

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Risk diplotypes promoting recombination error in hybrid and non-hybrid human lineages

A thesis by Róisín Ní Chuaig

Supervisor: Professor Russell McLaughlin

Trinity College Dublin

2026

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. I confirm that the work has been completed in full compliance with relevant university policies, including the Academic Integrity policy, the Trinity Policy on Good Research Practice, and the guidance on AI and GenAI in Teaching, Learning, Assessment and Research.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Signed 

Date 30/09/2025

Summary

This thesis investigates the impact of genetic variation on the meiotic recombination landscape and its consequences in modern human genomes. The central hypothesis involves PRDM9, a protein that directs the positioning of recombination hotspots by binding to specific DNA sequences. We posit that changes affecting the binding sites of PRDM9 can predispose the genome to non-allelic homologous recombination (NAHR), a primary mechanism that leads to copy number variants (CNVs). This is explored in the context of common polymorphisms within modern human populations as well as the large-scale divergence introduced through archaic introgression.

A comprehensive map of sequence variants that theoretically alter PRDM9 binding site motifs across diverse human populations was developed. This revealed that such variants are exceptionally common and are subject to purifying selection, indicating that they are generally deleterious. This selective pressure is particularly strong for variants that disrupt existing binding sites, compared to those that create new ones.

The thesis then builds on these principles by demonstrating that archaic introgressed segments act as persistent hotspots of genomic instability in the modern human genome. A key finding is that deletions are highly enriched at the boundaries of these archaic segments. This was supported through multiple lines of evidence, including patterns of read depth, the overlap of called deletions with regions flanking archaic segments, and the significant association between deletion breakpoints and the edges of archaic segments. This is the exact genomic pattern expected from NAHR events that occur between modern and archaic sequences.

This research also suggests that the distribution of archaic ancestry in modern human genomes may be influenced by recombination activity. Specifically, archaic segments are significantly depleted in regions of high recombination rate. This pattern is consistent with the possibility that the deletions observed at archaic segment boundaries potentially arising through NAHR may also contribute to the gradual, long-term purging of archaic sequences from the modern human genome.

This thesis links genetic variation to structural genomic instability through PRDM9-mediated recombination. It demonstrates how common polymorphisms may have the ability to create NAHR risk and are therefore evolutionarily constrained. Most significantly, it reveals that the legacy of archaic introgression may extend beyond the simple inheritance of genes. Archaic segments may be actively contributing to ongoing structural variation in the modern human genome by altering the PRDM9 binding landscape. The findings suggest that the admixture of human and Neanderthal DNA have introduced a source of meiotic instability, which has subsequently shaped both the structural landscape and the ancestry distribution of the modern human genome.

Table of Contents

1 Introduction 1

1.1 The Complex History of Human Evolution 1

1.2 The Genomic Revolution in Mapping Human Variation 6

1.3 Meiotic Recombination 9

1.4 Scope and Structure of Thesis 11

2 Datasets and General Methods 13

2.1 Summary of Terms used in Table 2.1 13

2.2 Dataset overview 15

3 SNP-Driven Changes in PRDM9 Binding Sites within Human Populations 18

3.1 Introduction 18

3.1.1 Meiotic Recombination and PRDM9 18
3.1.2 Structural Variation 20
3.1.3 Non-Allelic Homologous Recombination 22
3.1.4 Research Aims 27

3.2 Materials and Methods 29

3.2.1 Mapping Gains and Losses 29
3.2.2 Allele Frequencies 30
3.2.3 NAHR Risk Diplotypes in Human Populations 31
3.2.4 CNVs Explainable by Risk Diplotypes 31
3.2.5 Microdeletion and Microduplication Syndromes 32
3.2.6 Trio Analysis 33

3.3 Results 35

3.3.1 Allele Frequencies 35
3.3.2 NAHR Risk Diplotypes in Human Populations 39
3.3.3 CNVs Explainable by Risk Diplotypes 41
3.3.4 Microdeletion and Microduplication Syndromes 42
3.3.5 Trio Analysis 44

3.4 Discussion 48

4 The Genomic Legacy of Interbreeding between Humans and Neanderthals 53

4.1 Introduction 53

4.1.1 Archaic Genomes and their Modern Legacy 53
4.1.2 Inferring Archaic Introgression 54
4.1.3 Research Aims 57

4.2 Materials and Methods 58

4.2.1 Mapping Archaic Segments 58
4.2.2 Analysis of Archaic Deserts 61
4.2.3 Principal Component Analysis 63

4.3 Results 64

4.3.1 Preamble 64
4.3.2 Initial Analysis of Archaic Peaks 64
4.3.3 HGDP Dataset 68
4.3.4 Quality Control of the Irish Dataset 70
4.3.5 Identification and Analysis of Archaic Segments 74

4.3.6 Archaic Deserts 81
4.3.7 Principal Component Analysis 91

4.4 Discussion 101

5 Archaic SNPs as Drivers of NAHR in the Modern Human Genome 105

5.1 Introduction 105
5.1.1 Hybridisation 105
5.1.2 Recombination in Hybrids 106
5.1.3 Implications for Human-Neanderthal Hybrids 109
5.1.4 Research Aims 111

5.2 Materials and Methods 112
5.2.1 Private Variant Analysis 112
5.2.2 Deletions Flanking Archaic Segments 113
5.2.3 Archaic Segment Depletion in the Modern Human Genome 116

5.3 Results 118
5.3.1 Private Variant Analysis 118
5.3.2 Deletions Flanking Archaic Segments 121
5.3.3 Archaic Segment Depletion in the Modern Human Genome 127

5.4 Discussion 137

6 Discussion 142

6.1 Summary of Findings and their Significance 142

6.2 The Depletion of Neanderthal Ancestry in the Modern Human Genome 145

6.3 Limitations and Future Directions 148

6.4 Conclusion 153

Appendix A 155

Appendix B 184

Appendix C 206

References 210

List of Abbreviations

AB Allele balance

ALS Amyotrophic lateral sclerosis

ARI Adjusted Rand index

BDMI Bateson-Dobzhansky-Muller incompatibility

CEPH Centre d'Etude du Polymorphisme Humain

ChIP-seq Chromatin immunoprecipitation sequencing

cM centiMorgan

CNV Copy number variant

DAV Denisovan, Altai, and Vindija

DNA Deoxyribonucleic acid

DSB Double-strand break

FIMO Find individual motif occurrences

GQ Genotype quality

GRCh37 Human Genome Reference Consortium genome build '37'

GRCh38 Human Genome Reference Consortium genome build '38'

GWAS Genome-wide association study

H3K36me3 Trimethylation of histone H3 at lysine 36

H3K4me3 Trimethylation of histone H3 at lysine 4

HGDP Human Genome Diversity project

HGP Human Genome Project

HMM Hidden Markov model

HWE Hardy-Weinberg equilibrium

ILS Incomplete lineage sorting

Indel Insertion/deletion

LCR Low-copy repeat

LD Linkage disequilibrium

LoF Loss-of-function

MMS Microdeletion and microduplication syndromes

NAHR Non-allelic homologous recombination

NGS Next-generation sequencing

o/e observed/expected

PCA Principal component analysis

pLI Probability of loss-of-function intolerance

PRDM9 PR/SET domain 9

PV Private variant outside of segments

PVAS Private variant within an archaic segment

PWM Position weight matrix

QC Quality control

RD Read depth

SGDP Simons Genome Diversity Project

sLDSC Stratified linkage disequilibrium score regression

SNP Single nucleotide polymorphism

SNV Single nucleotide variant

SV Structural variant

UCSC University of South California, Santa Cruz

VQSR Variant Quality Score Recalibration

List of Key Terminology

Archaic count The number of individuals of a given dataset inferred to carry an archaic haplotype within each 1 kb window

Archaic desert A large region of the genome that is depleted of archaic introgression

Archaic segment A DNA segment in a modern human genome that has been inherited from an archaic hominin (Neanderthal or Denisovan) through introgression

Callable region A genomic region with sufficient quality and coverage (described in section 2.2)

***De novo* CNV** A deletion or duplication that is present in a child but absent from both parents' genomes

DSB hotspot Experimentally validated region in which meiotic recombination takes place (locations identified by Pratto et al. (2014))

Gain-causing SNP/Gain A SNP that creates a new potential PRDM9 binding site motif

Gain-gain diplotype A genotype where both chromosomal haplotypes carry a gain-causing SNP within 200 kb of each other, thereby increasing the risk of NAHR between these two novel PRDM9 binding sites

Gain-loss diplotype A genotype where one of the chromosomal haplotypes carries a gain-causing SNP and a nearby (within 200 kb) loss-causing SNP, thereby increasing the risk of NAHR between the novel PRDM9 binding site and the remaining PRDM9 binding site on the opposing haplotype

Loss-causing SNP/Loss A SNP that abolishes an existing PRDM9 binding site motif

Loss-loss diplotype A genotype where both chromosomal haplotypes carry a loss-causing SNP within 200 kb of each other, thereby increasing the risk of NAHR between the two remaining PRDM9 binding sites

PRDM9 motif A specific short DNA sequence that is recognised and bound by PRDM9, which defines the location of a potential recombination hotspot

Risk diplotype A specific combination of two haplotypes that increases the risk of NAHR and the generation of a *de novo* CNV

Acknowledgements

I would firstly like to thank my supervisor, Professor Russell McLaughlin, for his incredible guidance and support throughout my PhD. I feel very fortunate to have had the opportunity to work under his supervision and receive consistent encouragement and clear feedback that has been invaluable to me. This thesis would not have been possible without his direction and support. A special thank you to Ross Byrne, whose help along the way has been greatly appreciated.

I would also like to extend my appreciation to my friends and family for continuously encouraging me through this challenging journey. A particular thanks to Tim Steybe, who has been an incredible source of support from start to finish. Thank you to everyone who has contributed and supported me on this path, whether in an academic context or through moral support.

1 Introduction

The human genome is not static – neither over evolutionary time, nor from one generation to the next. Changes accumulate through mutation, recombination, structural rearrangements, and gene flow, leaving behind a record of evolutionary processes in our DNA. Understanding these processes has been one of the central challenges of modern genomics, which requires both a reconstruction of the deep evolutionary past of *Homo sapiens* and a detailed map of variation within and between populations today.

1.1 The Complex History of Human Evolution

The evolutionary history of *Homo sapiens* is complex, beginning in Africa approximately 2.5 million years ago. From this continent of origin, multiple waves of hominins dispersed into Eurasia. Our common ancestor was *Homo heidelbergensis*, a species whose European descendants evolved into the Neanderthals (*Homo neanderthalensis*) and whose Asian descendants gave rise to the Denisovans, groups that became adapted to the environments of Pleistocene Eurasia for several hundred thousand years (Meyer et al., 2016; Stringer, 2012). Meanwhile, in Africa, our own lineage was evolving towards modern humans. Fossil evidence from sites like Jebel Irhoud in Morocco (Hublin et al., 2017) and Omo Kibish in Ethiopia (McDougall et al., 2005) suggests that morphologically recognisable *Homo sapiens* were present by at least 300,000 years ago.

A major event in human history was the dispersal of modern humans out of Africa, which likely occurred in multiple stages but with a significant pulse around 60,000 years ago (Gronau et al., 2011; Li & Durbin, 2011; Schiffels & Durbin, 2014). This brought them into the regions long occupied by Neanderthals and Denisovans. For most of the twentieth century, the prevailing view was that these expansions led to the complete replacement of other hominins, involving little or no admixture (Stringer & Andrews, 1988). However, this idea has changed dramatically with the advent of ancient DNA studies.

The sequencing of the Neanderthal genome provided the first direct evidence of admixture between archaic and modern humans. Green et al. (2010) demonstrated that non-African populations today carry ~1-4% Neanderthal ancestry. Soon after, Denisovan DNA recovered from a finger bone in Denisova Cave, Siberia, revealed another archaic contributor to the modern human gene pool, with Denisovan ancestry persisting mostly in Oceania and some Asian populations (Meyer et al., 2012; Reich et al., 2010). Subsequent work has shown that these admixture events were not rare, isolated incidences but instead were part of a broader pattern of interbreeding among hominin groups across Eurasia (Browning et al., 2018; Prüfer et al., 2014; Vernot & Akey, 2014).

Following archaic admixture, the resulting introgressed genomic segments persist in modern human genomes in non-random patterns, revealing a complex history of natural selection acting as a potential editor of the hybrid modern human genome (Sankararaman et al., 2016; Telis et al., 2020). Signatures of both negative and positive selection have been described, determining which archaic alleles have been purged as detrimental and which were retained as beneficial (Dannemann & Kelso, 2017; Sankararaman et al., 2014).

Amongst the most widely cited evidence for widespread negative selection against introgressed material is the existence of large archaic deserts, i.e. genomic regions which are apparently completely devoid of Neanderthal and Denisovan ancestry. Genome-wide scans have identified these deserts as being significantly enriched on the X chromosome (Sankararaman et al., 2014) and in genes with critical biological functions, particularly those highly expressed in the testis (Sankararaman et al., 2016) and those involved in brain development (Reilly et al., 2022). This non-uniform distribution supports the role of Bateson-Dobzhansky-Muller incompatibilities (BDMIs), which describes independent variation in diverging lineages that interact negatively when brought together in hybrids (Bateson, 1909; Dobzhansky, 1934; Muller, 1942). When previously isolated genetic backgrounds are suddenly mixed through hybridisation, certain combinations of alleles can produce negative fitness consequences, resulting in reduced reproductive success in hybrid offspring. The significant reduction on the X chromosome (which is known as the "large X-effect") is a classic signature of BDMIs, as recessive incompatible alleles on the X chromosome are exposed in the hemizygous state in males (Harris & Nielsen, 2016; Reilly et al., 2022). Similarly, genes essential for male fertility (often involved in spermatogenesis) are some of the fastest-evolving genes in the genome and are frequent hotspots for hybrid incompatibilities, which would explain the strong selective pressure to remove archaic variants from these regions (Harris & Nielsen, 2016; Reilly et al., 2022).

Neanderthals lived in small, isolated populations with low genetic diversity for hundreds of thousands of years. As a result of this, their effective population size was much smaller than humans, estimated to be fewer than 3,500 in females (Sánchez-Quinto & Lalueza-Fox, 2015). Thus, genetic drift would have dominated over natural selection and allowed weakly deleterious alleles to accumulate in the population as if they were neutral. Simulations based

on the distribution of mutational fitness effects indicate that Neanderthals likely carried a significantly higher mutational load compared to modern human populations, with estimated fitness reductions of at least 40% compared to humans (Harris & Nielsen, 2016). When modern humans and Neanderthals interbred, this deleterious genetic burden was introduced into the modern human gene pool. However, the much larger effective population size of modern humans would have allowed natural selection to act more efficiently against these introgressed deleterious alleles. Harris & Nielsen (2016) calculated that Neanderthal introgression likely introduced a residual deleterious mutational load that would have reduced fitness in non-African populations by at least 0.5%. Studies suggest that negative selection acted immediately, with many deleterious Neanderthal alleles purged within just a few generations (Iasi et al., 2024; Sümer et al., 2025).

Adaptive Neanderthal alleles have also been identified, where alleles provided adaptive benefits to modern humans as they dispersed out of Africa into new Eurasian environments, to which Neanderthals had already been adapted for hundreds of thousands of years (Iasi et al., 2024; Reilly et al., 2022; Sümer et al., 2025). For example, Neanderthal alleles in genes like *BNC2, POU2F3,* and *MC1R* affect skin pigmentation, ultraviolet light sensitivity, and hair texture (Dannemann & Kelso, 2017; Reilly et al., 2022). These traits were likely advantageous for regulating vitamin D synthesis and thermoregulation in non-African latitudes with lower and more seasonal ultraviolet radiation. Another strong signal of adaptive introgression is found in the immune system, particularly the OAS gene cluster (which helps combat viral infections) and the TLR genes (involved in pathogen recognition) (Enard & Petrov, 2018; Sams et al., 2016). Neanderthals provided modern humans with alleles for fighting off Eurasian pathogens they had never before encountered. This effectively allowed modern humans to acquire beneficial immunity variants without waiting

for new mutations to arise. Furthermore, a Denisovan-like haplotype of the *EPAS1* gene, which regulates haemoglobin production in response to hypoxia, has been identified in Tibetans (Huerta-Sánchez et al., 2014). This is thought to be a key adaptation for life at high altitude.

The adaptive value of an introgressed allele, however, may not always have been immediate. The concept of delayed selection posits that some Neanderthal alleles were neutral or nearly neutral upon introgression and only became advantageous later, in response to novel environmental pressures or shifts in human lifestyle (Yair et al., 2021). This can be distinguished from immediate positive selection by examining the structure of the introgressed haplotypes. Immediate selection would have caused a long, uninterrupted Neanderthal haplotype to sweep to high frequency. In contrast, the signature of delayed selection would be a shorter and more refined haplotype, indicating that the allele was initially at low frequency for some time, allowing recombination to break up the original introgressed segment before recent positive selection acted on the beneficial core variant (Yair et al., 2021). This pattern has been observed in key regions, such as the *OAS1* locus (related to viral defence), suggesting its adaptive significance, such as protection against severe COVID-19, may be a more recent phenomenon (Zeberg & Pääbo, 2021). It is important to note that alternative forces, such as linkage to a recently mutated beneficial allele or the recessive nature of the archaic variant, can also produce similar patterns (Yair et al., 2021).

Finally, the genetic contribution of Neanderthals was not limited to novel alleles. They also served as a reservoir for ancestral variants that had been lost in the modern human lineage, likely due to genetic drift during the out-of-Africa bottleneck. Rinker et al. (2020)

demonstrated that these reintroduced ancestral alleles were more likely to be tolerated in the modern human genetic background than uniquely Neanderthal-derived alleles. This suggests that admixture acted as a partial reversal of the bottleneck, restoring ancient genetic diversity that had been lost and providing a broader range of variation upon which natural selection could act. Thus, the legacy of Neanderthal introgression is a complex mixture of loss, gain, and reacquisition.

The vast majority of our understanding of archaic introgression over the past 15 years has derived from technological advances in DNA sequencing, which has enabled the identification of genetic variation in human genomes. The primary method for detecting archaic segments in the modern human genome relies on analysis of this genetic variation to detect divergent haplotypes, discussed in more detail in chapter 4. By comparing the genomes of modern humans to the Neanderthal and Denisovan genomes, computational methods can identify genomic regions in modern humans where the allele frequencies and haplotype patterns are more similar to the archaic genomes than to the modern human background (Skov et al., 2020; Vernot & Akey, 2014). These methods have provided profound insights, such as archaic deserts and adaptive introgression, as well as allowing for a more precise estimation of the timing and number of admixture pulses between modern humans and archaic hominins. However, these insights are predominantly based on SNVs, leaving the role of archaic structural variation as an open question.

1.2 The Genomic Revolution in Mapping Human Variation

The journey from inferring history from fossils and modern populations to directly sequencing ancient genomes and mapping global diversity is one of the most significant scientific revolutions of the past half-century. This began with the monumental effort to

obtain a single reference genome and has now developed into large-scale population genomics, enabling the detailed cataloguing of human genetic variation on an unprecedented scale.

The first-ever reference sequence of a human genome was provided in the period between 2001 and 2003 with the completion of the Human Genome Project (HGP). The HGP's reference genome served as a coordinate system and baseline for all future comparisons (Lander et al., 2001). It revealed that the human genome is a vast and complex landscape, with approximately 3 billion base pairs and an unexpectedly small number of protein-coding genes (~20,000-25,000), which highlights the importance of non-coding regulatory regions. The International HapMap Project (2002-2009) began soon after, which created a catalogue of common SNPs and patterns of linkage disequilibrium (LD) across multiple populations (International HapMap Consortium, 2005). By identifying "tag SNPs" that could represent entire haplotype blocks, HapMap enabled the genome-wide association study (GWAS) era, which led to the discovery of thousands of genetic loci associated with complex traits and diseases (Sollis et al., 2023).

While HapMap and GWAS were revolutionary, their technologies limited the study of human genomics to common variants that were represented on pre-designed genotyping arrays. The development of next-generation sequencing (NGS) technologies in the mid-2000s enabled millions of parallel sequencing reactions to be performed on clonally amplified DNA fragments, which massively reduced the cost of sequencing and enabled a transition from genotyping known common variants to discovering the full spectrum of allele frequencies across a more diverse set of genetic variants. This includes small insertions and deletions (indels), structural variants (SVs), tandem repeats and repeat expansions. A major

limitation of short-read NGS, however, is the technology's precision in SV mapping, particularly in detecting large SVs. While it can detect SVs through indirect signals like read depth, split reads, and paired-end mappings, the resolution is often poor. Precise breakpoint determination and phasing of SVs are difficult, and sensitivity is further reduced in complex, repetitive genomic regions (Chaisson et al., 2019). Consequently, a significant portion of genomic diversity and its functional impact, especially from SVs, remained poorly characterised in early NGS-based maps.

