 5).
310. Van de Lagemaat, L. N., Medstrand, P. & Mager, D. L. Multiple effects govern endogenous retrovirus survival patterns in human gene introns. English. *Genome Biology* **7** (2006) (cit. on p. 52).
311. Van Den Bergh, G. D., Kaifu, Y., Kurniawan, I., *et al.* Homo floresiensis-like fossils from the early Middle Pleistocene of Flores. *Nature* **534**, 245–248 (2016) (cit. on p. 11).
312. Verdu, P., Austerlitz, F., Estoup, A., *et al.* Origins and Genetic Diversity of Pygmy Hunter-Gatherers from Western Central Africa. *Current Biology* (2009) (cit. on p. 90).
313. Vernot, B. & Akey, J. M. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343**, 1017–1021. arXiv: 20 (2014) (cit. on pp. 58, 60).
314. Vernot, B. & Akey, J. M. Complex History of Admixture between Modern Humans and Neandertals. *en. The American Journal of Human Genetics* **96**, 448–453 (2015) (cit. on pp. 10, 12, 13, 17, 70, 74, 86, 92, 109, 128, 134).
315. Vernot, B., Tucci, S., Kelso, J., *et al.* Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* (2016) (cit. on pp. 10, 58, 62, 63, 68, 70, 74, 81, 100, 134).
316. Villanea, F. A. & Schraiber, J. G. Multiple episodes of interbreeding between Neanderthal and modern humans. *Nature Ecology and Evolution* **3**, 39–44 (2019) (cit. on p. 17).
317. Villesen, P., Aagaard, L., Wiuf, C. & Pedersen, F. S. Identification of endogenous retroviral reading frames in the human genome. *en. Retrovirology* **1**, 32 (2004) (cit. on p. 3).
318. Voisset, C., Weiss, R. A. & Griffiths, D. J. Human RNA "Rumor" Viruses: the Search for Novel Human Retroviruses in Chronic Disease. English. *Microbiology and Molecular Biology Reviews* **72**, 157–196 (2008) (cit. on pp. 4, 5).
319. Wall, J. D., Lohmueller, K. E. & Plagnol, V. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Molecular Biology and Evolution* (2009) (cit. on pp. 17, 60, 93, 96, 114, 115, 128).
320. Wall, J. D., Yang, M. A., Jay, F., *et al.* Higher levels of Neanderthal ancestry in east Asians than in Europeans. *Genetics* (2013) (cit. on p. 62).
321. Wang, L., Rishishwar, L., Marino-Ramirez, L & Jordan, I. K. Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. *Nucleic Acids Research* **45**, 2318–2328 (2017) (cit. on pp. 88, 137).

322. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**, 256–76 (1975) (cit. on p. 47).
323. Wegmann, D., Kessner, D. E., Veeramah, K. R., *et al.* Recombination rates in admixed individuals identified by ancestry-based inference. *Nature Genetics* **43**, 847–853 (2011) (cit. on pp. 96, 116).
324. Weinstein, J. N., Collisson, E. A., Mills, G. B., *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* **45**, 1113–1120. arXiv: arXiv:1403.6652v2 (2013) (cit. on pp. 22, 36).
325. Weiss, R. A. & Stoye, J. P. Virology. Our viral inheritance. *Science* **340**, 820–821 (2013) (cit. on p. 3).
326. White, T. D., Asfaw, B., DeGusta, D., *et al.* Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. *Nature* (2003) (cit. on p. 18).
327. Wildschutte, J. H., Williams, Z. H., Montesion, M., *et al.* Discovery of unfixed endogenous retrovirus insertions in diverse human populations. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E2326–E2334 (2016) (cit. on pp. 6, 23, 30, 46–50, 52, 54–56, 132, 133).
328. Williamson, S. H., Hubisz, M. J., Clark, A. G., *et al.* Localizing Recent Adaptive Evolution in the Human Genome. *PLoS Genetics* **3**, e90 (2007) (cit. on p. 87).
329. Wolf, D. & Goff, S. P. Host Restriction Factors Blocking Retroviral Replication. *Annual Review of Genetics* (2008) (cit. on p. 57).
330. Wolpoff, M. H., Wu, X. Z. & Thorne, A. G. in *The Origins of Modern Humans. A World Survey of the Fossil Evidence* (1984) (cit. on pp. 13, 58).
331. Wood, B. Did early *Homo* migrate "out of" or "in to" Africa? *Proceedings of the National Academy of Sciences of the United States of America* **108**, 10375–10376 (2011) (cit. on p. 11).
332. Wu, J., Lee, W.-P., Ward, A., *et al.* Tangram: a comprehensive toolbox for mobile element insertion detection. English. *BMC Genomics* **15**, 795 (2014) (cit. on pp. 9, 23–25).
333. Xu, D., Pavlidis, P., Thamadilok, S., *et al.* Recent evolution of the salivary mucin MUC7. *Scientific Reports* **6**, 31791 (2016) (cit. on p. 137).
334. Xu, D., Pavlidis, P., Taskent, R. O., *et al.* Archaic Hominin Introgression in Africa Contributes to Functional Salivary MUC7 Genetic Variation. *Molecular Biology and Evolution* **34**, 2704–2715 (2017) (cit. on p. 137).
335. Yang, M. A., Malaspinas, A. S., Durand, E. Y. & Slatkin, M. Ancient structure in Africa unlikely to explain Neanderthal and non-African genetic similarity. *Molecular Biology and Evolution* **29**, 2987–2995 (2012) (cit. on p. 90).
336. Yang, M. A., Harris, K. & Slatkin, M. The projection of a test genome onto a reference population and applications to humans and archaic hominins. *Genetics* **198**, 1655–1670 (2014) (cit. on pp. 71, 86).
337. Zhao, M., Lee, W.-P., Garrison, E. P. & Marth, G. T. SSW Library: An SIMD Smith-Waterman C/C++ Library for Use in Genomic Applications. *PLoS ONE* **8** (ed Mariño-Ramírez, L.) e82138. arXiv: 1208.6350 (2013) (cit. on pp. 28, 33).

338. Zilhão, J., Anesin, D., Aubry, T., *et al.* Precise dating of the Middle-to-Upper Paleolithic transition in Murcia (Spain) supports late Neandertal persistence in Iberia. *Heliyon* **3**, e00435 (2017) (cit. on p. 12).

List of Figures

1.1	<i>Retrovirus life cycle</i> . Schematic from Denner & Tönjes (2012), Clinical Microbiology Reviews. The green single strands represent the RNA genomes that are converted by RT to the red DNA duplex, known as the provirus, that then integrates into the host cellular DNA. Through the normal course of host transcription, viral mRNA is produced along with host RNA. The synthesised viral proteins come together with the genomic RNA to form viral particles, which sequentially bud and cleave off.	2
1.2	Mechanism of retrovirus endogenisation (Dewannieux & Heidmann, 2013). In the course of a normal infection, retroviruses can reach the germline of the host where they integrate their genome as a provirus. The latter can then be passed down through Mendelian inheritance to the offspring, and also amplify with the genome, leading to the creation of a new family of ERVs.	3
1.3	<i>Structure of HK2 provirus and RNA transcripts</i> . Schematic from Hohn <i>et al.</i> (2013). The first transcript is the canonical <i>env</i> ORF. The second is HK2 type 2 provirus structure with the <i>rec</i> ORF. The third structure is of HK2 type 1 provirus with the 292 bp deletion at the 5' end of <i>env</i> ORF results in <i>np9</i> splice product.	6
1.4	<i>Ambiguities in aligning repeats</i> . Diagram from Treangen & Salzberg (2011). (A) Above are shown three different tandem repeats with 2 copies each. Beneath these repeats are reads that are sequenced from these regions. For each of the repeats, a single read has been zoomed. In the first case, the red read has a low confidence when mapping within the tandem repeat because it aligns well to both X_1 and X_2 . In the second case, there is a higher confidence in mapping the green reads due to a single nucleotide difference— Y_1 matches better than Y_2 . In the final example, the blue read aligns perfectly to Z_1 whereas against Z_2 , there are three mismatches. (B) The 13 bp shown aligns to 2 locations: "a" and "b". Location "a" has a mismatch and location "b" has a deletion. Depending on parameters used by an aligner, mismatches are often less costly than a deletion. The alignment program may map the read in location "a" but perhaps the source of DNA may have truly belonged to location "b".	8

1.5	Proposed models of human evolution. The left side depicts the two main human evolution models that were challenged and now replaced by an intermediate model (right side) which considers interbreeding with archaic hominins outside of Africa. From Gibbons (2011) as reported in Science (Gibbons, 2011). Reprinted with permission from AAAS.	14
1.6	<i>Simplified model of archaic hominin and modern human history from Nielsen et al. (2017) featured in Nature.</i> The relationships between populations today and the estimated times where they split off from the most recent common ancestor are shown here. The solid lines represent well-established admixture events whereas dashed lines stand for tentative admixture between modern and archaic humans. The model also displays the potential a small fraction of ancestry observed in the Oceanic population, derived from an early out-of-Africa event. In the small inset we see an insight into the history of ancestral Europeans which studies have found to be an admixture between ancient north Eurasian (ANE), early European farmer (EEF), and west European hunter-gatherer (WHG).	16
1.7	<i>Admixture versus ancestral population structure.</i> Expected lengths of archaic-like tracts for an admixture scenario (left panel) and an ancestral population structure scenario (right panel). Because there is less time for recombination to break down the migrated tracts (red) in the admixed population, the expected tract length in this case will be longer than in the case of ancestral population structure (Racimo <i>et al.</i> , 2015).	17
1.8	<i>Stratification and migration across the African continent.</i> (a) Map of major prefarming population stratification across the African continent, showing non-Africans (brown), East Africans (dark blue), West Africans (green), central African hunter-gatherers (light blue), and Khoe-San (red). (b) Migration routes related to the expansion of herders and crop farmers during Holocene times. The green arrows represent the Bantu Expansion, the brown arrows represent the Eurasian back-migration, the brown/blue arrow represents the southward migration of mixed East African/Eurasian pastoralists, and the blue/green arrow represents bidirectional migration across the Sahel belt. Abbreviation: BP, years before present. (Schlebusch & Jakobsson, 2018).	19

2.1	<i>Workflow of STEAK.</i> Processing data: All reads are locally aligned using the Smith Waterman algorithm and allowing mismatches when mapping reads against a TE reference (5' and 3'-ends and respective reverse complements). Reads that match with the TE are trimmed of the matching portion. Information on the trimmed reads and their mates, such as the original mapping positions, MAPQ, and sequence qualities, are kept in STEAK outputs. Detection Module: Trimmed reads can be remapped to the human reference either as single-end (trimmed read detection) or paired-end reads (guided detection).	29
2.2	<i>Parallelisation of STEAK processing.</i> A 50× coverage simulation of chromosome 1 was processed using our MPI-based software. The speedup as a function of the number of cores shows that the program scales well because few concurrency issues affect it. t_1 is the time of the problem computing at a single processor whereas t_n is the average time of the problem computing at one processor where N processors are run.	31
2.3	<i>Processing in STEAK.</i> (A) The software runs as several processes that read different parts of the input SAM or FASTQ file. (B) In the case of a BAM/CRAM file, because the data is read from a single point, a unique process is executed but it spawns several threads, one that accumulates the data in a buffer and the others that process them. TEs are specified in a FASTA file. The trimmed sequence is written in the first FASTQ file. The original sequence of the other read (or the trimmed sequence if there is a match as well) is placed in the second FASTQ file.	32
2.4	<i>Distributions for cluster read depth of known HK2 integrations.</i> Top: Cluster read depth for all known integrations. Middle: Cluster read depth for known integrations within repetitive elements. Bottom: Cluster read depth for known integrations that are not within repetitive elements.	34
2.5	<i>Distribution of twenty simulated HIV integrations within human reference (hg19) in chromosome 1.</i> Respective genomic coordinates can be found in Table 2.2.	37
2.6	<i>Comparative performance of HK2 detection in whole genome sequencing.</i> (A) Bar graph displays known non-reference integrations detected in WGS projects by each respective system. (B) Pyramid plot depicts number of polymorphic integrations detected per genome—number of present non-reference and number of absent reference integrations. Abbreviations correspond to samples in Table 2.3.	39
2.7	<i>Detection performance in target enrichment data.</i> The left facet depicts presence of non-reference integrations detected by each respective system. The right facet depicts the marking of presence of reference integrations. MELT results detect 5 reference presences and 137 absences.	40