Nonetheless, NGS has been used by several major projects to create comprehensive maps of human genetic diversity. The 1000 Genomes Project was the first large-scale effort to apply NGS to a diverse cohort. This project identified over 88 million variants from 2,504 individuals from 26 populations (Auton et al., 2015). While this was a huge development for SNVs and indels, its catalogue of SVs was less complete due to the limitations of short-read technology. The Human Genome Diversity Project (HGDP)–Centre d'Etude du Polymorphisme Humain (CEPH) and the Simons Genome Diversity Project (SGDP) focused on sequencing globally diverse populations, which are often underrepresented in large biobanks (Bergström et al., 2020; Mallick et al., 2016). Since then, these publicly available datasets have become essential resources for many population genetic and bioinformatic analyses. Furthermore, the work of deCODE genetics in Iceland has pioneered the model of population-scale sequencing. By sequencing a large proportion of the Icelandic population, deCODE demonstrated the ability to trace the history of mutations through pedigrees, identify rare variants with large effects on disease, and perform association studies with high resolution (Gudbjartsson et al., 2015). This model of genotyping in combination with phenotyping has since been used in other flagship projects such as the UK Biobank (Sudlow

et al., 2015; UK Biobank Whole-Genome Sequencing Consortium, 2025), FinnGen (Kurki et al., 2023), and others.

The limitations of short-read NGS technology mean that even these large-scale projects provided much less information about SVs than SNVs. Consequently, the overall understanding of the functional impact of SVs remains lower than that for SNPs. Nevertheless, researchers have made many attempts to overcome these hurdles. For example, studies have successfully linked *de novo* CNVs to neurodevelopmental disorders such as autism and schizophrenia (Jeste & Geschwind, 2014; Rosenfeld & Patel, 2017; Szecówka et al., 2023; Vicari et al., 2019). Furthermore, the recent whole-genome sequencing of 490,640 UK Biobank participants has provided enormous power to study SVs in different populations, which already revealed that the number of SVs varies significantly across different ancestral groups (UK Biobank Whole-Genome Sequencing Consortium, 2025). These studies are now beginning to connect SVs to complex traits, though the overall understanding of SVs still lags far behind that of SNPs due to detection challenges. Nevertheless, their prevalence and their generation through fundamental biological processes like meiotic recombination highlight a critical role in human health, disease, and evolution.

1.3 Meiotic Recombination

Large-scale sequencing projects and recently developed databases of human genetic variation have begun to characterise not only the variation that exists between and within human populations, but also the history and evolution of this variation. These developments now allow us to pursue more nuanced questions about how genetic variation may affect other

processes such as meiotic recombination, one of the most fundamental and important biological mechanisms.

Meiotic recombination is a highly orchestrated process concentrated in locations known as recombination hotspots. A key protein involved in the determination of these recombination hotspot locations is PRDM9. PRDM9 binds to specific short DNA motifs via its zinc-finger array and marks the site for the initiation of double-strand breaks (DSBs) (Altemose et al., 2017). High-throughput technologies, such as ChIP-seq and specialised assays like single-strand DNA sequencing, have successfully mapped PRDM9 binding sites and the direct locations of meiotic DSBs (Brick et al., 2012). Furthermore, the large datasets of human variation from projects like the 1000 Genomes Project have been used to infer historical recombination rates through statistical methods such as LD-based mapping (Zhou et al., 2020). These approaches have revealed that the zinc-finger domain of PRDM9 is one of the most rapidly evolving regions in the human genome, resulting in population-specific differences in hotspot usage due to changes in the protein's binding specificity (Oliver et al., 2009). In addition, the PRDM9 binding site motifs themselves are also subject to rapid evolution, a phenomenon known as hotspot erosion. These concepts are discussed in more detail in chapter 3.

Meiotic recombination can be a major source of genomic instability, with errors during the complex process of DSB repair often leading to the formation of *de novo* CNVs. Non-allelic homologous recombination (NAHR) is a well-established mechanism describing the misalignment between highly homologous sequences during meiosis that can lead to deletions, duplications, and inversions (Burssed et al., 2022). NAHR has been shown to cause a wide range of genomic disorders involving SVs (Carvalho & Lupski, 2016; Dittwald

et al., 2013; Parks et al., 2015), often recurring in unrelated individuals due to the predisposition of the region to ectopic recombination. The misalignment of sequences and consequent NAHR is generally attributed to segmental duplications in the genome. However, the advancement of sequencing technologies and the access to large genomic databases now allow us to investigate a more dynamic hypothesis: that specific genetic variants can actively modulate the fidelity of the meiotic process, potentially influencing an individual's risk of generating *de novo* CNVs.

1.4 Scope and Structure of Thesis

This thesis aims to explore the role that pre-existing sequence variation plays in eliciting recombination error when variants come together in specific combinations – either risk diplotypes effecting *de novo* CNV over a single generation, or longer-term association of copy number duplication and deletion with sequence divergence at hybrid sites incorporating lineages from multiple *Homo* subspecies. Chapter 3 focuses on single nucleotide variation within modern human populations. This chapter utilises multiple genomic datasets to identify SNPs that alter PRDM9 binding motif affinity and evaluate their potential role in eliciting the meiotic errors that lead to *de novo* CNVs. Chapter 4 shifts to an evolutionary perspective, inferring archaic introgression in modern human populations and analysing these segments for evidence of negative selection, thereby mapping the surviving legacy of ancient admixture. Finally, chapter 5 combines these themes by investigating the genomic differences between modern human and archaic DNA as potential drivers of NAHR and CNV formation. We test the hypothesis that the divergence between Neanderthals and humans created different PRDM9 binding site landscapes, and that their combination in a hybrid genome could predispose to ectopic recombination, resulting in deletions and duplications. In doing so, this work explores the dynamic nature of the human genome, the

biological impact of subspecies hybridisation and the evolutionary legacy of archaic introgression. By dissecting the interplay between sequence variation and genome stability, this thesis seeks to highlight biological processes that are fundamental to human evolution, health and disease.

2 Datasets and General Methods

Several genomic datasets were used in this study, each subjected to different quality control (QC) and filtering procedures. Each dataset and the filtering steps applied are detailed here along with their distinct name used throughout the thesis. Table 2.1 summarises the application of these steps for each dataset.

2.1 Summary of Terms used in Table 2.1

GATK HaplotypeCaller: A variant caller that uses a local *de novo* assembly approach to identify SNPs and indels, improving sensitivity for complex variants.

Variant quality score recalibration (VQSR): A dynamic filtering process that utilises two steps, GenomicsDBImport and GenotypeGVCFs. VQSR calculates a quality score for each variant by considering multiple annotated parameters (quality depth, mapping quality, etc.). This multi-dimensional assessment allows VQSR to evaluate the overall quality of the variant rather than relying on individual annotations. To retain the variants with the highest quality, a tranche sensitivity threshold of 99.7% was selected, meaning that 0.3% of variants with the lowest quality scores were removed.

ExcessHet filtering: The removal of sites with excessively high heterozygosity, which can indicate alignment or variant calling artifacts.

Hardy-Weinberg equilibrium (HWE) filtering: Variants that deviate significantly from HWE ($p < 0.001$) within any population were excluded, as this can indicate genotyping errors or selection pressures.

Genotype quality (GQ) filtering: Individual genotypes with low GQ scores were set to missing.

Read depth (RD) filtering: Genotypes with read depth exceeding 1.65x the genome-wide mean were removed.

Multi-allelic SNP filtering: At multi-allelic sites, only the two most common SNP alleles were retained, with rare or indel alleles masked as missing.

Sample Filtering: Samples with abnormal numbers of SNVs, high singleton counts, or extreme heterozygosity were removed. Samples with contamination >5%, chimeras >5%, or median insert size <250 bp were also excluded. Additional outlier detection was performed using ancestry-adjusted principal components.

Inbreeding coefficient filtering: Variants with a highly negative inbreeding coefficient were excluded, as this suggests an excess of heterozygotes, indicative of poor-quality sequencing.

Absence of high-quality genotypes: Variants were excluded if they had no high-quality genotypes (GQ ≥ 20, DP ≥ 10, AB ≥ 0.2 for heterozygotes).

Table 2.1, A summary of all processing and filtering steps applied to each dataset.

Dataset Name	*Reference Genome*	*Sample Size*	*Processing and Filtering Steps Applied*
1000G Trio	GRCh38	300	**Variant calling:** GATK HaplotypeCaller **Variant filtering:** VQSR, ExcessHet, HWE ($p < 0.001$) **Genotype filtering:** GQ < 10
HGDP	GRCh38	809	**Variant calling:** GATK HaplotypeCaller **Variant filtering:** VQSR, ExcessHet, HWE ($p < 0.001$), Multi-allelic SNP filtering **Genotype filtering:** GQ ≤ 20, RD filtering
HGDP+1000G	GRCh38	3,147	**Variant calling:** GATK HaplotypeCaller **Sample filtering** **Variant filtering:** VQSR, Inbreeding Coeff. < -0.3, absence of high-quality genotypes, HWE ($p < 0.001$) **Additional filters applied:** GQ < 10
Irish	GRCh37	408	**Variant calling:** GATK HaplotypeCaller **Variant filtering:** VQSR, ExcessHet, Inbreeding Coeff. < -0.8 **Genotype filtering:** GQ < 10

2.2 Dataset overview

1000G Trio Dataset

CRAM files for these individuals (Byrska-Bishop et al., 2022) were obtained from the International Genome Sample Resource (IGSR; https://www.internationalgenome.org) and 100 trios (300 individuals) were selected to be processed in-house as described above. SV calls for this dataset were obtained from a comprehensive CNV VCF file generated using DELLY (Rausch et al., 2012), an integrated SV discovery pipeline that combines paired-end mapping, split-read analysis, and RD information to identify genomic rearrangements.

HGDP Dataset

This dataset was obtained from the Human Genome Diversity Project (HGDP) via the IGSR and was pre-filtered according to the methods detailed by Bergström et al. (2020) and outlined in Table 2.1. Upon retrieval of the pre-filtered data, HWE filtering was applied to all populations.

HGDP+1000G Dataset

This merged dataset combines individuals from HGDP and the 1000 Genomes Project and was obtained from (Koenig et al., 2024) via gnomad (https://gnomad.broadinstitute.org/data). Their processing of the dataset is described in Table 2.1. Two additional filtering steps were carried out upon retrieval of the data: GQ and HWE filtering.

Irish Dataset

This dataset represents an Irish case/control cohort for the neurodegenerative disease amyotrophic lateral sclerosis (ALS; 272 cases, 136 controls) sequenced by Illumina, inc (San Diego) as part of Project MinE (Project MinE ALS Sequencing Consortium, 2018). Data were either used as provided by Illumina (individual-level VCF calls generated by the Isaac variant caller, later termed the "unfiltered" dataset) or reprocessed from raw BAM alignments in-house as described in Table 2.1 ("filtered" dataset).

Callability

Callability files were generated for both GRCh37 and GRCh38 genomes. Following recommendations from Skov et al. (2020), we incorporated a strict mask file from the 1000 Genomes Project (Zheng-Bradley et al., 2017) and a repeat mask file from the UCSC

Genome Browser (Hinrichs et al., 2006). Additionally, all autosomal regions spanning 10 kb and containing fewer than 200 variants in the 1000 Genomes database (Auton et al., 2015) were removed. Similarly, 10 kb regions on the X chromosome with fewer than 30 variants were discarded.

3 SNP-Driven Changes in PRDM9 Binding Sites within Human Populations

3.1 Introduction

3.1.1 Meiotic Recombination and PRDM9

Meiotic recombination is a fundamental biological process that ensures the creation of genetic diversity. The process is initiated by double-strand breaks (DSBs) in DNA, which are repaired using homologous chromosomes as templates. The initiation of DSB formation is regulated by PRDM9, a zinc finger protein that binds to specific DNA motifs to determine the locations of these breaks. PRDM9 contains a DNA-binding domain that recognises certain motifs in the genome, allowing it to direct DSBs to specific regions. The SET domain of PRDM9 trimethylates histone H3 at lysines 4 (H3K4me3) and 36 (H3K36me3) (Eram et al., 2014; Wu et al., 2013), which then act as epigenetic markers for the initiation of meiotic recombination by creating a central nucleosome-depleted region where the DSBs can occur (Parvanov et al., 2017). Following methylation by PRDM9, SPO11 is recruited to create these DSBs which are then resected to generate 3′ single-strand overhangs (Lange et al., 2016). These single-stranded ends are then bound by the meiotic recombinase DMC1 which facilitates strand invasion of the homologous chromatid (Baker et al., 2014). The majority of these breaks are repaired as non-crossover events, but a subset are resolved as crossovers, which are important for proper chromosome disjunction during meiosis.

The role of PRDM9 in recombination was first elucidated by the mapping of DMC1-associated DNA, which showed that nearly all DSBs overlap with sites of PRDM9-

dependent H3K4me3 (Brick et al., 2012; Smagulova et al., 2011). The zinc finger domain of PRDM9 directs the positioning of recombination hotspots by binding to specific DNA sequences. While the overall protein structure of PRDM9 is highly conserved across species, the zinc finger domain displays wide variation. This domain can consist of anywhere from 8 to over 20 zinc fingers, and the three amino acids within each finger that dictate DNA-binding specificity are highly polymorphic (Berg et al., 2010; Paigen & Petkov, 2018). Despite all these zinc fingers binding DNA, only 4 to 6 make a substantial contribution to recognising the motif, with the remaining fingers playing a secondary role in sequence specificity and stability of the binding interaction (Paigen & Petkov, 2018). PRDM9 binding sites are generally short, degenerate DNA sequences that are recognised by this zinc finger domain. The DSBs initiated by PRDM9 do not occur directly at the PRDM9 binding site motifs, but instead may occur up to 300 bp away from the motif (Lam & Keeney, 2015b; Pratto et al., 2014).

Studies in mouse models have revealed that while PRDM9 can methylate thousands of sites, only a fraction of these actually become recombination hotspots. Baker et al. (2014) demonstrated this phenomenon by analysing mouse strains with different *Prdm9* alleles, identifying approximately 4,700 sites marked by PRDM9-mediated methylation, yet only 250-400 of these locations actually acquired a DSB. In contrast, Altemose et al. (2017) performed ChIP-seq in a mitotic human cell line transfected with the human *PRDM9-B* allele and identified 170,198 PRDM9 binding peaks across the human genome, demonstrating that PRDM9 has the ability to bind to many sites, even outside of recombination hotspots. However, these results are not entirely informative of the behaviour of PRDM9 during meiosis as the mitotic chromatin environment in these cultured cells differs from that of a cell during meiosis, which likely influences PRDM9 binding patterns.

The interaction between PRDM9 and its binding sites creates an evolutionary paradox which results in the continuous turnover of recombination hotspots. As recombination involves the precise exchange of genetic material between paired chromosomes, a key mechanism within this exchange is gene conversion. This is a non-reciprocal process where a small stretch of DNA from one chromosome is used as a template to overwrite and correct the corresponding sequence on the partner chromosome. As PRDM9 directs the recombination to specific sequences, the resulting gene conversion events gradually mutate or overwrite the DNA motifs recognised by PRDM9, thereby eroding its binding sites (Auton et al., 2012). This process, called 'hotspot erosion', occurs relatively quickly, with Auton et al. (2012) demonstrating a complete lack of hotspot sharing between humans and chimpanzees. This creates a selective pressure for new *PRDM9* alleles that can recognise alternative sequences, leading to rapid evolution of the zinc finger domain through positive selection. This concept is known as the Red Queen hypothesis, which suggests that PRDM9 must continually evolve new binding specificities to compensate for lost target motifs (Boulton et al., 1997). Evidence for this model was provided by Schwartz et al. (2014), who sequenced the *PRDM9* zinc finger domain of 64 primates from 18 species. Their work not only confirmed rapid evolution of PRDM9 throughout primate lineages, but also identified novel zinc finger variants in Neanderthal and Denisovan genomes, presumably resulting in distinct recombination landscapes for each species.

3.1.2 Structural Variation

Structural variants (SVs) represent a major source of genetic variation and disease in the human genome. These include deletions, duplications, insertions, inversions, and translocations. A major subclass of SVs is copy number variants (CNVs), which alter gene dosage and are particularly significant due to their enrichment at loci underlying

neurodevelopmental disorders, immune traits, and variability in drug response (Conrad et al., 2010; Ho et al., 2020; Redon et al., 2006; Weischenfeldt et al., 2013).

Multiple DNA repair and replication processes can generate SVs, each with distinct breakpoint signatures. Non-homologous end joining, for example, repairs DSBs by directly ligating broken DNA ends. This is often error-prone and can join ends from different breaks, leading to translocations, inversions, and deletions, usually with little to no homology at the junctions and potentially small insertions of nucleotides (Lieber, 2010). A related pathway, termed microhomology-mediated end joining, uses 5-25 bp of microhomologous sequences flanking the break site to align the ends before joining. This results in SVs whose breakpoints can be distinguished by these short microhomologies and the frequent deletion of intervening sequence (McVey & Lee, 2008; Sfeir & Symington, 2015).

In contrast to the non-homologous mechanisms discussed above, SVs can also be generated through homology-based mechanisms. This includes fork stalling and template switching and the related microhomology-mediated break-induced replication. These are replication-based models where a stalled DNA replication fork disengages and restarts DNA synthesis using a nearby microhomology-containing template, often on a different chromosome. This process can result in complex genomic rearrangements with breakpoints that have short stretches of microhomology (typically 2-15 bp) (Lee et al., 2007). This category of homology-based mechanisms also includes non-allelic homologous recombination, which will be a large focus of this thesis.

3.1.3 Non-Allelic Homologous Recombination

Non-allelic homologous recombination (NAHR) occurs when recombination happens between two homologous regions that are not allelic but share high sequence similarity, known as segmental duplications or low-copy repeats (LCRs). During meiosis, these highly identical repeats on homologous chromosomes, or even on the same chromosome, can misalign. When these misaligned repeats are used for DSB repair, it results in NAHR. If the repeats are arranged in direct orientation on the same chromosome, NAHR can lead to deletions or duplications. If the repeats are inverted, NAHR can result in an inversion. NAHR can also produce translocations if the repeats are located on non-homologous chromosomes (Liu et al., 2012; Stankiewicz & Lupski, 2002). Mobile elements, such as the short interspersed element Alu and the long interspersed element-1 LINE-1, are also capable of causing NAHR (Burssed et al., 2022), given the homologous sequences share at least 96% sequence identity (Startek et al., 2015). The three types of NAHR are demonstrated in Figure 3.1.

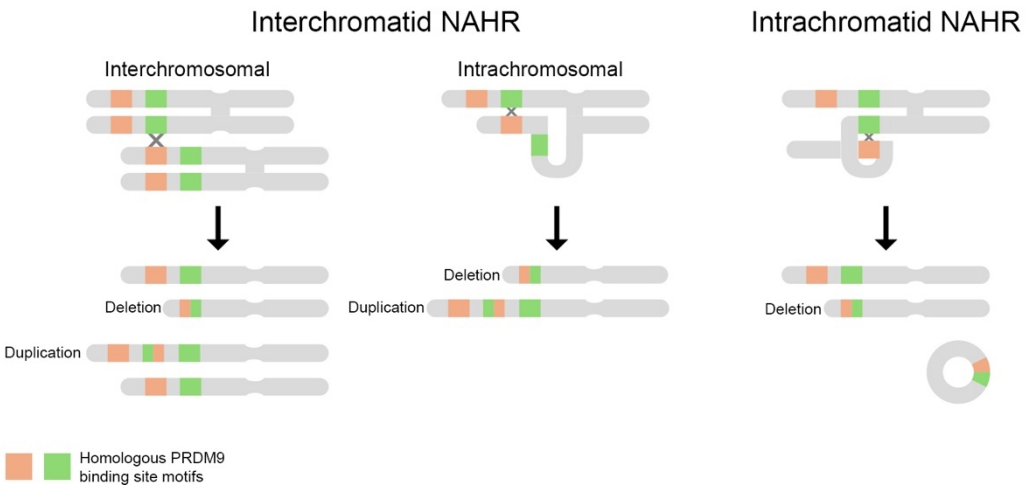


Figure 3.1, Mechanisms of non-allelic homologous recombination. Misalignment of homologous chromosomes or sister chromatids can lead to three types of recombination events. Interchromosomal NAHR and intrachromosomal NAHR occur between homologous chromosomes and between sister chromatids, respectively, and both produce a deletion and duplication. Intrachromatid NAHR occurs within the same chromatid, and results in a single deletion event, with the looped out sequence being lost.