3.1	Trajectory of <i>HERV-K (HML-2)</i> integration rate (μ) across primate evolution. Upper panel depicts phylogeny of primate species where HK2 colonised at different time points in the last 41 million years. The dotted grey lines connect phylogeny to times of HK2 amplification. The appearance of alternative splicing product <i>np9</i> is pointed to in red. Dates and integration rates older than 5 MYA were calculated from Subramanian <i>et al.</i> (2011).	51
3.2	Estimated allele frequencies of polymorphic HK2 integrations distributed world-wide. Each map represents the frequency of an integration type: (A) Solo-LTRs, (B) 2-LTRs, (C) Presence of the full-length provirus. The difference between (B) and (C) is that (B) reflects the presence of the integration through LTR detection while (C) reflects the presence of the full-provirus using unique k-mers to reconstruct the viral genes.	53
3.3	Inferred frequencies of unfixed HK2 integrations in germline across individuals from 1000 KGP and patients with different cancers. Figure (A) shows the frequency of solo-LTRs and (B) the frequencies of 2-LTR integrations. Frequency estimates for Americans of African ancestry in Southwest of the US (ASW), Utah residents with Northern and Western European ancestries (CEU), and Southern Han Chinese ancestries (CHS) (Wildschutte <i>et al.</i> , 2016) are displayed at the bottom-end for comparison purposes (i.e. TCGA samples were from self-reported "White", "Asian", and "African-American" individuals). Only integrations searched for both the Wildschutte <i>et al.</i> (2016) study and this study are displayed. TCGA abbreviations: Bladder Urothelial Carcinoma (BLCA), Breast invasive carcinoma (BRCA), Colon adenocarcinoma (COAD), Esophageal carcinoma (ESCA), Lymphoid Neoplasm Diffuse Large B-cell Lymphoma (DLBC), Glioblastoma multiforme (GBM), Kidney Chromophobe (KICH), Kidney renal papillary cell carcinoma (KIRP), Liver hepatocellular carcinoma (LIHC), Lung adenocarcinoma (LUAD), Ovarian serous cystadenocarcinoma (OV), Prostate adenocarcinoma (PRAD), Skin Cutaneous Melanoma (SKCM), Stomach adenocarcinoma (STAD), Thyroid carcinoma (THCA), Uterine Corpus Endometrial Carcinoma (UCEC).	55

- 4.1 *Schematic for S^* scoring.* Here we see phased chromosomes and a test population that is represented in red and the outgroup in grey. S^* attempts to search for the most optimal sums of scores for an overall subset of SNPs at a given locus. It rewards fully linked pairs of sites, in other words where two successive SNP positions do not differ, and the increase in that reward is proportional to the distance between the positions. Consequently, S^* values increase with increasing LD within a window. The SNP positions that provide the optimal score are taken note of and can then be used to provide the delimited region that gives that optimal score. The lower bottom of the figure depicts an example of how S^* goes about calculating the most optimal sums of scores. Seven positions within a window of 50,000 bp and only SNPs not present in the outgroup are considered. Sites within the haplotype are labelled as a grey star (i.e. ancestral) or a white star (i.e. derived), those highlighted in red (target) are calculated to give the most optimal solution and that no other set of SNPs gives a higher score. The first 10,000 bp displays higher linkage than the following 10,000 bp where there is a change in genotype leading to a penalty of -10,000. We can see that in the calculation that follows the positions from about 30,000 bp onwards are rewarded for being linked and outweigh any penalty from before. The scoring function works in rewarding linked sites with a score of 5,000 plus the distance between sites, penalties are given -10,000 for up to 5 mismatches in the above example. Regions with more than 5 mismatches or no linked sites are $-\infty$, or essentially no score. The total S^* score for this example is 30,000. 61

- 4.2 *Overview of Hidden Markov model and conditional random field framework.* Here is a depiction of how probabilistic models, such as HMM and CRF, can estimate the ancestry (x_i) of a SNP in a genome sequence (i). The possible hidden states for x_i are either introgressed (archaic) or not introgressed (not archaic). In the example, y_i is a matrix composed of individuals from a target population, an archaic population, and two outgroup populations that act as references for modern human variation. These are considered at three SNP positions (below model). Sites which are consistent with introgression (x_1) are determined by the derived allele of the target population also present in the archaic population but not in the reference. In the case of x_2 , the derived allele is only in one of the reference populations suggesting that the derived allele is modern human variation and therefore inconsistent with introgression. In x_3 , both target and reference populations share the derived allele but absent in the archaic population which is uninformative. Hidden states (ancestry) are connected to the observed data (y_i) through emission probabilities (red) for HMM or emission functions (vertical red and diagonal yellow) for CRF. These emission functions in a CRF model can be used to evaluate whether a site is consistent or inconsistent with introgression, like in HMM. Additionally they can also score whether a haplotype overall is closer to the archaic sequence than to the reference haplotypes. Horizontal connections between x_i and $x_i + 1$ denote transition probabilities (HMM) or transition functions (CRF) which model linkage between ancestral states along the genome. These parameters depend on the recombination, admixture proportion, and time of admixture. 65
- 4.3 *Demographic models used in simulations.* The schematic of (A) the base population model showing the main points where parameters varied in each of the simulations ran. Referred to parameters can be found in Table 4.1. Parameter (1) specifies the date of divergence between humans and the archaic, (2) controls the split time between Africans [A] and Eurasians [E] and (3) dictates the date where admixture between an archaic and African humans occurred. Illustration based off of schematic from Galway-Witham & Stringer (2018) (Galway-Witham & Stringer, 2018). Schematics representing the (B-E) demographic events explored in the simulations. Table 4.1, under the "Demography" column, indicates which simulation was run with a simulated African population under (B) exponential growth and a Eurasian population with a bottleneck, (C) a hunter-gatherer [HG] African population and a farmer African population under logistic growth, (D) a model like (B) but where only Eurasians may receive archaic introgression and finally (E) a model where a simulated Papuan [P] population passes through two bottlenecks and may undergo archaic admixture. . . . 73

4.4	<i>Admixture versus no admixture.</i> Density plots displaying skew in the full distribution of S^* scores when comparing a population simulated under the null hypothesis (red) versus a population simulated under the null hypothesis with 4% admixture (blue).	74
4.5	<i>Admixture proportions.</i> Pairwise comparison of simulations without admixture and with admixture in different percentages. (A) Full distribution of S^* scores and, (B) the top distribution of the S^* scores ($> 400,000$) when admixture is present.	75
4.6	<i>Constant versus variable recombination.</i> Models with admixture and without admixture with either constant or variable recombination rate. (Above) Full distribution of S^* scores when admixture is present and, (below) the top distribution of the S^* scores (i.e. $> 100,000$ instead of 400,000 to visualise differences in no admixture scenarios).	76
4.7	<i>Exponential versus logistic growth.</i> Pairwise comparisons of full distribution of S^* scores for simulations with exponential growth (blue) or logistic growth (yellow) in scenarios without (left) and with (right) admixture. P -values indicate how S^* distributions are statistically significantly different from each other when comparing growth models regardless of whether archaic introgression is present or absent.	78
4.8	<i>Admixture under logistic growth.</i> Pairwise comparison of the full distribution of S^* scores for the null hypothesis model which contains no admixture (Sim Null) versus a simulation of a scenario with 40% of archaic admixture but with logistic growth in the target population (Sim 16).	79
4.9	<i>S^* scores for bottleneck scenarios with and without admixture.</i> (A) Density plot of S^* scores for a 1-bottleneck scenario (Demographic model D) with and without admixture. (B) Density plot of S^* scores for a 2-bottleneck scenario (Demographic model E) with and without admixture. (C) S^* score pairwise comparison between both bottleneck scenarios in absence of admixture. (D) Pairwise comparison of all S^* scores for both bottleneck scenarios with admixture. (E) Bottom density plot displaying distributions for top S^* scores in both bottleneck scenarios with admixture. Upper right-hand boxplot showing a pairwise comparison for the top S^* scores with 2% admixture, 1-bottleneck and 6% admixture, 2-bottlenecks scenarios, respectively. Top scores can only be obtained for admixture scenarios. . . .	80