NAHR occupies a central role in recurrent CNVs and can be responsible for a plethora of genomic disorders. These disorders typically map to genomic regions flanked by LCRs, which predispose the locus to recurrent rearrangements due to misalignment during meiosis. Examples of this include Williams-Beuren syndrome, which results from a recurrent 1.5 Mb deletion at 7q11.23, mediated by NAHR between LCR blocks (Osborne et al., 2001). Similarly, Charcot-Marie-Tooth disease type 1A is caused by reciprocal duplication of a 1.4 Mb region on chromosome 17 (Lupski et al., 2010). A reciprocal deletion at the same location causes hereditary neuropathy with liability to pressure palsies (Carvalho & Lupski, 2016). Another well-studied case is the 22q11.2 deletion syndrome (DiGeorge/velocardiofacial syndrome), where NAHR between multiple LCRs on 22q11 drives recurrent deletions of variable size (McDonald-McGinn et al., 2015; Shaikh et al., 2000). These recurrent CNVs often occur with relatively high mutation rates compared to point mutations, and they can recur in unrelated individuals due to the architectural predisposition of the locus (Carvalho & Lupski, 2016).

3.1.3.1 PRDM9 Binding Site Alterations

PRDM9 determines the genomic locations at which meiotic DSBs are formed, and NAHR is initiated when DSB repair engages homologous but non-allelic templates, typically within segmental duplications. Thus, the altering of DSB positioning has the potential to alter the likelihood of NAHR. Because NAHR is a recombination-driven process, any factor that repositions recombination initiation sites could, in principle, modify susceptibility to recurrent structural variation.

PRDM9 binding motifs are rapidly evolving features of the genome. As described above, hotspot erosion leads to the progressive loss of PRDM9 motifs through biased gene

conversion (Auton et al., 2012). Furthermore, population-level variation in PRDM9 zinc finger alleles alters hotspot usage within humans (Berg et al., 2010; Pratto et al., 2014). Together, these findings show that the binding motifs recognised by PRDM9 are subject to continual change.

Meiotic DSBs occurring within or near segmental duplications can promote NAHR-mediated rearrangements. CNV breakpoints cluster within recombination-active regions (Carvalho & Lupski, 2016; Lupski et al., 2010), and DSB induction within duplicated sequences has been shown to increase the probability of non-allelic exchange (Liu et al., 2012; Startek et al., 2015). These observations collectively suggest that the positioning of PRDM9-directed DSBs can influence the probability that homologous repair engages allelic versus non-allelic templates.

Despite these findings, a gap remains in the understanding of how alterations to PRDM9 binding site motifs affects individual susceptibility to NAHR. As PRDM9's major role involves specifying DSB locations, and NAHR is known to arise from misrepair of meiotic DSBs within segmental duplications, it is possible that naturally occurring polymorphisms that modify PRDM9 binding motifs also alter individual susceptibility to NAHR-mediated CNV formation. This thesis proposes and tests a theory that SNPs which alter PRDM9 binding affinity create 'risk diplotypes' that predispose to NAHR-mediated structural variation, and that these SNPs are under purifying selection in human populations.

A SNP within a PRDM9 binding motif could potentially disrupt an existing site (a 'loss'), reducing or abolishing PRDM9's ability to bind and initiate a DSB at that site. Alternatively, a base change could create a novel DNA sequence that now constitutes a high-affinity

binding motif for an individual's specific *PRDM9* allele, potentially creating a *de novo* recombination site (a 'gain'). Both types of alterations could reconfigure the recombination landscape in an individual's germline, creating new pairs of homologous sequences (risk diplotypes) that could mispair during meiosis and create a *de novo* CNV event in their offspring.

Importantly, NAHR would require that a binding-site-modifying SNP occur in the vicinity of at least one other PRDM9 motif. This could be another existing PRDM9 binding site motif, or another gain or loss. Thus, different diplotype configurations become possible. Two independent loss variants in *trans* (a *trans*-heterozygote loss-loss diplotype) may eliminate binding at two nearby pairs of motifs, directing the recombination machinery to use non-allelic templates. Similarly, two independent gain variants in *trans* (gain-gain diplotype) could create novel motifs that recombine with one another. A *cis* gain-loss diplotype could also be possible, where one SNP creates a new binding motif while another disrupts a pre-existing one, thereby allowing recombination between the gained binding site and the nearby existing homologue of the lost binding site. These configurations, displayed in Table 3.1 along with other possible combinations, provide a mechanistic framework for how common genetic polymorphisms could influence CNV formation. It should be noted that inversions are also a potential outcome of NAHR, though this thesis will focus exclusively on deletions and duplications.

Table 3.1, NAHR risk diplotype configurations.

		NAHR risk	
Configuration	Diplotype	Inter	Intra
Non-PRDM9 homozygote	-0---0- -0---0-	Low	Low
Non-PRDM9 bizygote	-0---0- -0---1-	Low	Low
Non-PRDM9 bizygote	-0---1- -0---0-	Low	Low
Mixed homozygote	-0---1- -0---1-	Low	Low
Non-PRDM9 bizygote	-0---0- -1---0-	Low	Low
Cis-heterozygote	**-0---0-** **-1---1-**	Low	High
Trans-heterozygote	**-0---1-** **-1---0-**	High	Low
PRDM9 bizygote	**-0---1-** **-1---1-**	Medium	Medium
Non-PRDM9 bizygote	-1---0- -0---0-	Low	Low
Trans-heterozygote	**-1---0-** **-0---1-**	High	Low
Cis-heterozygote	**-1---1-** **-0---0-**	Low	High
PRDM9 bizygote	**-1---1-** **-0---1-**	Medium	Medium
Mixed homozygote	-1---0- -1---0-	Low	Low
PRDM9 bizygote	**-1---0-** **-1---1-**	Medium	Medium
PRDM9 bizygote	**-1---1-** **-1---0-**	Medium	Medium
PRDM9 homozygote	**-1---1-** **-1---1-**	Med-low	Med-low

'1' represents a PRDM9 binding site motif and '0' represents the absence of a PRDM9 binding site motif. Linked alleles are joined by dashed lines. The interchromosomal and intrachromosomal NAHR risk is estimated for each combination based on the PRDM9 binding possibilities for that diplotype.

Loss-causing SNPs are expected to be more strongly associated with NAHR than gain-causing SNPs. This is based on the cellular requirements for a functional recombination hotspot, as a PRDM9 binding motif in the DNA sequence is not sufficient to create a DSB on its own. In addition to restricted sequence motifs, PRDM9 binding sites are also characterised by specific histone modifications, such as H3K4me3 and H3K36me3 deposited by PRDM9 itself (Borde & de Massy, 2013; Lam & Keeney, 2015a). These marks create open chromatin and recruit DSB machinery, particularly SPO11 (Lange et al., 2016). Therefore, a loss-causing SNP has a more immediate impact than a gain-causing SNP as it disrupts a motif within a locus that is already epigenetically primed for recombination. A loss-loss diplotype would further increase this probability by fully disabling the allelic homologous repair pathway for both motifs.

In contrast, gain-causing SNPs are less likely to have functional consequences. While they can create new motifs, these motifs must also exist within a chromatin state that is accessible to PRDM9 and permissive for recombination initiation. The genomic region might lack the necessary cofactors and chromatin remodellers required to open the locus and establish the nucleosome-depleted region required for DSB formation (Brick et al., 2012). Consequently, the penetrance of gain variants as NAHR risk factors is expected to be lower than that of loss variants.

3.1.4 Research Aims

This thesis chapter seeks to identify and characterise SNPs that alter PRDM9 binding site motifs. We hypothesise that such SNPs are not evolutionarily neutral but are subject to purifying selection due to their potential to predispose to NAHR events. Furthermore, we posit that these gain- and loss-causing SNPs are enriched at the breakpoints of CNVs and

can help explain a fraction of *de novo* genomic rearrangements. To explore this, we first identify all potential gain- and loss-causing SNPs in the HGDP+1000G dataset. We conduct broad analyses to test whether CNV breakpoints are significantly enriched for these gain- and loss-causing SNPs, as well as population genetic analyses to investigate the evolutionary pressures acting on these variants. We then analyse parent-offspring trios, identify *de novo* CNVs in the offspring and examine the breakpoints in the parental genomes to search for heterozygous SNPs that could create or disrupt PRDM9 binding motifs.

3.2 Materials and Methods

3.2.1 Mapping Gains and Losses

A comprehensive catalogue of potential PRDM9 binding sites was generated using a modified GRCh38 reference sequence in which the reference allele at each SNP position in the HGDP+1000G dataset was replaced with the inferred ancestral allele. The ancestral states were obtained using the Ensembl Variant Effect Predictor (McLaren et al., 2016). These potential PRDM9 binding sites were identified using the computational tool FIMO (Find Individual Motif Occurrences), which is part of the MEME suite (Grant et al., 2011). FIMO scans DNA sequences to identify statistically significant matches to a library of known sequence motifs, which are represented as position weight matrices (PWMs). A PWM captures the probability of each nucleotide at each position within a binding site, providing a quantitative model of the binding preference. Our analysis incorporated a set of 20 PRDM9 binding motifs to capture the known diversity and specificity of PRDM9 recognition. 17 motifs were obtained from Altemose et al. (2017), with three additional motifs from Patel et al. (2016). A significance threshold of $p < 1 \times 10^{-4}$ was applied for all FIMO scans. This initial scan of the reference genome provided a baseline set of all potential PRDM9 binding sites, against which all subsequent genetic variation was compared.

Customised FASTA files were then generated for each set of SNPs of interest. Each FASTA file was created by introducing the alternate alleles of the SNPs into the reference genome sequence, thereby simulating an individual's haplotype at those positions. FIMO was then rerun on these modified FASTA files using the same 20 PRDM9 motifs as input. Gains and losses could then be defined by comparing the results of motif scanning between the modified haplotype and the reference genome. A gain was defined as the presence of a

PRDM9 motif in the modified FASTA file without any overlapping motif of the same type and strand in the reference genome. This indicates that a SNP has introduced a new motif that was not previously present in the reference sequence. A loss was defined as a PRDM9 motif in the reference genome that does not overlap with a motif of the same type and strand in the SNP-modified sequence. This indicates that the SNP has disrupted a motif, potentially abolishing PRDM9 binding at that site.

3.2.2 Allele Frequencies

European allele frequencies for all potential gain- and loss-causing SNPs were extracted from the HGDP+1000G dataset, with each SNP's frequency compared to that of its nearest neighbouring SNP within the same population. This comparative approach used the closest SNP as a control, which served as a proxy for the expected allele frequency distribution in the local genomic context, with the assumption that nearby SNPs experience similar evolutionary pressures except for the specific effects of the gain/loss alteration. One-sided paired Wilcoxon signed-rank tests were conducted on each chromosome as well as for the entire genome. The primary hypothesis tested was whether gain- and loss-causing SNPs exhibit significantly lower allele frequencies than their nearest neighbouring SNPs, which would suggest purifying selection against these variants. This analysis was repeated only considering potential gain- and loss-causing SNPs within 1 kb of a DSB hotspot. These DSB locations were obtained from Pratto et al. (2014), who generated comprehensive maps of meiotic recombination initiation in five males with different PRDM9 alleles.

A broader analysis was then conducted where all potential loss-causing SNPs were separated into two categories based on their proximity to DSB hotspots. The first category included all potential loss-causing SNPs located within 1 kb of a DSB hotspot, and all remaining loss-

causing SNPs located further away were assigned to the second category. An equal number of near-hotspot and far-from-hotspot SNPs were randomly selected to ensure balanced comparison. Allele frequency distributions were compared between the two SNP groups using the Wilcoxon rank sum test, which was implemented as a one-tailed test to specifically test the hypothesis that near-hotspot SNPs would show lower allele frequencies.

3.2.3 NAHR Risk Diplotypes in Human Populations

The identification of potential PRDM9 binding site alterations was performed according to section 3.2.1. All biallelic SNPs from the HGDP+1000G dataset were used as SNPs of interest to identify potential gains and losses. To focus on alterations with potential to mediate NAHR, a proximity filter was applied where only potential gains or losses located within 200 kb (but at least 10 bp apart) from another existing potential PRDM9 binding site motif of the same type and strand were retained. Nearby motif pairs were then categorised into three risk diplotypes: gain-loss (one gain and one loss with same motif type and strand within 200 kb), gain-gain (two gains), or loss-loss (two losses).

3.2.4 CNVs Explainable by Risk Diplotypes

CNV calls from the 1000 Genomes Project were obtained from https://registry.opendata.aws/ilmn-dragen-1kgp. To investigate the relationship between CNVs and risk diplotypes, the regions flanking these CNVs (± 100 bp) were analysed for the presence of gain- or loss-causing SNPs. Each CNV was classified into one of five categories based on the motifs at its breakpoints: Gain-Loss (gain on one end, loss on the other), Gain-Gain (gain on both ends), Loss-Loss (loss on both ends), Gain (a gain on one end and an existing PRDM9 binding motif on the other), Loss (a loss on one end and an

existing PRDM9 binding motif on the other), or None. Motif and strand concordance were required for classification in all cases. When multiple categories were present, a hierarchy was used: Gain-Loss was prioritised, followed by Gain-Gain or Loss-Loss (selected by motif p-value strength), then Gain or Loss (similarly chosen by p-value).

1,000 sets of random genomic regions were then generated that matched the CNV size distribution. Each control set was subjected to the same motif analysis and classification as described above, and was then compared to the classification of the CNV set.

3.2.5 Microdeletion and Microduplication Syndromes

A comprehensive database of CNV breakpoints involved in microdeletion and microduplication syndromes (MMS) was obtained from Wetzel & Darbro (2022), comprising 192 recurrent and non-recurrent genomic disorders. All CNV coordinates were converted to GRCh38 using the UCSC Genome Browser LiftOver tool (Hinrichs et al., 2006).

For each CNV, all potential PRDM9 binding site motifs located within ± 100 bp of each breakpoint were extracted. To establish a null distribution, 1,000 sets of random PRDM9 binding site motifs (matched in number to the CNV-associated set) from the human reference genome were generated for comparison. The number of loss-causing SNPs from the HGDP+1000G dataset occurring within PRDM9 binding sites near CNV breakpoints was then compared against this random background distribution.

Given that many PRDM9 binding motifs throughout the genome may not be functionally active, the analysis was then refined by limiting the binding site motifs to those overlapping

with DSB hotspots. These DSB locations were obtained from Pratto et al. (2014), who generated comprehensive maps of meiotic recombination initiation in five males with different *PRDM9* alleles. This filtered analysis was performed in two ways: 1) by restricting the random binding site motifs to those overlapping with DSB hotspots while retaining all CNV-associated binding sites, and 2) by requiring that both the random binding sites and the CNV-associated binding sites overlap DSB hotspots.

3.2.6 Trio Analysis

This analysis focused exclusively on deletions and duplications from the 1000G Trio CNVs (section 2). *De novo* CNVs were defined as variants present in a child but absent in both parents, with no overlapping CNV of matching type detected in either parental genome. To minimise false positives and exclude potentially inherited or recurrent variants, the dataset was further filtered by removing any CNV observed more than twice across the 300 individuals in the 1000G Trio cohort.

For each qualifying CNV, the region flanking (± 400 bp) each breakpoint was examined for heterozygous SNPs in the parental genomes. Only CNVs with a maximum of one heterozygous SNP per breakpoint in either parent were retained for subsequent analysis. This threshold was implemented because of the unphased nature of the VCF files, as multiple heterozygous SNPs at a breakpoint could reside on either the same or opposing haplotypes, which would complicate the reconstruction of sequences. For these breakpoints, the standard gain/loss analysis (as defined in section 3.2.1) was performed by comparing the parental haplotype containing the alternate allele of the heterozygous SNP to the human reference genome. However, in the case where a homozygous SNP was also present within the flanking region, the standard comparison to the human reference genome was invalid, as the reference

allele would not be present in the parent. To address this, the sequence containing both the homozygous SNP and the heterozygous SNP was compared to a reference sequence that included the homozygous SNP but not the heterozygous one. This comparison isolates the specific effect of the heterozygous SNP on the PRDM9 binding motif within the genetic context of that specific parent and provides a more accurate assessment of its potential to create or disrupt a binding site. To ensure biological relevance, gains and losses were only considered when they were paired with another PRDM9 binding motif (either existing, gained, or lost) within the opposing breakpoint region.

All CNV coordinates, along with their associated PRDM9 binding site alterations (gains, losses, and existing motifs), were mapped for visualisation. To validate the CNV calls and assess inheritance patterns, read depth coverage from the original BAM files was extracted and analysed for both parents and the child. This enabled direct visualisation of the putative *de novo* duplication or deletion events in the offspring relative to the parental genomes.

3.3 Results

3.3.1 Allele Frequencies

A comparative analysis between the European allele frequency of all potential gain- and loss-causing SNPs and their nearest SNP revealed significantly lower allele frequencies in loss-causing SNPs compared to their neighbouring control SNPs (Figure 3.2). Despite considerable heterogeneity among individual chromosomes (Table 3.2), the consistent genome-wide signal suggests that while the strength of selection may vary regionally, the overall trend supports the hypothesis that SNPs capable of altering PRDM9 binding sites may be deleterious and subject to purifying selection.

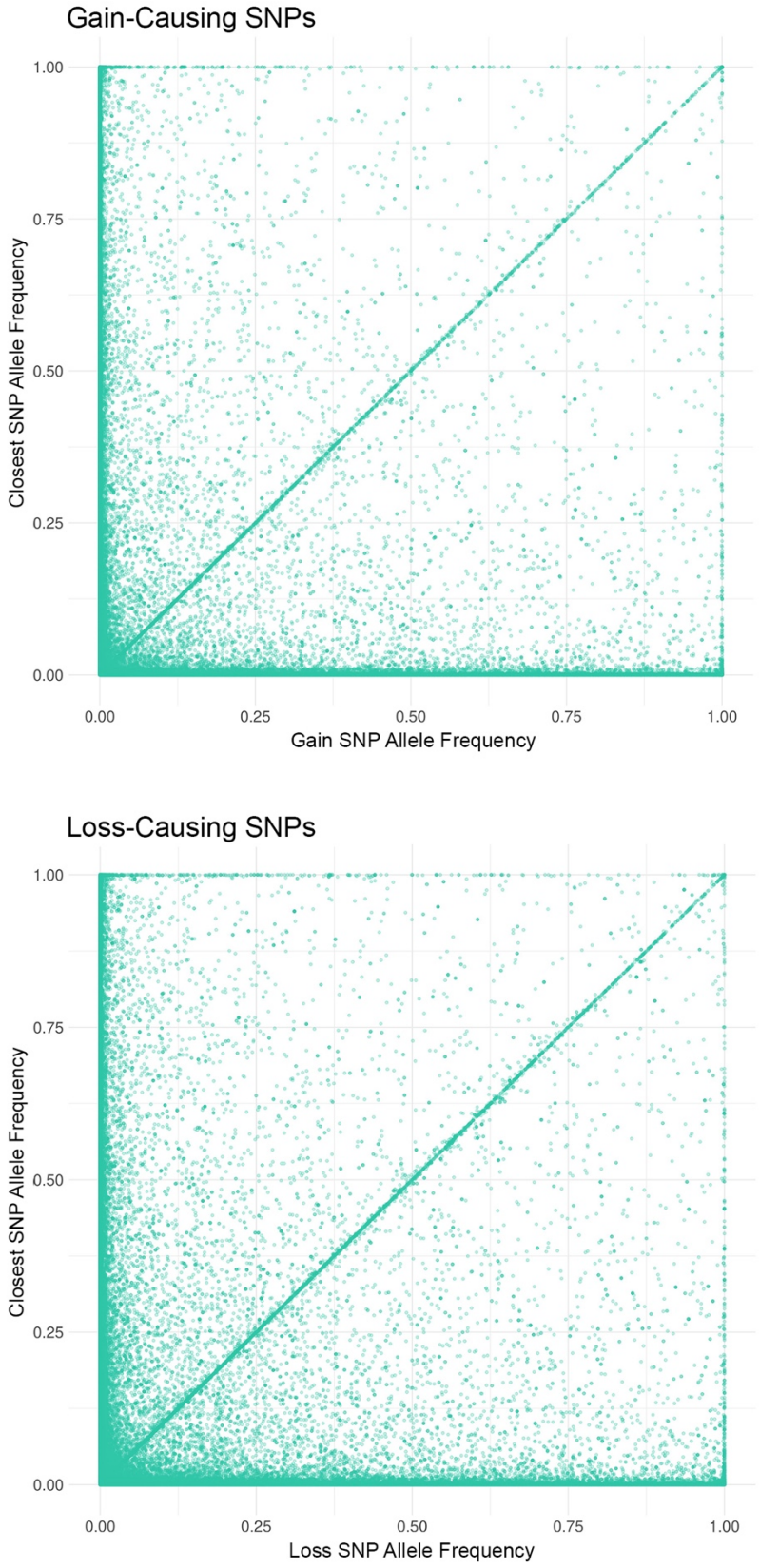


Figure 3.2, Gain- and loss-causing SNP AF comparison. Each gain/loss-causing SNP is represented and their European AF is compared to that of their closest SNP.