4.10	<i>Pairwise comparisons and multigroup comparison of simulations with differing split times and admixture times. (A) Full distribution of S^* scores and (B) the top distribution of the S^* scores ($> 400,000$) for scenarios where 2% admixture took place at 35 KYA but archaic and human divergence took place at 350 KYA, 700 KYA, and 1 MYA, respectively. (C) The full distribution of S^* scores and (D) the top distribution of the S^* scores for scenarios where splits took place 1 MYA but 2% admixture between humans and an archaic took place at different time points (35 KYA and 100 KYA).</i>	81
4.11	<i>TMRCa versus S^*. Positive correlation coefficient (R) between S^* and TMRCa (in years) in scenarios without admixture (left) and with 4% admixture (right).</i>	82
4.12	<i>Pairwise comparison of exponential growth simulations with admixture and without admixture for both full and top distributions of S^* candidates. (A) TMRCa in years of candidate regions and (B) S^* scores for candidate regions.</i>	83
4.13	<i>Pairwise comparison of TMRCa in years for S^* for exponential (blue) and logistic (yellow) scenarios. (A) Comparison of the growth scenarios in absence of (left) and with admixture (right) for full S^* distribution. (B) Comparison of TMRCa in years for top S^* scores ($> 400,000$) when admixture is present. Under a no admixture scenario, top scores can only be obtained for exponential growth.</i>	84
4.14	<i>TMRCa in years for bottleneck scenarios with and without admixture. (A) Pairwise comparisons between the full distribution of S^* candidates for a 1-bottleneck scenario (Demographic model D) and a 2-bottleneck scenario (Demographic model E) with and without admixture. (B) Pairwise comparison of TMRCa in years for top S^* candidates for both scenarios with 2% admixture, 1-bottleneck and 6% admixture, 2-bottlenecks, respectively. Top scores could only be obtained for admixture scenarios.</i>	85
5.1	<i>Score and sample size. Pairwise comparisons between different sample sizes for top S^* and Sprime candidates. Top candidates are haplotypes that surpass the 400,000 score threshold.</i>	101
5.2	<i>TMRCa, S^*, Sprime and whole genome. Pairwise comparison of the time to most recent common ancestor between whole genome, top candidates of S^* and Sprime segments across different sample sizes.</i>	102
5.3	<i>TMRCa and sample size. Pairwise comparison of the time to most recent common ancestor between different sample sizes for estimates across the genome and top candidates detected by S^* and Sprime.</i>	103