Table 3.2, Wilcoxon signed-rank one-sided tests for gain/loss-causing SNP allele frequencies.

	All Gains	*All Losses*	*DSB Hotspot Gains*	*DSB Hotspot Losses*
Chr1	0.18	4.20×10^{-4}	0.96	0.30
Chr2	0.85	6.83×10^{-9}	0.86	0.12
Chr3	0.29	2.08×10^{-3}	0.32	0.28
Chr4	0.16	0.44	0.85	0.22
Chr5	0.17	0.47	0.99	0.44
Chr6	0.50	2.44×10^{-3}	0.75	0.92
Chr7	0.54	7.74×10^{-2}	0.76	0.33
Chr8	0.34	1.59×10^{-2}	0.11	5.33×10^{-2}
Chr9	0.18	0.18	0.15	0.35
Chr10	0.12	0.05	0.99	5.03×10^{-3}
Chr11	0.56	1.03×10^{-5}	0.97	9.46×10^{-2}
Chr12	0.10	0.22	0.27	1.71×10^{-4}
Chr13	0.55	0.33	6.23×10^{-2}	0.18
Chr14	0.14	2.62×10^{-2}	0.95	5.88×10^{-2}
Chr15	0.71	0.15	1.58×10^{-2}	0.43
Chr16	0.84	0.32	0.99	8.72×10^{-2}
Chr17	0.01	0.86	0.20	0.76
Chr18	0.75	0.81	0.96	0.39
Chr19	1.29×10^{-4}	0.97	0.47	0.80
Chr20	0.40	2.94×10^{-4}	0.38	0.52
Chr21	0.94	6.13×10^{-2}	0.30	0.25
Chr22	0.88	0.41	0.88	0.33
Genome	**0.02**	**5.29×10^{-14}**	**0.99**	**3.69×10^{-4}**

The p-values obtained from Wilcoxon signed-rank tests evaluate whether the allele frequency of gain- and loss-causing SNPs are significantly lower than that of their nearest neighbouring SNP. All gains and losses were tested (left) as well as those within 1 kb of a DSB hotspot (right). The colours represent those found to be significant (orange: $p < 0.05$, blue: $p < 4.3 \times 10^{-4}$, representing a conservative Bonferroni-corrected threshold accounting for 92 tests).

When restricting analysis to potential gain- and loss-causing SNPs within 1 kb of a DSB hotspot, loss-causing SNPs maintained significant allele frequency depletion ($p = 3.69 \times 10^{-4}$), which reinforces their potential functional impact on PRDM9 binding and NAHR susceptibility. However, gain-causing SNPs near hotspots showed no significant frequency difference from their neighbouring SNPs ($p = 0.99$). This aligns with the model where loss-causing variants have more immediate consequences for recombination, while gain-causing variants may fail to create new functional sites without the appropriate chromatin context or cofactors for recombination initiation. The hotspot-restricted analysis therefore may reveal the more specific signature of selection acting solely on PRDM9's DSB-determining function, whereas the genome-wide analysis may capture selection pressures from all PRDM9's activities. The reduced statistical significance for both SNP categories in the DSB hotspot analysis may also reflect the much smaller SNP set size following the DSB hotspot proximity filtering, reducing the power of this analysis.

A comparative analysis was then conducted between the allele frequencies of 286,113 loss-causing SNPs within 1 kb of a DSB hotspot and a matching number of loss-causing SNPs further away from DSB hotspots (Figure 3.3). The analysis revealed no detectable difference in allele frequency distributions between the two groups of loss-causing SNPs (Wilcoxon rank sum test, $p = 0.9815$). This indicates that while loss-causing SNPs show genome-wide evidence of negative selection, this selective pressure does not appear intensified near recombination hotspots. Purifying selection appears to act against PRDM9 binding site disruption consistently across the genome, without hotspot-specific effects detectable within the region analysed.

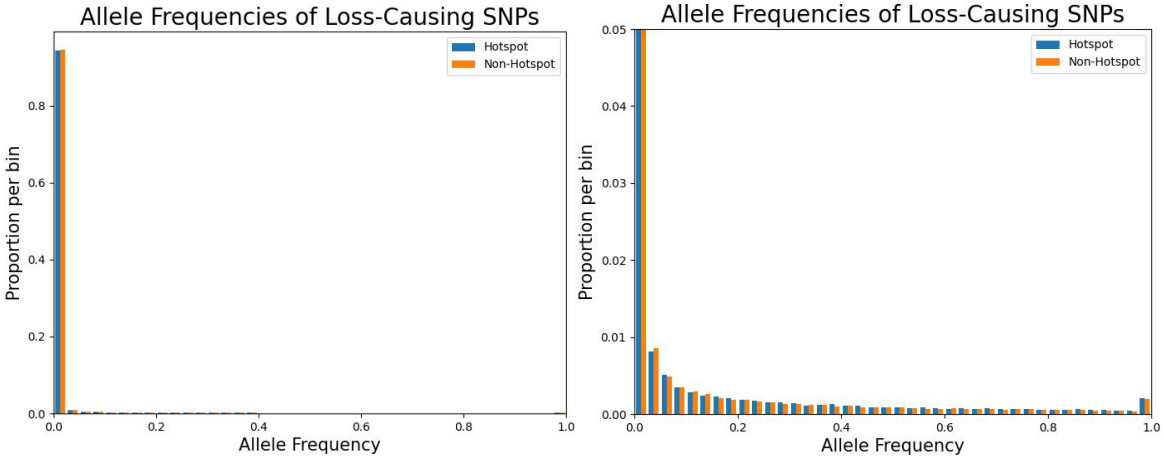


Figure 3.3, Allele frequency comparison of loss-causing SNPs based on their distance to the closest DSB hotspot. These bar plots show the proportion of allele frequencies of 286,113 loss-causing SNPs within 1 kb of a DSB hotspot ('Hotspot') compared to 286,113 loss-causing SNPs further away from DSB hotspots ('Non-Hotspot'), grouped into bins of 2.5% allele frequency intervals. For each bin, bars represent the proportion of SNPs within each category that fall into that allele frequency range. The left panel displays the full distribution, while the right panel shows a zoomed-in view with the y-axis limited to a maximum of 5% to highlight differences at lower proportions.

3.3.2 NAHR Risk Diplotypes in Human Populations

Analysis of the HGDP+1000G variants revealed substantial variation in potential PRDM9 binding sites across human populations, with 4,574,110 potential gains and 10,041,450 potential losses identified. The ~50% fewer potential gains compared to potential losses suggests that SNPs are more likely to disrupt existing PRDM9 binding motifs than create new ones, possibly reflecting the more stringent sequence requirements for *de novo* motif formation compared to simple disruption of existing sites. When examining the relationships between these alterations, an abundance of potential NAHR risk diplotypes were found: 553 million gain-loss combinations, 81 million gain-gain combinations, and 309 million loss-loss combinations were identified within the 200 kb window (Figure 3.4).

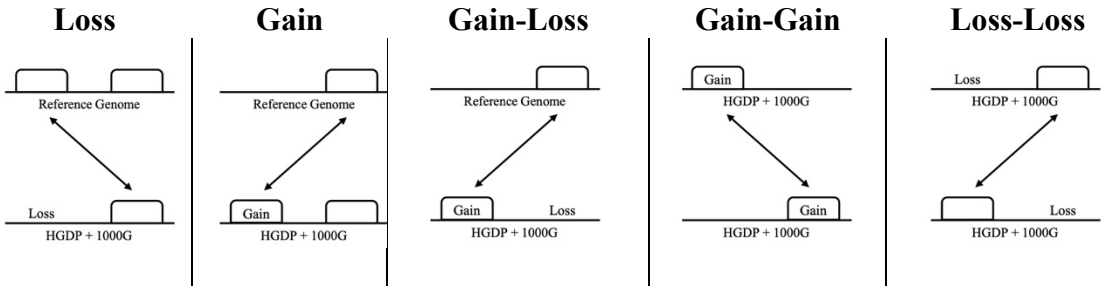


Figure 3.4, NAHR risk diplotypes in the HGDP+1000G dataset. Illustrated here are the 5 risk diplotypes and the number of matches found when considering all biallelic SNPs from the HGDP+1000G dataset in comparison to the ancestral alleles.

The large number of matches arise partially because the number of possible combinations increases rapidly as additional gains and losses are identified. Any potential gain or loss can theoretically pair with any other compatible alteration within the 200 kb window, meaning that even moderate numbers of gains and losses can generate extremely large numbers of potential combinations. To illustrate the genomic distribution of these potential NAHR risk diplotypes, the densities of gain-gain, gain-loss, and loss-loss combinations across each chromosome are shown in Appendix A.

Additionally, the enormous total numbers of gains and losses must be interpreted cautiously. While they suggest widespread sequence-level potential for NAHR risk diplotypes, the majority likely represent non-functional motifs. This analysis only considers sequence characteristics without accounting for crucial factors like chromatin accessibility, histone modifications, or the cellular and developmental context of PRDM9 expression. Many of the predicted sites likely lack the epigenetic environment required for actual PRDM9 binding, while others may represent marginal affinity sites that rarely bind the protein *in vivo*. Successful recombination also requires extensive sequence homology in the regions flanking the motifs to facilitate strand invasion and repair (Startek et al., 2015). A

comprehensive analysis to identify all pairs of motifs with sufficient surrounding homology for NAHR would be computationally prohibitive at this scale, as it would involve pairwise alignment for billions of potential combinations. These results therefore serve primarily to identify genomic regions where sequence variation could theoretically alter PRDM9 binding site motifs.

3.3.3 CNVs Explainable by Risk Diplotypes

To further investigate the relationship between CNVs and risk diplotypes, the regions flanking all CNVs (± 100 bp) in the 1000G Trio dataset were analysed for the presence of gain- or loss-causing SNPs. Each CNV was classified into one of five categories based on the motifs at its breakpoints: Gain-Loss (gain on one end, loss on the other), Gain-Gain (gain on both ends), Loss-Loss (loss on both ends), Gain (a gain on one end and an existing PRDM9 binding motif on the other), Loss (a loss on one end and an existing PRDM9 binding motif on the other), or None.

To statistically analyse the counts of these combinations, 1,000 comparative random region sets matching the CNV lengths were generated and subjected to the same classification. Comparisons between observed CNVs and these random sets (Figure 3.5) revealed a consistent pattern across all gain/loss categories. Gain-Loss, Gain-Gain, Loss-Loss, as well as the single Gain and Loss categories, were all underrepresented at CNV breakpoints relative to the distributions observed in the random regions. Consequently, the None category was overrepresented among the CNVs. This pattern was observed for both deletions and duplications, indicating that CNV breakpoints in this dataset were less likely than expected by chance to be associated with SNP-driven gains or losses of PRDM9 binding site motifs within the ±100 bp breakpoint regions.

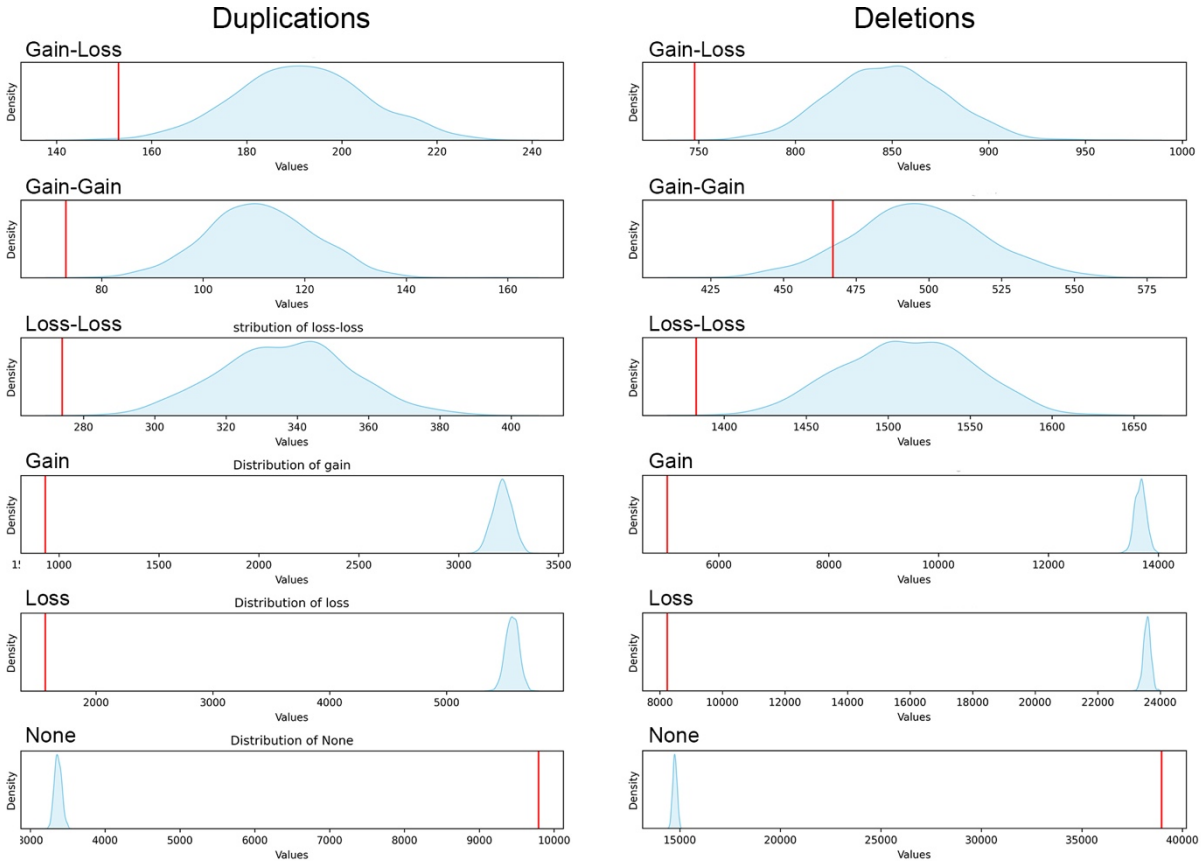


Figure 3.5, Gain/Loss Categories of CNVs vs Random Regions. The observed number of duplication (left) and deletion (right) breakpoint regions falling into each gain/loss category (red line) is compared to a null distribution from 1,000 matched random region sets (blue distribution).

3.3.4 Microdeletion and Microduplication Syndromes

A total of 5,896 potential PRDM9 binding site motifs were identified ± 100 bp of the breakpoints of 192 MMS CNVs catalogued by Wetzel & Darbro (2022). Of these, 905 motifs overlapped with loss-causing SNPs in the HGDP+1000G dataset. This was then compared to a null distribution generated from 1,000 random sets of 5,896 PRDM9 binding site motifs from the human reference genome (Figure 3.6). In every case, the random sets contained more loss-causing SNPs than the MMS-associated set. This consistent depletion suggests that PRDM9 motifs at MMS breakpoints may be subject to stronger functional constraints than motifs elsewhere in the genome. Motifs located at CNV breakpoints are more likely to

be functionally active in meiotic recombination, given their enrichment in regions associated with recurrent genomic instability. Selection may therefore act to reduce the frequency of such SNPs specifically at MMS-associated binding motifs to reduce the risk of NAHR-mediated rearrangements. By contrast, the randomly chosen PRDM9 motifs were taken from across the genome, the majority of which are unlikely to be in recombinationally active regions. These random loss-causing SNPs are therefore less likely to be functionally constrained and purged by selection.

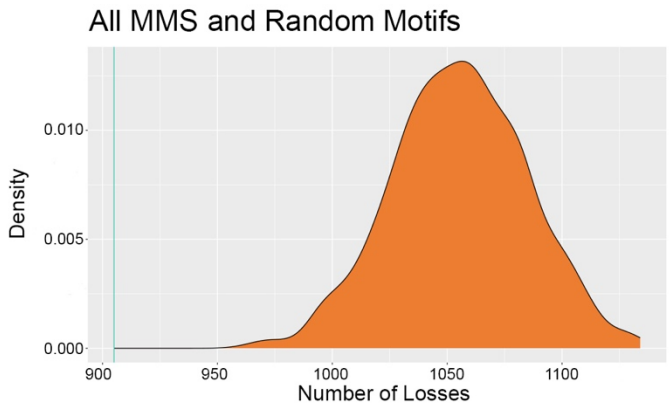


Figure 3.6, Enrichment of losses at MMS breakpoints. The observed number of loss-causing SNPs among the 5,896 PRDM9 motifs within ±100 bp of MMS breakpoints (green line) is compared to a null distribution generated from 1,000 random sets of 5,896 PRDM9 motifs (orange).

To account for this potential bias, two refined analyses were performed specifically focusing on DSB hotspot regions where PRDM9 binding is experimentally validated. In the first analysis comparing all MMS-associated motifs against hotspot-restricted random motifs, the same depletion pattern was observed, where all 1,000 control sets contained more loss-causing SNPs (Figure 3.7). This persistent observation suggests that even among potentially functional PRDM9 binding site motifs, those near MMS breakpoints still experience more selective constraints. The second analysis required both MMS and random motifs to reside in DSB hotspots. In this case, there was no significant difference between the number of loss-causing SNPs in the MMS set compared to the random regions ($p = 0.685$; Figure 3.8).

This is likely due to reduced statistical power from the smaller sample size, as the filtering retained only 199 motifs in the MMS set.

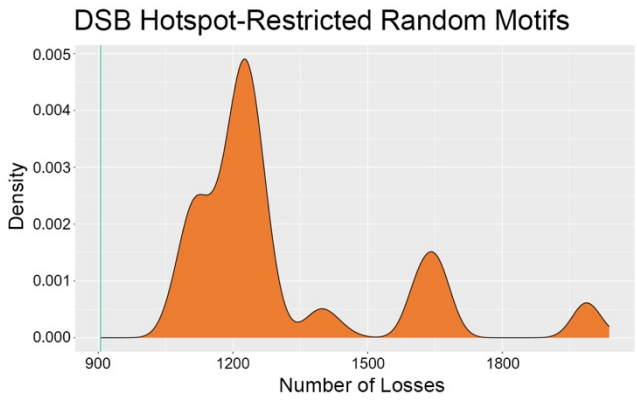


Figure 3.7, Enrichment of losses at MMS breakpoints compared to DSB-hotspot restricted random motifs. The observed number of loss-causing SNPs among the 5,896 PRDM9 motifs within ±100 bp of MMS breakpoints (green line) is compared to a null distribution generated from 1,000 random sets of 5,896 PRDM9 motifs within DSB hotspots (orange).

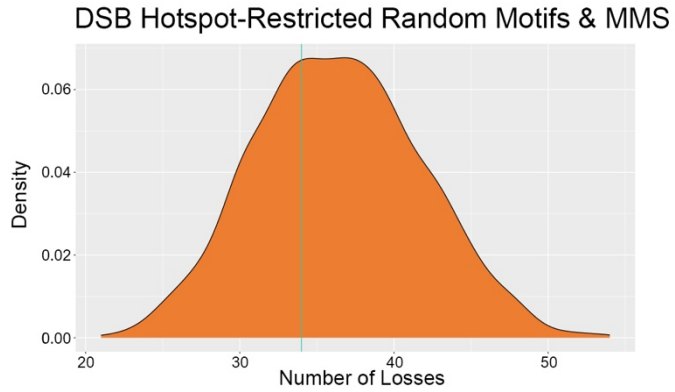


Figure 3.8, Enrichment of losses within DSB hotspots at MMS breakpoints. The observed number of loss-causing SNPs among the 199 PRDM9 motifs within ±100 bp of MMS breakpoints and overlapping a DSB hotspot (green line) is compared to a null distribution generated from 1,000 random sets of 199 PRDM9 motifs within DSB hotspots (orange).