5.4	<i>Comparison of S^* and Sprime on Papuan WGS from SGDP Data.</i> (A) Overlap between S^* and Sprime for different sample sizes in overall candidates, candidates that were above the score of 100,000, and in top candidates. Example: In a sample size of 2, 17.2% of S^* region overlap with Sprime detected regions overall and conversely, 88% of Sprime regions overlap with regions detected by S^* overall. (B) Overlapped with regions retrieved by Browning <i>et al.</i> (2018); Overall: all regions recovered; >100k: regions with an absolute score higher than 100,000; Top 1%: top 1% scoring fragments.	104
5.5	<i>TMRCa versus Sprime in Papuans.</i> Correlation coefficient (R) between Sprime score and TMRCa in Papuans for detected haplotypes which have been confirmed to be introgressed from Neanderthal and/or Denisova (left) and candidates which were found to have no match with either archaic reference (right).	106
5.6	<i>S^* and TMRCa across all Papuan and African SGDP samples.</i> (A) Multiple comparison of top S^* candidates for Papuans (n=15) and African populations (n=44) from SGDP Data. Top S^* candidates are haplotypes that surpass the 400,000 score threshold.	108
5.7	<i>S^* and TMRCa across subsamples Papuan and African SGDP samples.</i> (A) S^* score and (B) TMRCa multiple comparison (Kruskal-Wallis test) of top S^* candidates for Papuans and African populations. Groupings were subset to N=6 (Table S5.1; Supplementary Table). Top S^* candidates are windows that surpass the 400,000 score threshold.	109
5.8	<i>S^* and TMRCa distributions.</i> (A) Density plot for the full distribution of S^* scores for each grouping and (B) TMRCa pairwise comparison for all S^* candidates versus top S^* candidates. Top S^* candidates are haplotypes that surpass the 400,000 score threshold.	110
5.9	<i>Sprime and TMRCa.</i> (A) Sprime score and (B) TMRCa multiple comparison (Kruskal-Wallis test) of top Sprime candidates for Papuans and African populations. Groupings were subset to N=6 (Table X; Supplementary Table). Top Sprime candidates are haplotypes that surpass the 400,000 score threshold.	111
5.10	<i>TMRCa: Full versus Top.</i> TMRCa pairwise comparison for all Sprime detected haplotypes and top haplotypes detected by Sprime per grouping (N=6).	112
5.11	<i>TMRCa: Top shared regions.</i> TMRCa multiple comparison of top Sprime candidates for shared regions (red) and unique regions per African grouping.	113
5.12	<i>TMRCa: Shared versus unique regions.</i> Paired TMRCa (years) estimates, (Wilcoxon tests) for shared (red) and unique top Sprime candidate regions in San, Pygmies, West and East groupings.	114

5.13	<i>Sprime: HapMap versus AE maps.</i> Pairwise comparison of scores for top Sprime candidates detected using the African Enriched (AE; red) and HapMap YRI (YRI; blue) genetic maps.	117
5.14	<i>TMRCAs: HapMap versus AE maps.</i> Pairwise comparison (Wilcoxon test) of TMRCAs for top Sprime candidates detected using the African Enriched (AE; red) and HapMap YRI (YRI; blue) genetic maps.	118
5.15	<i>Length: HapMap versus AE maps.</i> Pairwise comparison of lengths (bp) for top Sprime candidates detected using the African Enriched (AE; red) and HapMap YRI (YRI; blue) genetic maps.	119
5.16	<i>D-statistics comparing two individuals to Vindija Neanderthal with the reconstructed common ancestor (Ancestral) as the outgroup.</i> The heatmap indicates whether individual on the Y-axis shares an excess of Neanderthal alleles (red) or an absence of them (blue).	121
5.17	<i>D-statistics comparing two individuals to Denisovans with Ancestral as the outgroup.</i> The heatmap indicates whether individual on the Y-axis shares an excess of Denisovan alleles (red) or an absence of them (blue).	122
5.18	<i>D-statistics comparing Neanderthal and Denisova to a modern human population with Ancestral as the outgroup.</i> The heatmap indicates which human populations share an excess of alleles with Denisova (red) or Neanderthal (blue).	124
5.19	<i>D-statistics for a pair of individuals to observe excess allele sharing to an archaic while the Ancestral remains as the outgroup.</i> Only derived mutations shared exclusively by both Denisova and Neanderthals were used here. If the value is positive (red) the individual on the Y-axis is closer to an archaic than the individual on the X-axis.	125
5.20	<i>D-statistics for a pair of individuals to observe excess allele sharing with Denisovans while the Ancestral remains as the outgroup.</i> Only positions where Denisovans had privately derived mutations were used. If the value is positive (red) the individual on the Y-axis is closer to Denisovans than the individual on the X-axis.	126
5.21	<i>D-statistics for a pair of individuals to observe excess allele sharing with Vindija while the Ancestral remains as the outgroup.</i> Only positions where Vindija had privately derived mutations were used. If the value is positive (red) the individual on the Y-axis is closer to Vindija than the individual on the X-axis.	127