3.3.5 Trio Analysis

Analysis of the 100 parent-child trios from the 1000G Trio dataset identified 1,869 *de novo* CNVs in the offspring, comprising 1,679 deletions, 155 duplications, and 35 inversions. Inversions were excluded from subsequent analyses, and the remaining CNVs were only

considered when at least one parent carried a heterozygous SNP at a breakpoint region (± 400 bp of the breakpoint). This was designed to observe scenarios in which a parental SNP could generate or disrupt a PRDM9 binding site motif, thereby enabling NAHR with an existing motif at the other breakpoint. Using this approach, 45 *de novo* CNV events met the criteria of having either a gain-causing or loss-causing SNP at a breakpoint, alongside at least one existing PRDM9 binding site motif at the opposing breakpoint. These results suggest that SNPs that are capable of altering PRDM9 binding motifs in parents could potentially participate in recombination events creating *de novo* CNVs in their offspring. However, it remains difficult to directly attribute causality to these SNPs without functional evidence supporting the SNP's role in recombination error.

Further investigation into the validity of these 45 candidate CNVs was performed by visually inspecting RD profiles in the BAM files for each trio. This revealed a major limitation of the dataset. For the majority of cases, the short-read sequencing coverage did not qualitatively support the presence of the reported CNV in the child. While a few deletions showed the expected drop in coverage in the offspring (which was absent in the parents) consistent with a true *de novo* deletion, most candidate CNVs showed no significant change in coverage at the region of interest. Representative examples of a well-supported deletion and a likely false-positive CNV call are shown in Figure 3.9. The remainder of the 45 *de novo* CNVs of interest are presented in Appendix A.

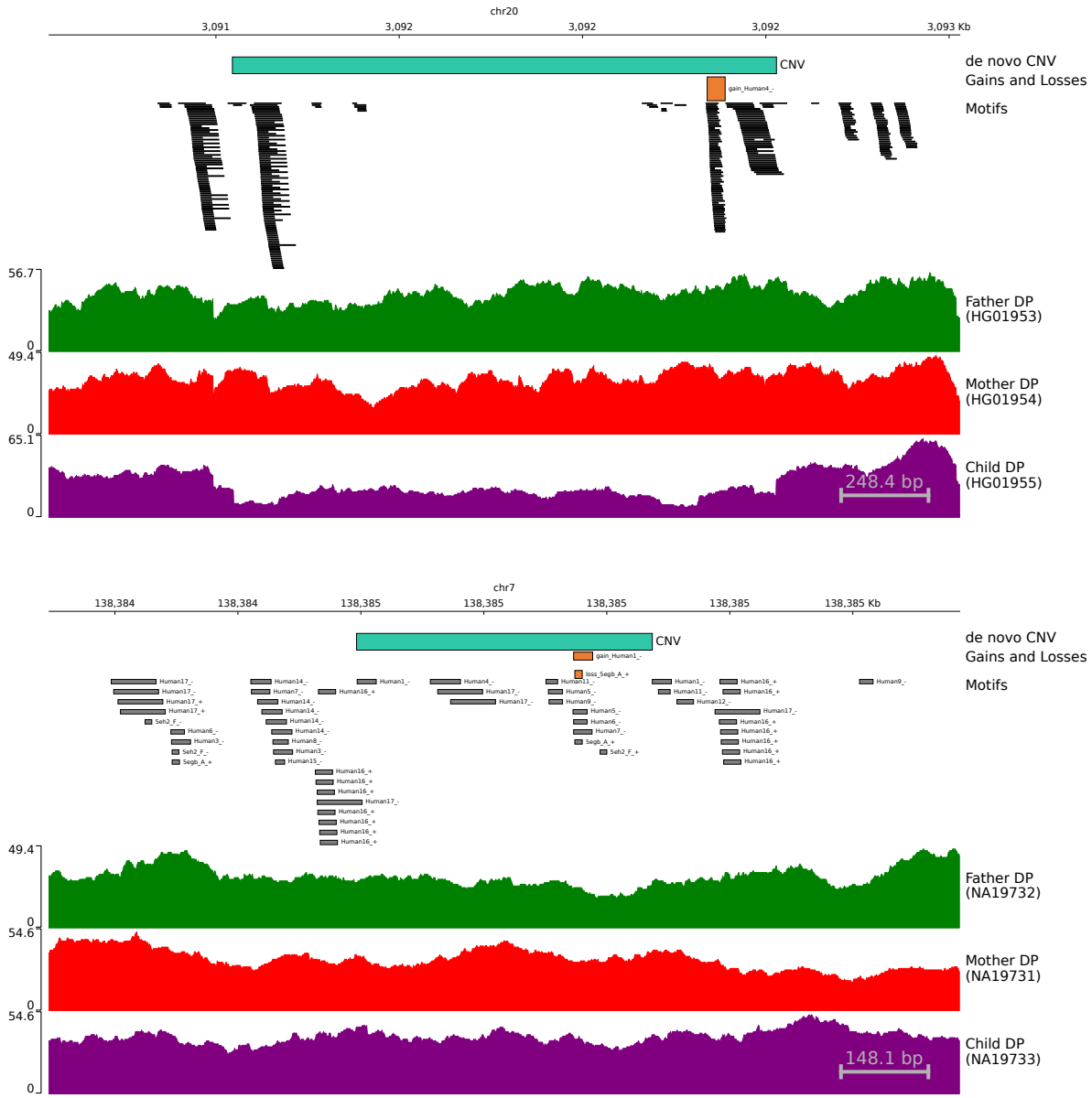


Figure 3.9, de novo deletion plots. Two candidate de novo deletions are displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed. The top plot illustrates a de novo deletion confirmed by a clear drop in child read depth. The bottom plot is an example of an unsupported candidate where qualitative assessment of read depth fails to validate the call.

The lack of RD evidence for many of these events highlights a challenge with CNV calling from short-read sequencing data. Previous studies have demonstrated that CNV detection is particularly unreliable when using short-read data, only allowing 10%-70% of CNVs to be detected, with up to 89% false positive rates (Gong et al., 2020; Sedlazeck et al., 2018). RD methods also generally have lower sensitivity for small CNVs (<10 kb) and can misinterpret

local fluctuations in coverage as structural variation. This calls for cautious interpretation of the results and suggests that future studies incorporating higher-resolution approaches, particularly long-read sequencing, will be necessary to accurately assess the role of such SNPs in NAHR events.

3.4 Discussion

A central hypothesis of this work is that SNPs and small indels which modify PRDM9 motifs can reshape the recombination landscape in ways that increase the risk of NAHR. The population-scale analysis revealed over 942 million pairs of risk configurations that are theoretically compatible with NAHR, though only a minority of the identified PRDM9 binding motifs would be functionally active. Nevertheless, the fact that such configurations exist across all chromosomes provides a foundation upon which SNP-mediated NAHR could act. In particular, the observation that loss-causing SNPs outnumber gains suggests that it is easier to disrupt a motif than to create a new one. This difference might also be partially explained by the continuous erosion of hotspots observed in human populations and the evolutionary pressure for new PRDM9 alleles (Auton et al., 2012; Schwartz et al., 2014).

The trio-based analysis demonstrates the potential of this model in real-world meiosis, with the identification of 45 *de novo* CNV events that coincide with parental SNPs altering PRDM9 motifs. However, the limited validation from read depth coverage highlights the difficulty of connecting individual variants to specific CNV events. SV callers such as DELLY, Manta, and LUMPY rely on combinations of read-pair information (discordant orientations or distances between paired reads), split-read mapping (where a single read aligns to two separate genomic locations), and local fluctuations in RD to infer breakpoints (X. Chen et al., 2016; Layer et al., 2014; Rausch et al., 2012). Each of these signals is compromised by the properties of short-read sequencing. The limited length of Illumina reads (~150 bp) means that they often don't span entire CNVs, and stochastic variance in coverage makes it difficult to distinguish genuine CNVs from sequencing noise. Repetitive regions can also cause ambiguous mapping. These technical challenges account for the high

false positive and false negative rates that persist in short-read based CNV discovery (Gong et al., 2020; Sedlazeck et al., 2018).

For the current analysis, apparent CNVs often represent artefacts of mapping or depth fluctuations. Although some deletions exhibited clear offspring-specific loss of coverage, most showed no discernible signal, which suggests that a substantial fraction of the CNVs in this dataset may be spurious. Considering that the majority of our *de novo* CNVs were relatively short (mean CNV length = 1.2 kb, median CNV length = 43 bp), it is highly likely that a large fraction of them are false positives caused by natural fluctuations in read depth due to mapping biases. This limitation highlights the need for higher-resolution technologies in future studies. Long-read sequencing platforms, such as those from Oxford Nanopore Technologies and Pacific Biosciences, are perfectly suited for this. These technologies can produce read lengths exceeding 10 kb, with ultra-long reads often exceeding 100 kb (Jain et al., 2018; Wenger et al., 2019). This allows individual reads to span entire CNVs and their breakpoints, which provides clear evidence for the variant's presence and enables precise determination of its boundaries. Long reads also preserve haplotype information, which would allow us to directly observe whether a candidate gain- or loss-causing SNP is physically located on the same haplotype as the CNV breakpoint. The current analysis was restricted by unphased VCFs, which necessitated conservative thresholds (e.g. excluding breakpoints with more than one heterozygous SNP). While this reduced false positives, it likely eliminated biologically interesting cases where multiple SNPs cooperatively alter PRDM9 binding.

All categories of risk diplotype configurations (loss-loss, gain-loss, gain-gain, as well as single gain and single loss categories) were underrepresented at CNV breakpoints relative to

random genomic regions. This pattern should be interpreted cautiously given the uncertainty surrounding CNV detection in short-read sequence data. As discussed above, many of these CNVs likely represent false positive calls, which would reduce the power to detect any biological enrichment around true breakpoints. Another important consideration is that NAHR represents only one of several mechanisms capable of generating CNVs. While NAHR plays a major role in recurrent genomic rearrangements mediated by large segmental duplications, many CNVs arise through alternative pathways such as non-homologous end joining, microhomology-mediated end joining, or replication-based mechanisms including fork stalling and template switching and microhomology-mediated break-induced replication (Lee et al., 2007; Lieber, 2010; McVey & Lee, 2008; Sfeir & Symington, 2015). Genome-wide analyses suggest that NAHR accounts for only a minority of SVs overall, estimated at ~13.5% of events, with the remainder attributed to non-homologous or replication-based mechanisms (Conrad et al., 2010; Redon et al., 2006). Consequently, even if gain- or loss-causing SNPs do influence NAHR susceptibility, their signal would be expected to be diluted in a set of CNVs that were generated through unrelated pathways. SVs arising through replication-based or non-homologous mechanisms would not necessarily be associated with PRDM9 binding motifs, and therefore would be less likely to exhibit gains or losses of motifs near their breakpoints. The overall depletion of gains and losses may therefore reflect the predominance of non-NAHR CNVs in the analysed dataset. Together, these considerations highlight the importance of distinguishing NAHR-driven CNVs from those produced by alternative mutational processes when evaluating the role of PRDM9 motif-modifying variants.

Genome-wide, loss-causing SNPs at biologically validated recombination sites exhibit significantly lower allele frequencies than their nearest neighbouring SNPs, suggesting that

these variants may not be evolutionarily neutral and could be subject to purifying selection. These observations raise the possibility that alterations to PRDM9 binding site motifs could have functional consequences in certain contexts, potentially by destabilising meiotic recombination and therefore increasing the risk of SVs. Loss-causing SNPs show stronger frequency differences than gain-causing SNPs, which aligns with our proposed model. The disruption of an active hotspot could have immediate and deleterious consequences, while creating a new motif would often be evolutionarily neutral unless it happens to land in a permissive chromatin context. These findings align with what is already known about the evolutionary consequences of CNVs themselves. Compared to SNPs, CNVs tend to have larger effect sizes, are more frequently associated with developmental disorders and disease phenotypes, and are under stronger purifying selection (Itsara et al., 2010; Sudmant et al., 2015; Zarrei et al., 2015). Recurrent CNVs, in particular, are disproportionately represented in neurodevelopmental disorders, congenital anomalies, and immune-related variation (Coe et al., 2019). The observation that SNPs disrupting PRDM9 motifs can increase the likelihood of NAHR suggests that the same selective pressures acting directly on CNVs may also extend indirectly to the SNPs that predispose them. The signature of purifying selection we observe on the SNPs themselves may therefore be connected to the negative selection known to act against the CNVs they potentially generate.

This chapter has established that common genetic variation has the potential to alter the meiotic recombination landscape and predispose to NAHR, and that these variants are selected against in human populations. The principles established here may extend beyond common polymorphisms within a population to larger-scale genomic differences between species. A particularly compelling context for this is the introgression of Neanderthal and Denisovan segments into the modern human genome. The ~500,000 years of evolutionary

divergence between modern humans and Neanderthals would have provided an opportunity for the accumulation of genetic variants, likely including many mutations that would be classified as gain- or loss-causing SNPs for the modern human PRDM9 protein. These differences in motif composition between archaic and modern human haplotypes could therefore increase the frequency of mismatched motif pairs, creating more opportunities for NAHR. To explore this, the following chapter will focus on inferring and characterising archaic segments in modern human populations. By identifying introgressed haplotypes, we can then test in the subsequent chapter whether they harbour an excess of motif-altering SNPs, whether their local recombination landscapes differ from non-introgressed regions, and whether they overlap recurrent CNV hotspots. This has the potential to reveal how ancient admixture reshaped not just the genetic variation present in humans but also the mutational processes that continue to generate structural diversity today.

4 The Genomic Legacy of Interbreeding between Humans and Neanderthals

4.1 Introduction

4.1.1 Archaic Genomes and their Modern Legacy

A major breakthrough in human paleogenomics was the first draft sequence of the Neanderthal genome, which was obtained from DNA isolated from fossil bones found in Vindija Cave, Croatia (Green et al., 2010). This was followed by the discovery of a distinct archaic lineage from sequencing and analysis of DNA from a single finger bone in Denisova Cave, Siberia, now named Denisovans (Reich et al., 2010). These achievements provided a new set of reference points against which modern human genomes could be compared to identify introgressed sequences.

Comparative analyses of human and Neanderthal sequence data revealed that the two subspecies share a common ancestor that lived approximately 550,000 to 765,000 years ago (Prüfer et al., 2014). Neanderthals subsequently adapted to the environmental pressures of Eurasia and evolved distinct morphological and physiological traits. Analysis of human and Neanderthal sequences using allele sharing statistics such as the ABBA-BABA test (discussed below) revealed that non-African modern humans share approximately 1-4% of their DNA with Neanderthals (Green et al., 2010; Sankararaman et al., 2014). This introgression is not uniform between non-African subpopulations; East Asian populations

carry up to 21% more Neanderthal sequence than Europeans (Vernot & Akey, 2014), which suggests possible multiple pulses of interbreeding or dilution in the West through later migrations. The distribution of Neanderthal DNA is also uneven across the modern genome, as it is absent from large regions of the X chromosome and genes highly expressed in the male germline, a pattern suggestive of hybrid sterility issues (known as the 'large X-effect) (Sankararaman et al., 2014).

Denisovans are more distantly related to modern humans than Neanderthals, as they diverged from a common ancestor with Neanderthals around 400,000 years ago (Prüfer et al., 2014). Their genetic legacy is also more geographically constrained than for Neanderthals. The highest levels of introgressed Denisovan ancestry (~4-6%) are found in Indigenous populations from Melanesia and Australia, with smaller amounts distributed across Southeast Asia and East Asia (Qin & Stoneking, 2015; Reich et al., 2010). Interestingly, South Asian and West Eurasian populations show minimal to no Denisovan ancestry. This pattern suggests that the interbreeding with Denisovans likely occurred in Southeast Asia after the divergence of the ancestors of East Asians and Melanesians, with the latter experiencing less dilution from later population movements.

4.1.2 Inferring Archaic Introgression

Identifying segments of DNA that introgressed from an archaic hominin tens of thousands of years ago is a computational challenge with multiple potential sources of error. Inferring introgression is not a direct observation but a statistical conclusion drawn from patterns of genetic variation, which, if not properly framed, may be sensitive to potential confounders. A primary challenge is distinguishing true introgression from incomplete lineage sorting (ILS) (Fang et al., 2024; Huang et al., 2025; Zhou et al., 2017). ILS occurs when genetic

variants are shared between species due to the persistence of ancestral polymorphism from a population's common ancestor, rather than from recent gene flow. For instance, a modern human and a Neanderthal might share an allele because their common ancestor carried multiple alleles, and the lineages coalesce before the population split. Disentangling this ancient shared history from recent interbreeding requires sophisticated statistical models that incorporate accurate demographic parameters, such as population split times and sizes, as an incorrect model can readily cause ILS to be misinterpreted as introgression, or vice versa (Zhou et al., 2017).

The quality and nature of the ancient DNA data representing the source of introgression introduces another layer of difficulty. Ancient DNA is notoriously degraded and susceptible to chemical damage, which results in characteristic patterns like cytosine-to-thymine misincorporations at the ends of sequence fragments (Dabney et al., 2013). While these patterns can be used to authenticate ancient sequences, if not properly accounted for in bioinformatic pipelines, they can also introduce systematic biases. Another issue known as reference bias arises because most genomic analyses align sequencing reads to the modern human reference genome. This reference is more phylogenetically similar to non-archaic modern human sequences, which can cause truly divergent archaic sequences to be underrepresented or misaligned, leading to an underestimation of introgression levels in the most divergent regions.

To overcome these challenges, a suite of complementary statistical methods has been developed. The initial evidence for gene flow was established through population genetic methods like the D-statistic (or ABBA-BABA test), which compares patterns of allele sharing between four populations (e.g. two modern human populations, an archaic genome,

and an outgroup) to detect a genome-wide signal of admixture (Durand et al., 2011; Green et al., 2010). While this is powerful for confirming the presence of admixture, it cannot identify the specific introgressed segments within individual genomes. For that purpose, haplotype-based methods have been developed. These approaches, such as the S* statistic, utilise the fact that a recently introgressed haplotype from a deeply divergent lineage will appear as a long, distinct block of linked alleles that has not yet been broken down by recombination (Plagnol & Wall, 2006; Vernot & Akey, 2014). More recent implementations use hidden Markov models (HMMs) to probabilistically identify these segments based on patterns of LD and divergence, as used in this chapter (Skov et al., 2022). An alternative and more direct approach is archaic haplotype matching, where a modern individual's genome is compared to the high-coverage Neanderthal and Denisovan reference genomes to identify regions of unusually high similarity (Browning et al., 2018). However, this method is highly sensitive to the quality and completeness of the archaic reference genome and can be confounded by deep ILS.

Finally, the field is increasingly turning to machine learning and simulation-based approaches to handle the complexity of the task. Supervised methods, such as convolutional neural networks, are trained on vast amounts of simulated genomic data generated under models with and without introgression (Gower et al., 2021). Tools like ArchIE (Durvasula & Sankararaman, 2019) learn the multidimensional patterns of haplotype length, divergence, and LD that distinguish introgressed tracts from regions of ILS. These methods can show greater power and accuracy but are dependent on the biological realism of the simulations used for training, meaning their conclusions are only as robust as the underlying demographic model. Therefore, the gold standard in the field is the convergence of evidence from multiple independent methods which, when they agree, provide high confidence that

an identified segment is a true archaic segment. The continued refinement of these techniques is crucial for precisely mapping our archaic legacy, allowing for more confident inferences about the evolutionary impact of these ancient encounters.

4.1.3 Research Aims

This chapter aims to generate a precise map of archaic segments in modern human genomes, focusing initially on the Irish dataset. This requires a thorough examination of archaic maps and comparisons between different introgression detection methods, followed by stringent filtering criteria on the input data. We then identify genomic deserts in this dataset, hypothesising that these regions may be functionally constrained. The strength of negative selection against these deserts is explored, as well as the general distribution of archaic ancestry across the genome and across different populations.

4.2 Materials and Methods

Initial validation of the introgression detection method and subsequent analysis of archaic segments were performed using the Irish dataset (section 2). The term ‘unfiltered dataset’ refers to the original VCF files obtained from Illumina without any additional processing. The ‘filtered dataset’ refers to the data that underwent the QC procedures detailed in section 2.

4.2.1 Mapping Archaic Segments

Archaic segments were inferred from the filtered dataset of 408 Irish genomes using a HMM approach developed by Skov et al. (2020). An overview of the model is shown in Figure 4.1a. This method identifies introgressed sequences by analysing the differential density of private variants, i.e. variants present in the ingroup (Irish) but absent in an outgroup. The outgroup consisted of three sub-Saharan African populations sequenced as part of the 1000 Genomes Project (Yoruba, Mende, and Esan), which, under the out-of-Africa model of human dispersal, are expected to be largely devoid of Neanderthal introgression. All variants shared between the ingroup and outgroup were disregarded, as they likely originated prior to the ingroup-outgroup divergence. The remaining variants, private to the ingroup, are those that accumulated either on the ingroup branch after its divergence from the outgroup or on the deeper archaic branch prior to introgression (Figure 4.1a).

The genome was separated into non-overlapping 1 kb windows, and the density of private variants was calculated for each window (Figure 4.1b). The HMM classifies each window as ‘human’ or ‘archaic’ based on: i) the observed private variant density, ii) the prior probability of each state, and iii) the probability of transitioning between states in adjacent

windows. Archaic segments were defined as contiguous windows with a posterior probability of being archaic ≥0.90. To minimise edge effect false positives, 5 kb was trimmed from both ends of each inferred archaic segment. The remaining segments were then further refined by truncating them to the genomic span between the first and last private variant, following the protocol of Skov et al. (2020).

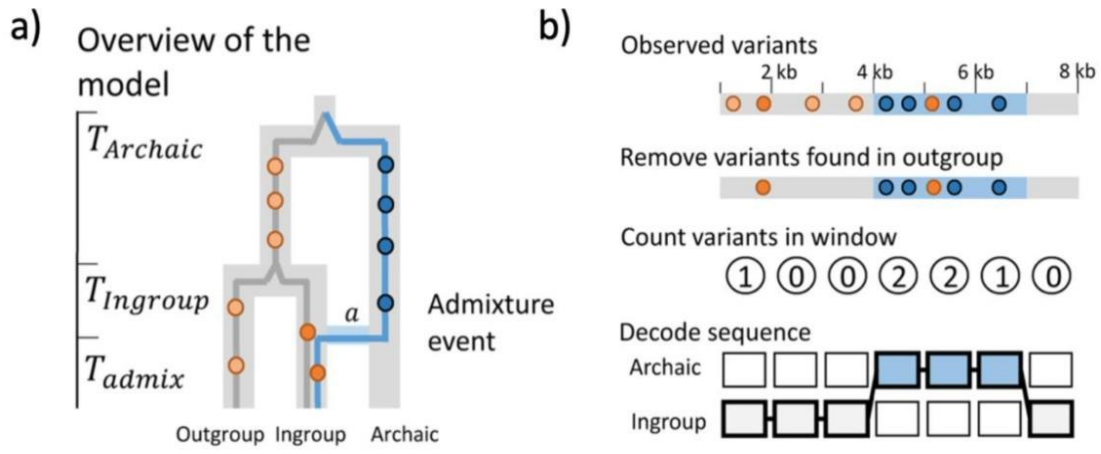


Figure 4.1, An overview of the introgression detection model. a) Schematic of the phylogenetic model. Light orange variants are shared with the outgroup; dark orange variants are ingroup-private, arising on the ingroup branch ($T_{Ingroup}$); blue variants are archaic, introgressed at time T_{admix}. b) Methodological workflow. Variants present in the outgroup are filtered out. The density of remaining private variants is calculated in 1 kb windows; windows with high density are classified as archaic. Figure taken from Skov et al. (2020).

To validate the method, private variants (variants not present in the outgroup) were categorised into three groups: 1) those from the unfiltered Irish dataset, 2) those from the filtered Irish dataset, and 3) filtered private variants located within inferred archaic segments. The variants in each group were intersected with variants from three high-coverage archaic genomes (Altai Neanderthal, Denisovan, and Vindija Neanderthal), obtained from the Neanderthal Genome Project (Meyer et al., 2012; Prüfer et al., 2014, 2017). The number of shared variants was tallied for each group.

The emission values optimised by the HMM were used to estimate coalescence times using the following equations from Skov et al. (2018):

$$T_{Ingroup} = \frac{\lambda_{Ingroup}}{2 \cdot \mu \cdot L}$$

$$T_{Archaic} = \frac{\lambda_{Archaic}}{\mu \cdot L} - T_{Ingroup}$$

where $T_{Ingroup}$ and $T_{Archaic}$ represent the mean coalescence times in generations between the ingroup and the outgroup, and between the outgroup and Neanderthals, respectively, $\lambda_{Ingroup}$ and $\lambda_{Archaic}$ represent the emission values for human and archaic segments, μ denotes the average mutation rate, and L represents the window size (Skov et al., 2018).

Using a window size of 1 kb and an average mutation rate of 1.25×10^{-8} mutations per base pair per generation, the coalescence times for the ingroup and the outgroup, as well as for the outgroup and Neanderthals, were estimated using the initial optimised parameters obtained from the unfiltered Irish dataset and those obtained from the filtered Irish dataset.

For comparative purposes, the Sprime algorithm (Browning et al., 2018) was run on the unfiltered Irish dataset. Sprime uses a composite likelihood approach to detect archaic segments by evaluating linkage patterns of alleles across all individuals simultaneously. It generates a score for SNP pairs within 10 bp to 20 kb of each other, and these scores are summed to define segments. Segments with a combined score exceeding the default threshold of 100,000 are classified as archaic. Unlike the HMM method, Sprime analysis is restricted to autosomes.

4.2.2 Analysis of Archaic Deserts

Archaic deserts were defined as genomic regions ≥1 Mb in length with ≥100,000 callable bases per 1 Mb, where no individual in the dataset carried an archaic segment (with a probability ≥0.90). To identify deserts of high confidence, we integrated results from multiple independent studies and datasets:

1. Sprime analysis of 5,639 Eurasian and Oceanian genomes (Browning et al., 2018)
2. Conditional Random Field (CRF) analysis of 1,004 genomes from the 1000 Genomes Project (Sankararaman et al., 2014)
3. IBDMix analysis of 1,843 non-African genomes (Chen et al., 2020)
4. HMM-based analysis of SGDP and HGDP genomes (Coll Macià et al., 2021)
5. HMM-based analysis of 27,566 Icelandic genomes (Skov et al., 2020)
6. The Irish dataset analysed in this chapter

The final set of 'overlapping deserts' consisted only of regions classified as deserts across all these independent analyses.

4.2.2.1 Gene Density

To compare the gene densities of archaic deserts and non-deserts, each 1 kb window of the genome was categorised as genic or non-genic based on gene annotations from the UCSC genome browser (genome.ucsc.edu). The gene density of the overlapping archaic deserts was calculated as the proportion of genic windows within them. To assess the significance of this observed density, 1,000 random sets of genomic regions were generated, each set matching the number and size distribution of the overlapping deserts. The gene density of these random sets served as a null distribution for comparison.

4.2.2.2 Functional Constraint

A list of 19,704 genes with functional constraint metrics (observed/expected (o/e) ratios for loss-of-function, missense, and synonymous mutations, and probability of loss-of-function intolerance (pLI)) was obtained from gnomAD (gnomad.broadinstitute.org). Each gene was categorised based on its overlap with the overlapping archaic deserts: 'Deserts' were defined as genes completely spanned by a desert, 'partial deserts' are genes partially spanned by a desert, and 'non-deserts' are genes that had no overlap with any desert. The o/e ratios and pLI scores were compared across these three categories to assess differences in functional constraint.

4.2.2.3 Stratified Linkage Disequilibrium Score Regression

Stratified linkage disequilibrium score regression (sLDSC) was used to test for enrichment of trait heritability within archaic deserts. Linkage disequilibrium scores for the overlapping archaic deserts were calculated by the ldsc tool developed by Bulik-Sullivan et al. (2015) using Europeans from the 1000 Genomes Project. Publicly available GWAS summary statistics for 196 traits were downloaded from the GWAS Catalog (www.ebi.ac.uk/gwas/). These summary statistics were pre-processed and then used to compute partitioned heritability estimates and enrichment scores with p-values for archaic deserts.

4.2.2.4 Private Variants vs. Archaic Variants

To test for functional constraint in archaic genomes, we analysed the density of SNPs from three archaic genomes (Altai Neanderthal, Vindija Neanderthal, and Denisova; obtained from the Neanderthal Genome Project (Meyer et al., 2012; Prüfer et al., 2014, 2017)) within desert and non-desert regions. All callable windows with a minimum of 100 aligned reads in

the archaic genome were classified as either desert (archaic count = 0 in the Irish dataset) or non-desert (archaic count ≥1). The average number of archaic variants per window was calculated for both desert and non-desert windows for all three archaic genomes. The number of variants in the archaic genomes per window was also compared to the archaic count in the corresponding window in the Irish dataset.

4.2.3 Principal Component Analysis

A principal component analysis (PCA) was performed in R (Team, 2021) using two types of pairwise similarity matrices constructed from the Irish dataset. The first pairwise matrix was constructed to capture the proportion of shared archaic sequences between each pair of individuals, excluding the X chromosome. This proportion was calculated as the total length of shared archaic sequence divided by the total length of the autosomes. The matrix was then double-centred by transforming the data to have zero-mean rows and columns. A second matrix was generated to assess the fraction of shared private variants within the total shared archaic sequence between each pair of individuals. This fraction was calculated as the total number of shared private variants located within a shared archaic segment divided by the total length of the shared archaic sequence between each pair of individuals. Similarly to the previous matrix, normalisation was performed by double-centring the matrix. The PCA was conducted on both matrices, as well as on each chromosome separately for the latter matrix.

4.3 Results

4.3.1 Preamble

The following sections describe the inference and detailed analysis of introgressed archaic segments in modern Irish genomes and comparison with segments inferred in diverse worldwide populations. The early stages of this process involved the establishment of suitable variant filtration and quality control parameters to provide reliably called archaic segments, which itself relied on first inferring archaic segments. As this was by definition an iterative process, there is no perfect or correct order in which to present the results. Here, we first describe the initial analysis of archaic segments which led to the establishment of quality control procedures defining the filtered dataset. Although this followed the mapping of archaic segments, description of this process and subsequent analyses begins at section 4.3.5.

4.3.2 Initial Analysis of Archaic Peaks

The filtering of the Irish dataset (and later, subsequent datasets) was a crucial process in order to improve the accuracy of the inference of archaic segments, as is evident by the results obtained prior to filtering. Initially, mapping introgressed segments in the unfiltered data suggested regions of the genome in which an improbably large number of individuals were deemed to carry archaic sequence. In particular, there were three regions in which all 408 Irish genomes had been inferred as archaic. To validate these findings, the Sprime algorithm was also applied, which uses a distinct, population-wide approach. Despite the differences in the methods used, a substantial overlap was observed, particularly within these peaks (Figure 4.2). This improbably high level of introgression relative to the genomic

average suggested that these peaks were likely anomalous and attributable to properties of the input data rather than a model-specific artifact.

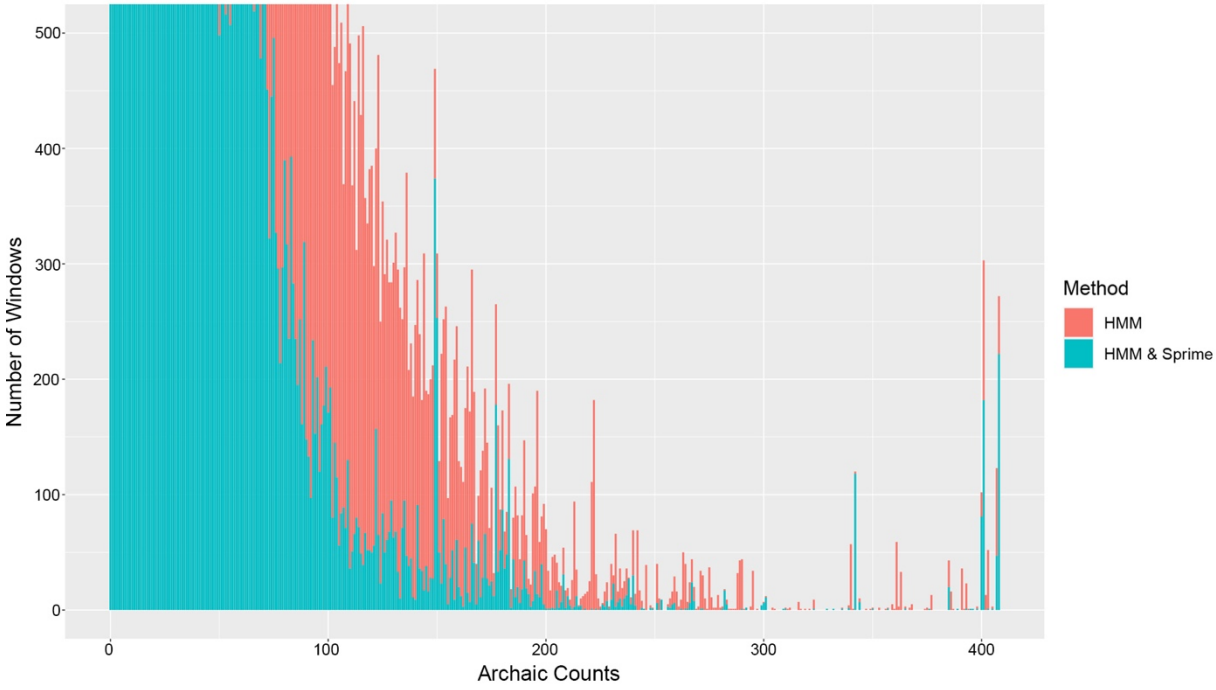


Figure 4.2, The distribution of archaic counts using the HMM method and their agreement with the sprime method. The x-axis represents the archaic count, defined as the number of individuals inferred to carry an archaic haplotype within each 1 kb window. Each bar is subdivided by colour to indicate whether the archaic signal in those windows was also independently identified by the sprime algorithm. It is important to note that sprime does not produce archaic counts because it analyses all genomes simultaneously and outputs a single list of archaic segments. To enhance the visibility of archaic peaks, the y-axis is limited to 500.

Consequently, the characteristics of the private variants located in the archaic peaks were then investigated, focusing on the top three peaks that were inferred as archaic in all individuals. If these segments were truly archaic and fixed in the population, we would expect high homozygosity in their variants. Contrary to this expectation, an initial assessment revealed substantial heterozygosity. This was assessed using an Exact SNP test for Hardy-Weinberg Equilibrium (HWE) (Wigginton et al., 2005), the results of which are shown in Figure 4.3. Compared to the rest of the genome, all three archaic peaks contain a considerable number of private variants that deviate significantly from HWE, indicating an excess of heterozygosity.

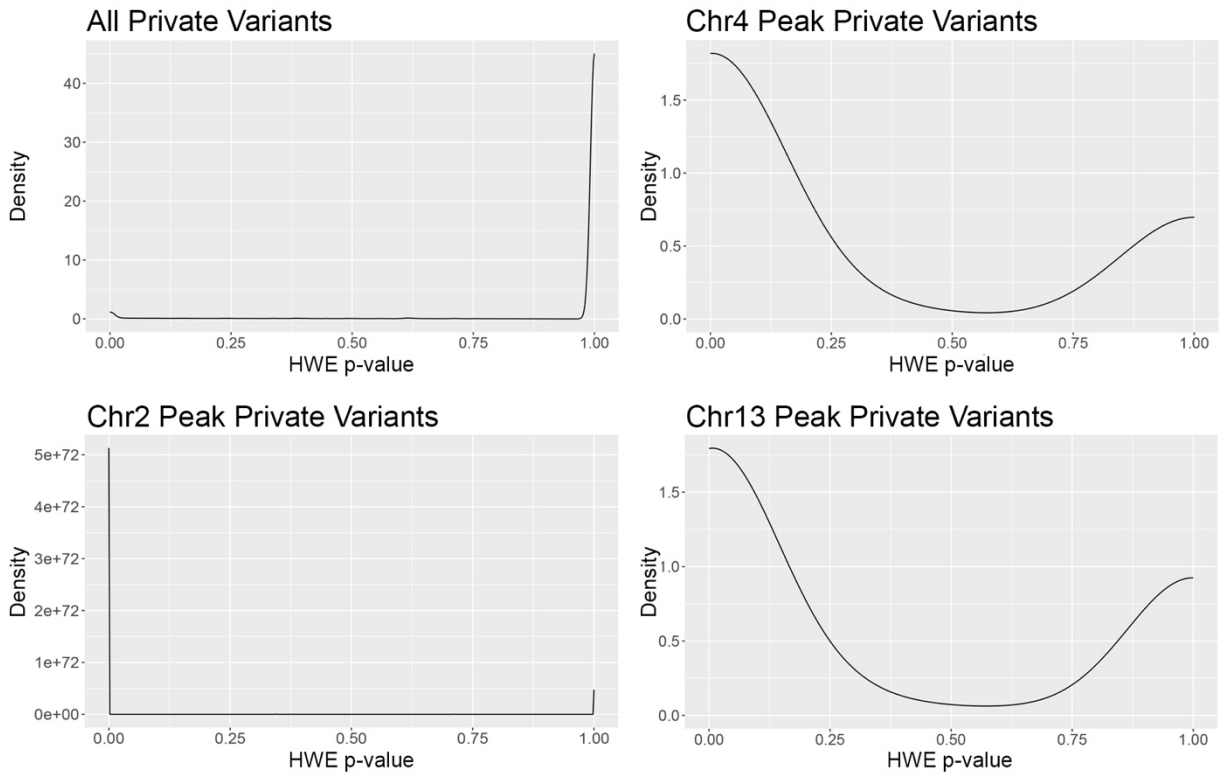


Figure 4.3, HWE deviation in archaic peaks. The distribution of HWE p-values in all private variants is compared to the private variants in each of the top three peaks, where significant p-values indicate a deviation from HWE.

The distribution of these private variants were then visualised using chromoMap (Anand & Rodriguez Lopez, 2022) (Figures B1-B4). A representative archaic peak (chr4) and a control genomic region are shown in Figure 4.4.

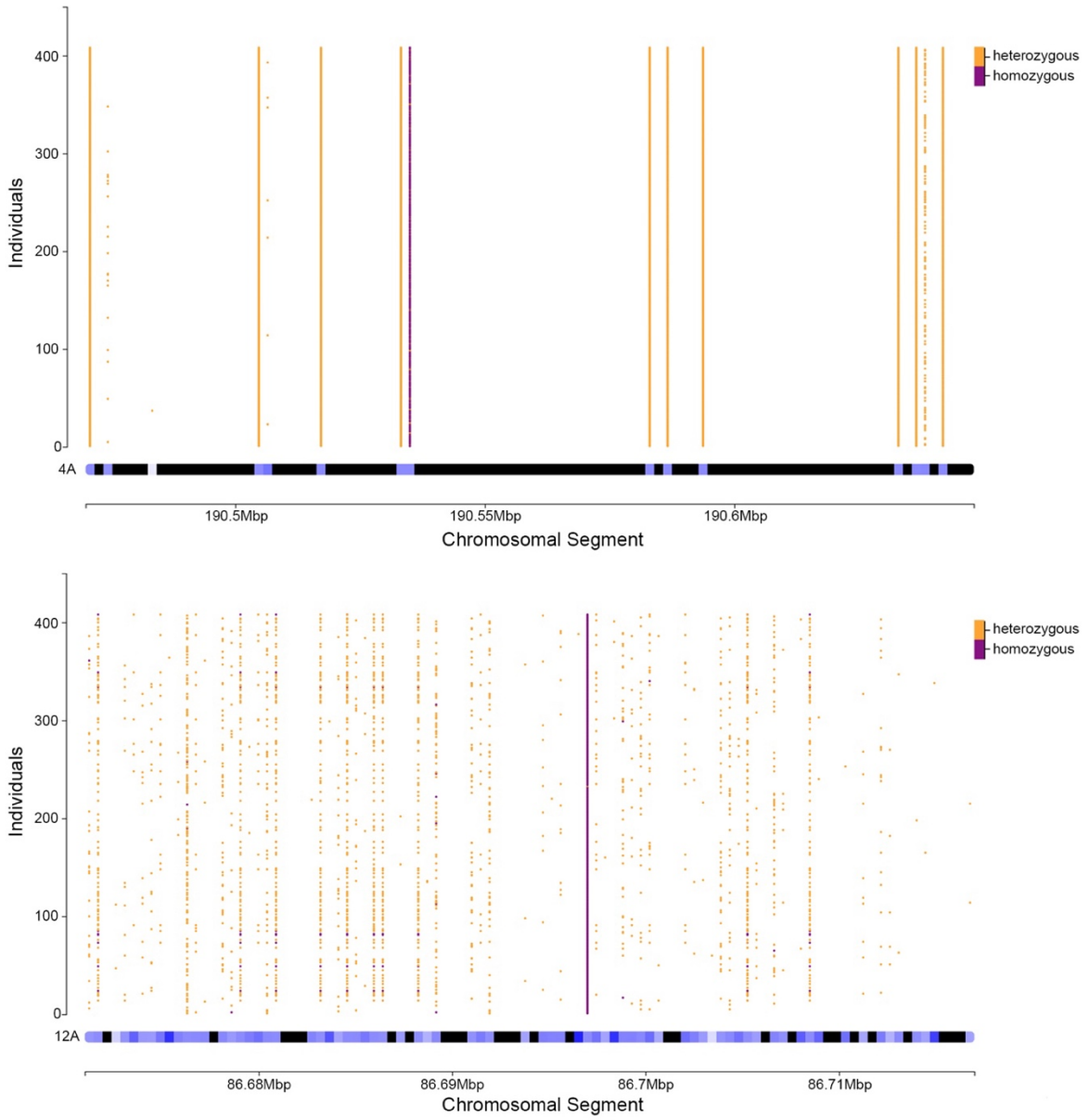


Figure 4.4, Heterozygote comparison between an archaic peak (top) and a control region (bottom). These chromoMap visualisations include an archaic peak (top) and a random 46 kb archaic region (bottom). Each individual has been assigned a number, 0-407, displayed on the y-axis. For each private variant that is present in an individual, a yellow or purple dot is assigned to their corresponding number, depending on whether their genotype is homozygous (purple) or heterozygous (yellow).