5.22	<i>Proposed model.</i> Schematic depicting combined demographic histories of both modern humans and Eurasian archaics which may best explain the strong presence of ancient shared polymorphisms in African populations with largest effective population size. Allele sharing with Denisovans more than Neanderthal for some African populations may be explained by the stronger bottleneck in Neanderthal (Prufer <i>et al.</i> , 2014; Prufer <i>et al.</i> , 2017). Triangles indicate the occurrence of bottlenecks. Dotted lines suggest potential gene flow at the level of the basal lineages.	131
A.1	Density plots and corresponding boxplots for S^* distributions of simulations (A) without admixture, (B) with admixture, and for (C) SGDP data. f : full distribution, t : top candidate distribution.	164
A.2	Density plots and corresponding boxplots for TMRCA distributions of simulations (A) without admixture, (B) with admixture, and for (C) SGDP data. f : full S^* distribution, t : top candidate distribution.	165
A.3	D -statistics comparing two African individuals to a Non-African (English) with the reconstructed common ancestor (Ancestral) as the outgroup. The heatmap indicates whether individual on the Y shares an excess of alleles (red) or an absence of them (blue) with the English individual.	168
A.4	Correlation of D -statistics comparing Non-African and Vindija shared ancestry across African populations. Each point is a pairing of $P3$ versus $P4$ where the represented population is $P4$ in the calculation of $D(\text{Ancestral}, P2, P3, P4)$	169

List of Tables

2.1	Software for detecting TEs and viruses in WGS data.	24
2.2	Twenty simulated HIV integrations into human chromosome 1.	35
2.3	The sequencing samples analysed in benchmarking.	36
4.1	Parameters of the 26 simulations that were generated and analysed. The first column numbers the simulation modelled and the subsequent columns delineate the status of other parameters used. The demography column refers to the basis model used (Figure 4.3B-E) and which is the target population in the simulation. If a recombination map was used to simulate local recombination rate it was labelled as variable, whereas if a single recombination rate was used across the simulations, it was labelled constant. The sample column reports the number of diploid individuals simulated in each model. Dates of divergence between human and archaic as well as between Africans and Eurasians are given in years in separate columns. Date of admixture is also provided in years followed by a final column that reports proportion of archaic admixture simulated in the model. Simulations with no archaic admixture have "NA" under admixture date and 0 under admixture proportion (Null, 3, 11, 19). A: Archaic; H: Human Af: African; Eu: European.	72
5.1	Grouping of the 44 African individuals from the Simons Genome Diversity Project.	107
5.2	Putatively introgressed regions detected in other studies in both genic and non-genic regions. The "non-fully shared" section corresponds to Hsieh <i>et al.</i> (2016a) candidate regions which in this study were not shared by all African populations. Bottom section lists loci unrecovered by Sprime. All coordinates correspond to hg19 reference positions for each chromosome. The "Gene Locus" column denotes the genic region it is in or the known chromosomal arm notation from a previous study. Columns for each geographical or population grouping indicate whether these previously detected candidate regions in other studies are present (+) in this study or absent (-).	115
A.1	The Cancer Genome Atlas 70 patients analysed in this study.	162
A.2	Polymorphic integrations found in 70 human genomes.	163

A.3	Subsets of African individuals from the Simons Genome Diversity Project per target grouping.	166
A.4	Eurasian outgroup used when performing S^* and Sprime on African target populations.	167