These visualisations provided insight into the distinct patterns observed within the archaic peaks. Many of the private variants within the peaks appear in a heterozygous manner across all individuals, which is not observed in the randomly selected regions. This occurrence is rather unusual since a variant present in all individuals would typically be observed in a

homozygous state unless it is not tolerated as a homozygote, suggesting these patterns may be artefactual.

4.3.3 HGDP Dataset

To provide a comparative analysis and investigate whether archaic peaks appear in other populations, the HGDP dataset was analysed. The HGDP genomes were separated into populations, and their archaic proportions are depicted in Figure 4.5. Consistent with expectations, Oceanian individuals exhibit the highest archaic proportion, reflecting the stronger impact of Denisovan introgression on their genomes. East Asians displayed higher archaic proportions compared to Europeans, which is also consistent with previous observations (Vernot & Akey, 2014; Vilanea & Schraiber, 2019; Witt et al., 2022). To internally assess the performance of the HMM, African groups from the HGDP dataset were included in this evaluation; these exhibited a scattered distribution of inferred archaic proportion, which is to be expected given the method's reliance on sub-Saharan African genomes as an outgroup. HGDP African genomes are excluded from subsequent analyses.

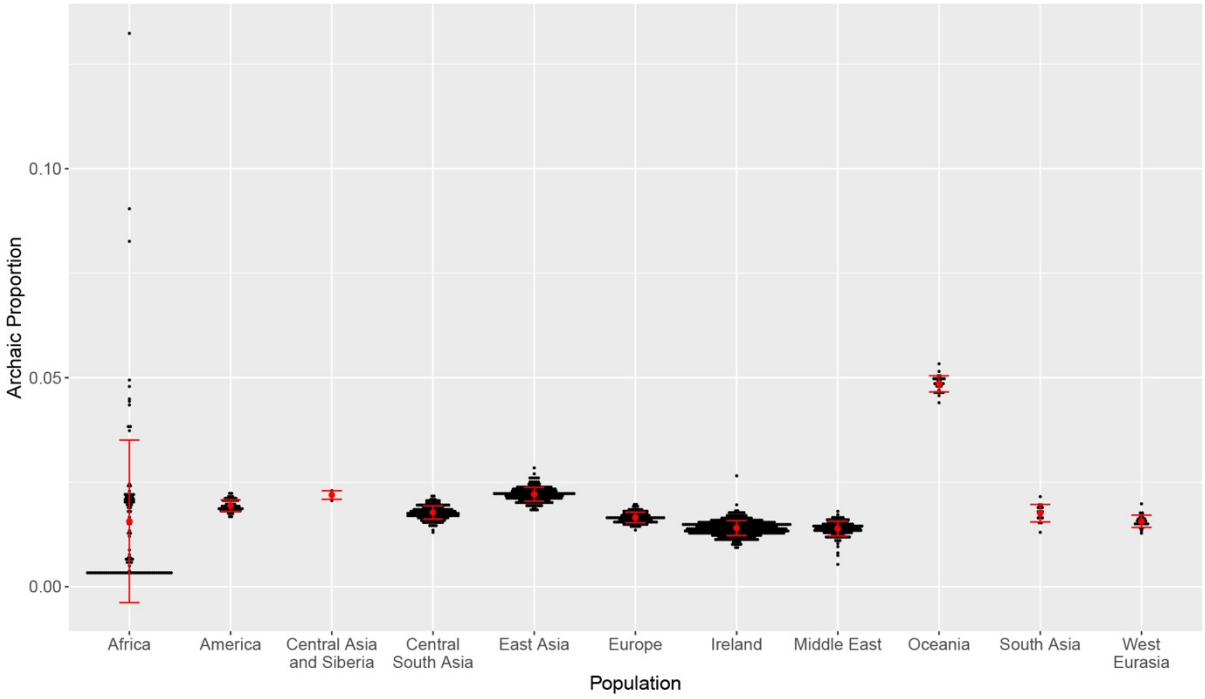


Figure 4.5, Archaic proportions of all HGDP populations and individuals from the Irish dataset. The distributions of archaic proportions in each population are shown here, with each individual represented as a black dot and the standard error of the mean as red bars.

Following the conversion of the coordinates of all HGDP archaic segments from GRCh38 to GRCh37 using liftOver (Hinrichs et al., 2006), a comparison was conducted between the archaic counts of all genomic windows in the Irish dataset and the corresponding archaic counts in the HGDP dataset. This analysis included the entire HGDP dataset as well as their individual populations, as depicted in Figure 4.6. Correlation coefficients indicated that the Irish dataset was most similar to European, Middle Eastern and South Central Asian HGDP populations but less similar to East Asian and American populations. Additionally, the high density of windows along the x- and y-axes suggests the presence of many archaic segments with high archaic counts in one population but absent in the other. This observation may reflect population-specific archaic ancestry, a phenomenon previously highlighted by Witt et al. (2022), but it could also be confounded in part by the possibility of sequencing errors or other artifacts in the Irish data, given that the archaic peaks observed in the Irish genomes are not present in the HGDP genomes.

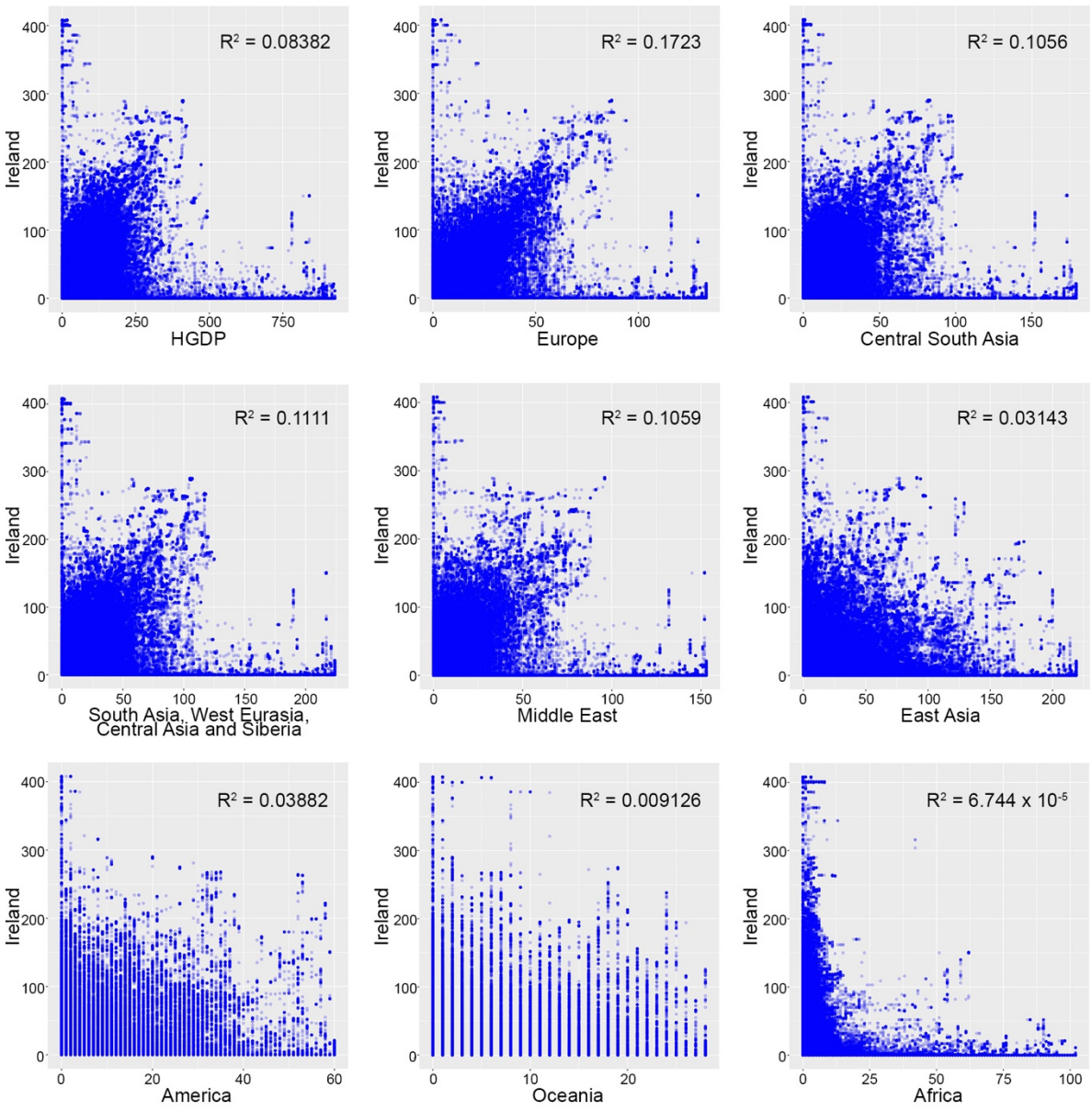


Figure 4.6, Archaic count comparisons between the Irish dataset and all HGDP populations. The archaic count for each callable window in the Irish genome is compared to the corresponding archaic count in the entire HGDP dataset, as well as each population separately. The correlation coefficients for each comparison have also been provided.

4.3.4 Quality Control of the Irish Dataset

To assess the quality of the variants contributing to the three archaic peaks highlighted in section 4.3.2, individual-level genotyping metrics were examined, specifically focusing on allele balance (AB) and genotype quality (GQ). The excess of heterozygotes observed

within the peak regions prompted an investigation into AB, which represents the ratio of reads aligned at a variant locus that support the alternate allele. We hypothesised that if the heterozygote reads within peak regions were false positives, there would be a reduced number of reads supporting the alternate allele at these sites compared to the rest of the genome. This was confirmed, as depicted in Figure 4.7. A one-sided Wilcoxon rank sum test demonstrated that the AB within the peaks was significantly lower than in non-peaks (p-value $< 2.2 \times 10^{-16}$), indicating that these reads are likely false positives. To address this issue, a filtering step was implemented, retaining only genotypes with an AB between 0.25 and 0.75. This criterion resulted in the removal of 28,132,163 heterozygote genotypes.

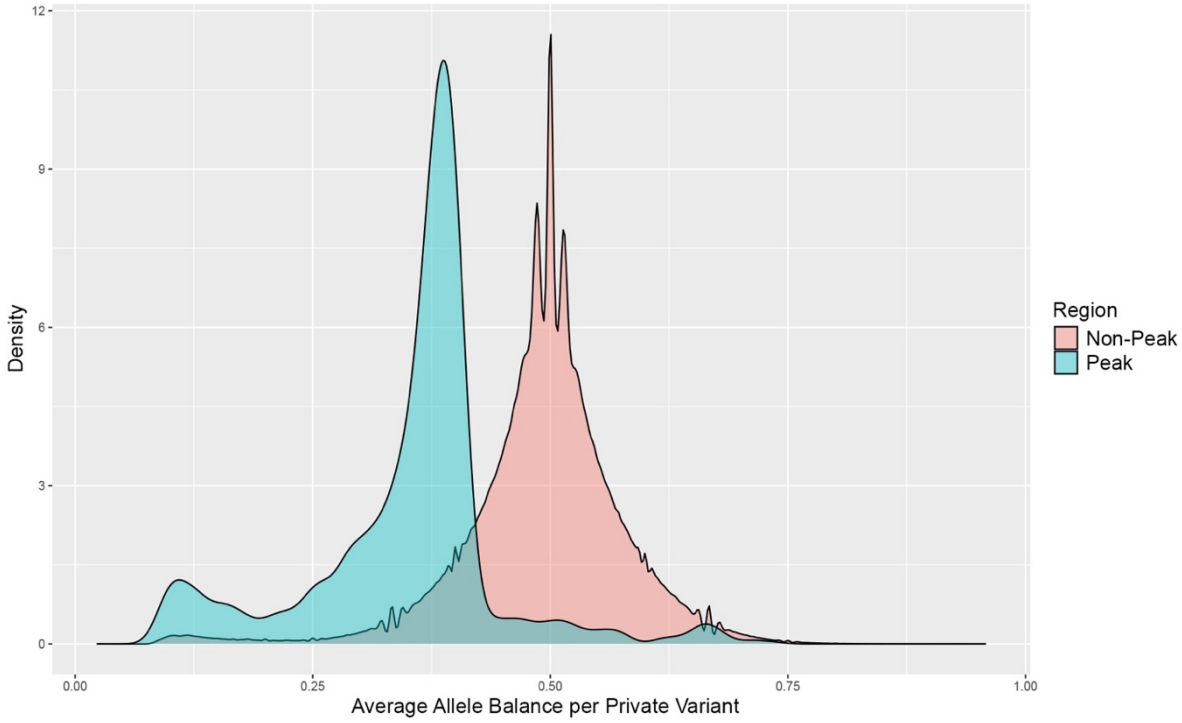


Figure 4.7, AB comparison between archaic peak and non-peak regions. The average AB of all heterozygote genotypes within the three top archaic peaks is compared to the distribution of heterozygote AB throughout the rest of the genome.

The subsequent analysis of GQ scores revealed that genotypes (excluding those that are homozygous reference) within archaic peak regions harbour lower average GQ scores compared to nonpeak regions (Figure 4.8). This difference was statistically significant according to a one-sided Wilcoxon rank sum test (p-value $< 2.2 \times 10^{-16}$). Given that Skov et

al. (2020) removed variants where the probability of the genotype being correct was <90%, genotypes with a GQ score lower than 10 were removed, corresponding to a probability of less than 90%. This filtering step eliminated 90,383,720 genotypes post-AB filtering.

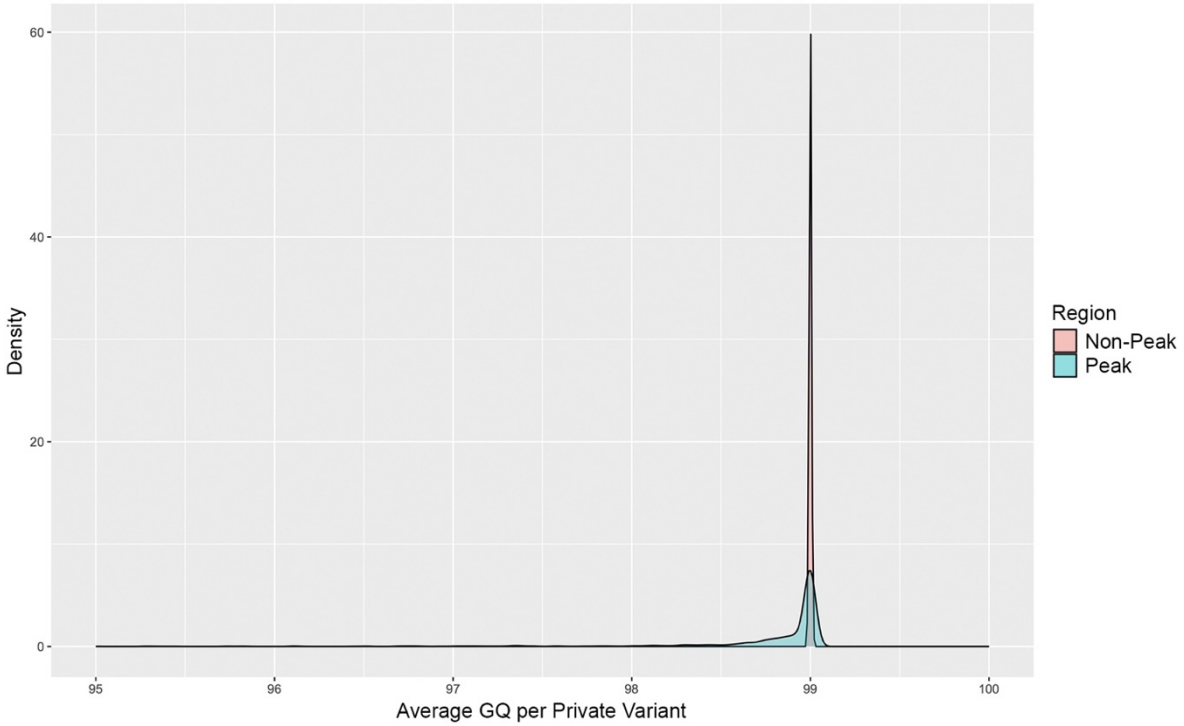


Figure 4.8, GQ comparison between archaic peak and non-peak regions. The average GQ scores of all genotypes (excluding those that are homozygous reference) for all private variants within the top three archaic peaks are compared to those for all remaining private variants.

Variant-wide analysis revealed a significant discrepancy in inbreeding coefficients between peak and non-peak regions (one-sided Wilcoxon rank sum test, $p < 2.2 \times 10^{-16}$; Figure 4.9), consistent with the observed heterozygosity excess. Variants with an inbreeding coefficient below -0.8 were subsequently removed, resulting in the elimination of 52,994 variants.

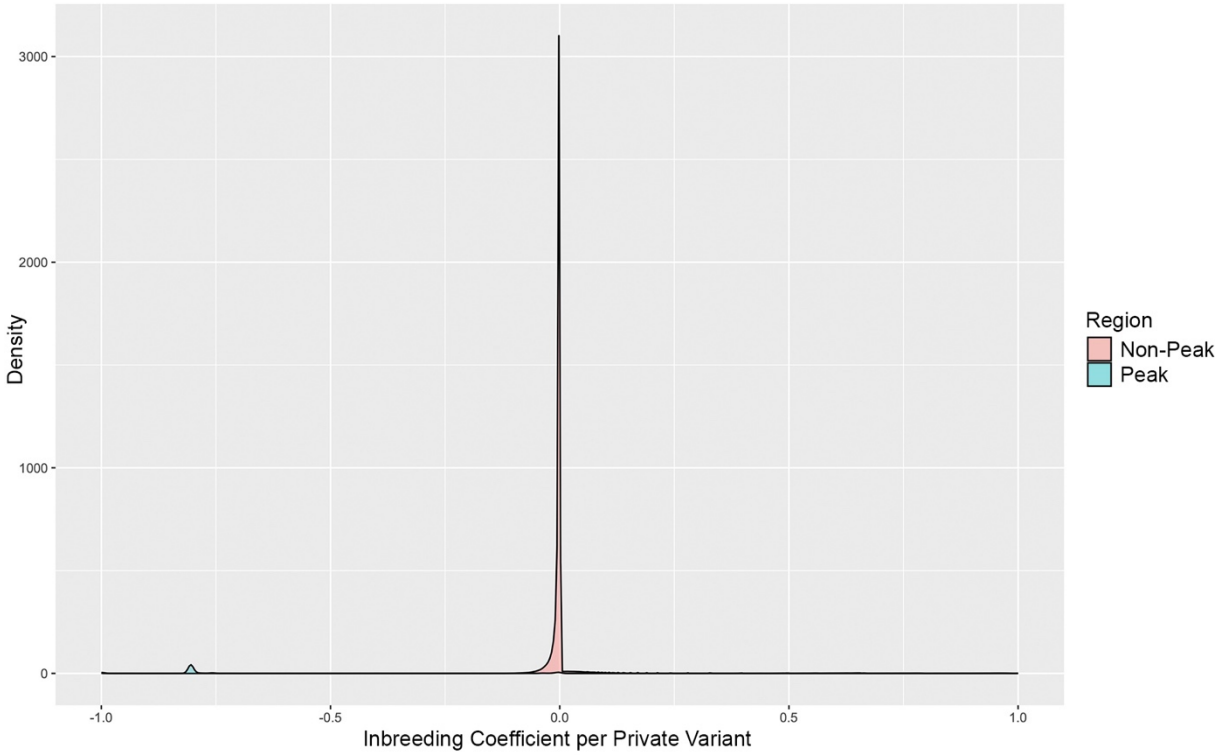


Figure 4.9, Inbreeding coefficient comparison between archaic peak and non-peak regions. The inbreeding coefficients for all private variants within the top three archaic peaks are compared to those for all remaining private variants.

Overall, this filtering process had a significant impact on the distribution of archaic counts generated by the HMM analysis. The effectiveness of the filtering is shown in Figure 4.10, which displays a substantial reduction in the number of private variants located within archaic peaks after applying the filtering thresholds, while retaining the majority of the remaining private variants. This is especially evident for the AB and inbreeding coefficient thresholds, which were most effective at removing private variants within archaic peaks without disrupting variants elsewhere. The collaborative effect of these filters is shown in Table 4.1; many variants failed multiple quality thresholds, strongly supporting the hypothesis that they were systematic false positives.

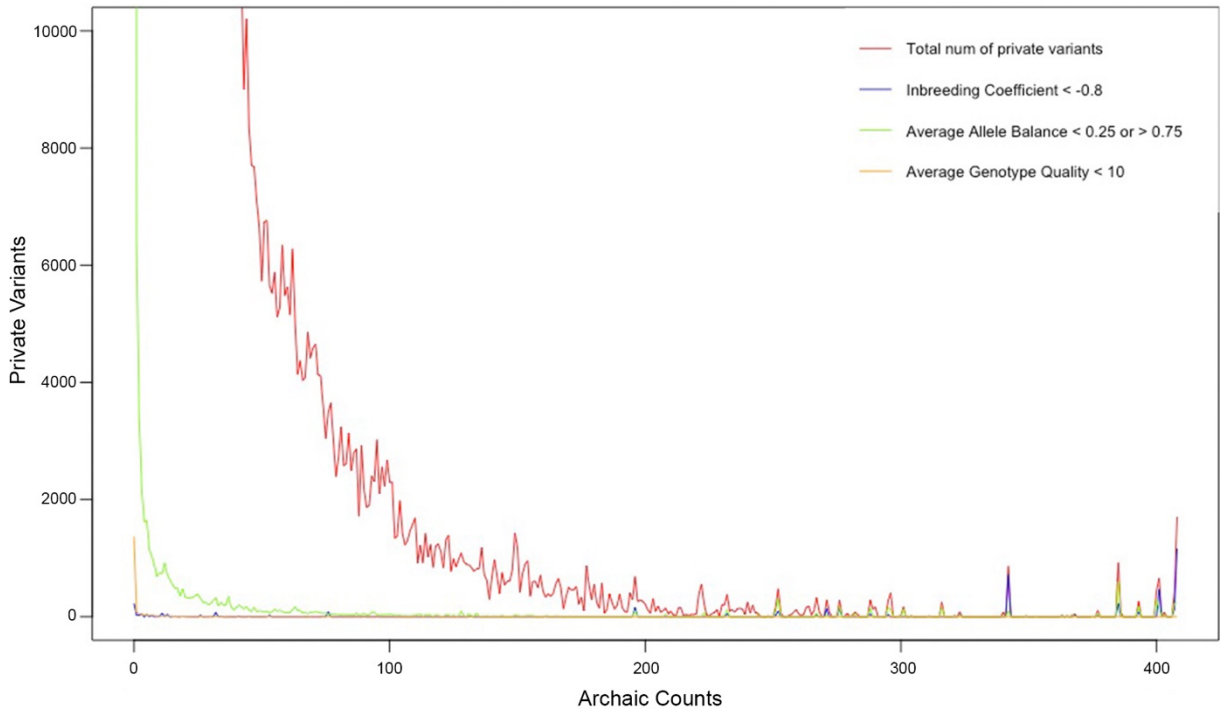


Figure 4.10, Private variants stratified by archaic count. This illustrates the number of private variants per archaic count, as well as the number of those private variants that failed the three QC thresholds.

Table 4.1, Filtering efficacy.

Category	*Not filtered*	*AB*	*GQ*	*INB*	*AB+GQ*	*AB+INB*	*GQ+INB*	*AB, GQ + INB*
Num. variants	13,108,499	3,422,336	2,044,401	5,254	1,585,767	19,422	83	15,713

The number of variants failing to meet the QC threshold for AB, GQ, and inbreeding coefficient (INB) alone are displayed, as well as those that failed multiple QC thresholds. A variant is considered to be filtered by AB or GQ if at least one genotype within the variant is removed.

4.3.5 Identification and Analysis of Archaic Segments

4.3.5.1 Archaic Segments

The HMM requires three parameter sets: starting probabilities (the probability of starting in either the human or archaic state), a transition matrix (the probabilities of remaining in the current state or transitioning to the other state), and emission values (mean private variants

per window in human/archaic states) (Table 4.2). The input values used for these parameters were those employed by Skov et al. (2020) in their study.

Table 4.2, HMM parameters.

		Human	*Archaic*
Starting Probabilities		0.98	0.02
		0.968	0.032
		0.952	0.048
Transition Probabilities	Probability of remaining in current state	0.9995	0.98
		0.9988	0.9637
		0.9989	0.9798
	Probability of transitioning state	0.0005	0.012
		0.0011	0.0363
		0.0010	0.0202
Emission values		0.04	0.1
		0.0715	0.6501
		0.0676	0.4136

The parameters that were used as initial estimates (red) are provided, along with the optimised parameters for the unfiltered Irish dataset (blue) and those for the filtered Irish dataset (green).

The filtered data yielded a higher average archaic proportion per individual (1.5%) than the unfiltered data (1.4%). This is attributed to the optimised emission value for the archaic state, which was lower in the filtered dataset (0.4136 vs. 0.6501), which allowed for the detection of more segments despite a lower absolute number of private variants. Trimming 5 kb from each segment end reduced the average archaic fraction to 1.3%. The distributions of archaic fractions in the unfiltered and filtered (after trimming 5 kb ends) datasets are depicted in Figure 4.11, illustrating that the filtered dataset is more normally distributed. This was confirmed by a Shapiro-Wilk test, which assesses the normality of a dataset based on its p-value. The Shapiro-Wilk test yielded a significant p-value for the unfiltered dataset

(2.291×10^{-9}), indicating a departure from normality, whereas the filtered dataset did not exhibit significant deviation from normality ($p = 0.0655$).

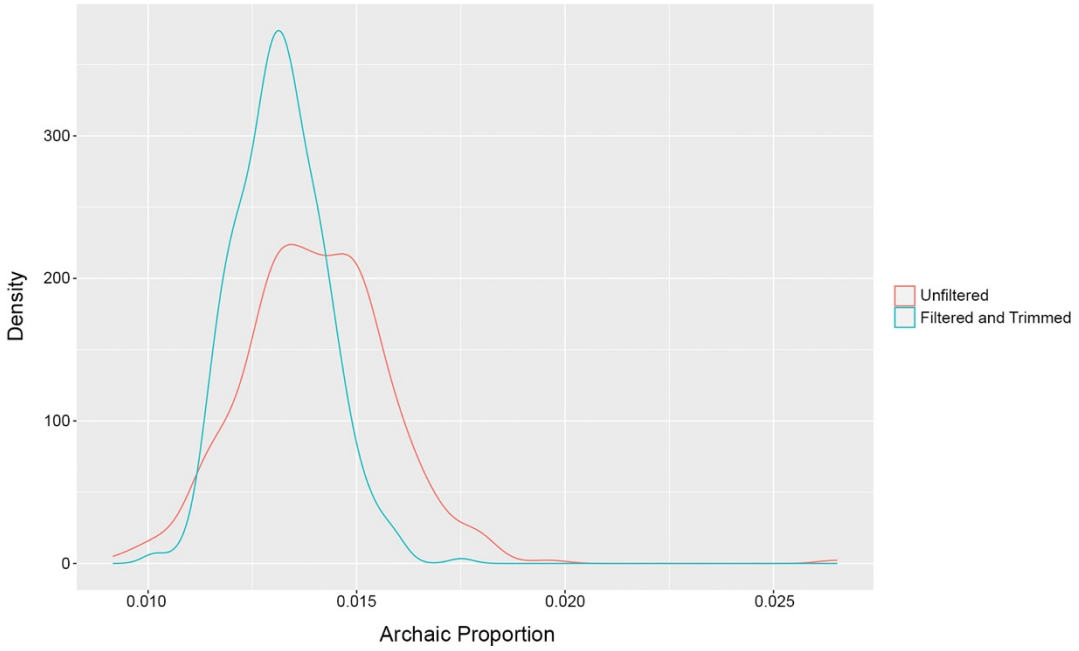


Figure 4.11, Archaic proportion distributions. The distribution of archaic proportions from the unfiltered dataset is displayed alongside the distribution of archaic proportions from the filtered and trimmed Irish dataset.

The archaic segments spanned a total of 816 Mb in the filtered dataset (890 Mb before trimming the ends), compared to 799 Mb in the unfiltered dataset. The increased coverage of archaic ancestry in the filtered dataset can be attributed to the ability to infer more archaic ancestry throughout the genome. In the filtered data, on average, each individual was found to harbour 40 Mb of archaic sequence, exceeding the average length of archaic sequence detected per individual in the study by Skov et al. (2020) (33 Mb). The average archaic segment length was 126,879 bp. This is likely an overestimate of the mean length of archaic segments in the Irish genome, as the method employed in this study is less sensitive for short archaic fragments, resulting in concatenation of adjacent segments. Considering that gene flow occurred approximately 2,000 generations ago, archaic segments in the modern human genome are expected to have a length of around 50 kb (Sankararaman et al., 2014).

The comparison of archaic counts in each callable window of the genome between the filtered and unfiltered datasets is depicted in Figure 4.12. As anticipated, a strong linear correlation is observed (R^2 = 0.6099), which indicates that the filtering process did not significantly affect the archaic count for most windows. The excess of points along the x-axis signifies the complete removal of artefactual archaic peak regions called in the unfiltered data. One notable exception is a large peak spanning 39 kb on chromosome 2, which remains after filtering. In the unfiltered dataset, this peak had an archaic count of 408 out of 408 individuals, while in the filtered dataset, the archaic count of the region ranges from 308 to 346. This suggests that this particular region may indeed represent a genuine archaic peak since it persisted after the filtering process. Alternatively, it raises the possibility that the filtering steps employed were not sufficiently stringent to eliminate all peaks. Interestingly, the next largest peaks identified in the filtered dataset are exclusive to the filtered dataset. This implies that the original unfiltered dataset generated not only false positives indicated by archaic peaks but also potential false negatives, as the emission values used may have been too stringent to identify these regions as archaic segments.

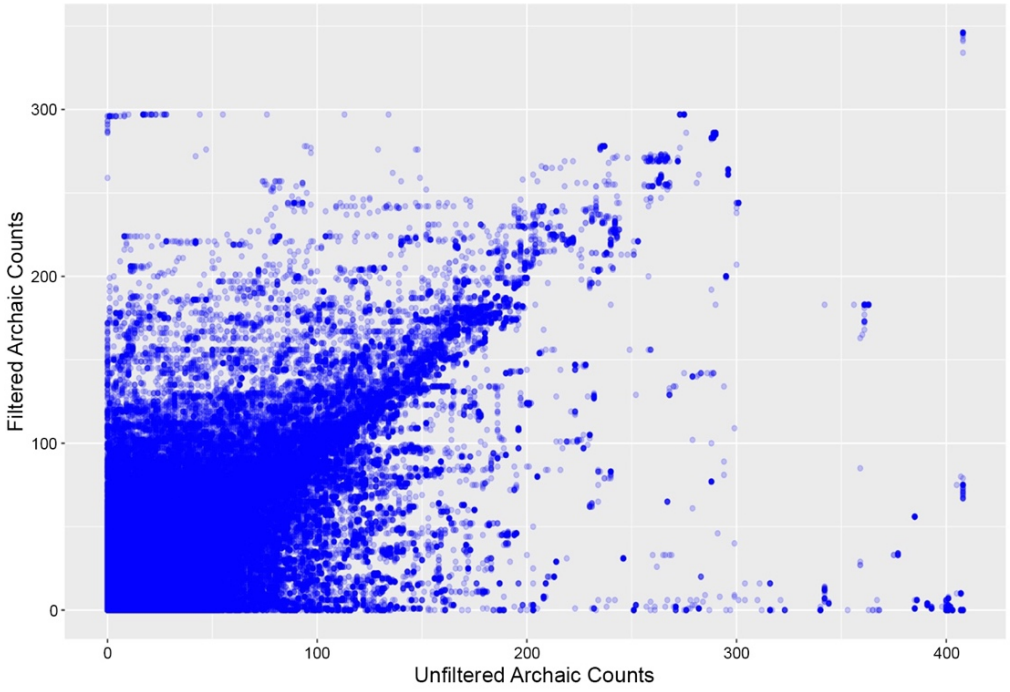


Figure 4.12, Filtered vs unfiltered archaic counts. The archaic count of each window in the callable genome is compared between the filtered and unfiltered Irish dataset.

4.3.5.2 Divergence Time Estimation

Using the optimised parameters, we estimated divergence times between the ingroup and the outgroup, as well as between the outgroup and Neanderthals. Assuming a generation time of 30 years, the unfiltered dataset produced a coalescence time of 85,800 years (2,860 generations) between the ingroup and the outgroup, and a coalescence time of 1,474,440 years (49,148 generations) between the outgroup and Neanderthals. The filtered dataset produced a coalescence time of 81,120 years (2,704 generations) between the ingroup and the outgroup, and a coalescence time of 911,520 years (30,384 generations) between the outgroup and Neanderthals. The inflated outgroup-Neanderthal time in the unfiltered data is due to false positive private variants increasing the apparent divergence. The filtered estimates are more consistent with published divergence times (~62 kya for sub-Saharan Africans/non-Africans and <800 kya for modern humans/Neanderthals (Fu et al., 2013; Gómez-Robles, 2019; Tassi et al., 2015)).

4.3.5.3 Private Variants vs. Archaic Variants

To confirm the veracity of inferred archaic segments and explore the intersection between private variants in these segments and directly observed archaic variants in ancient sequences, three archaic genomes (Altai, Denisova, and Vindija) were examined across three categories of private variants: those present in the unfiltered Irish dataset, those present in the filtered Irish dataset, and those present in the filtered Irish dataset and located within archaic segments. These categories were selected based on the expectation that the filtered dataset would have a higher degree of similarity to the Neanderthal variants since it contains a more refined set of private variants due to the removal of false positives, and that those in archaic segments would have an even greater overlap. Consistent with expectations, the proportion of private variants overlapping archaic variants was higher in the filtered dataset compared to the unfiltered dataset, although the difference was not substantial (Table 4.3, Figure 4.13). More notably, when considering private variants specifically within archaic segments, the proportion of Altai and Vindija Neanderthal variants overlapping with private variants increased by over two-fold. This observation lends support to the validity and efficacy of the introgression detection method used. Denisovan overlap was low across all categories, consistent with their relative absence in European genomes.

Table 4.3, Overlap of private variants with archaic genomes.

	All Private Variants (unfiltered dataset)	*Private Variants (filtered dataset)*	*Private Variants in archaic segments*
Total	8,258,583	4,716,039	1,594,170
Altai	183,146	116,378	90,199
Denisovan	74,275	46,618	30,222
Vindija	206,602	131,274	102,483
DAV	226,952	144,418	110,883

Three categories of private variants were analysed: all private variants in the unfiltered dataset, all private variants in the filtered dataset, and those from the filtered dataset found within an archaic segment. The total number is provided for each category, as well as the number of those found in one of the archaic genomes or in all three archaic genomes (DAV).

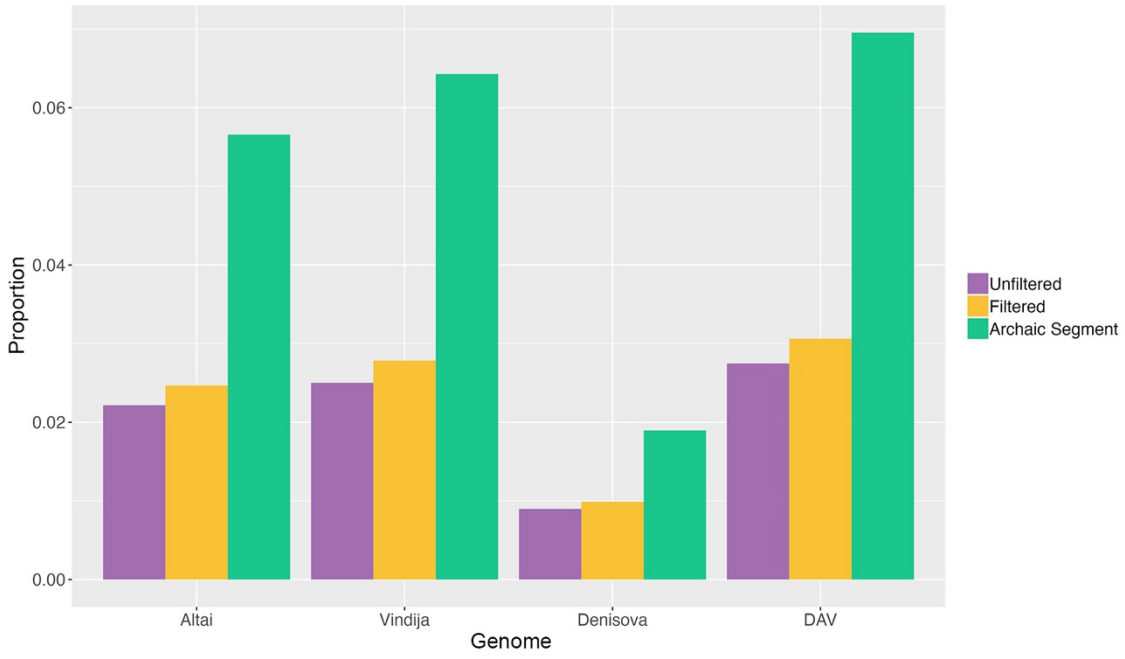


Figure 4.13, Private variant proportion overlap with archaic genomes. The plot shows the proportion of private variants (unfiltered, filtered, and filtered variants within archaic segments) that are also present in individual archaic genomes or shared by all three (DAV).

4.3.6 Archaic Deserts

521 archaic deserts were identified in the filtered Irish dataset, with a total length of 1.1 Gb. This number is notably higher than in other datasets, even those that contain fewer individuals. The overlap of the deserts from multiple datasets shown in Table 4.4 resulted in 38 deserts with a cumulative length of 24.6 Mb. The distribution of these deserts along each chromosome is displayed in Figure 4.14. The X chromosome, a major source of deserts in the Irish data, was excluded from this overlap analysis due to its absence in some comparator datasets.

Table 4.4, Summary statistics of multiple datasets of inferred archaic segments.

	No. genomes analysed	*No. archaic counts*	*Average no. archaic counts per genome*	*No. deserts*	*Total desert length*	*Average desert length*	*Reference*
*Irish dataset**	408	16,316,824	39,992	521	1.1 Gb	2.11 Mb	NA
HGDP	825	38,465,820	46,625	102	205 Mb	2.01 Mb	NA
*Icelandic dataset**	27,566	NA	NA	268	0.53 Gb	2.00 Mb	(Skov et al., 2020)
*SGDP**	209	9,921,656	47,472	374	860 Mb	2.30 Mb	(Coll Macià et al., 2021)
*HGDP**	115	5,523,747	48,032	433	949 Mb	2.19 Mb	(Coll Macià et al., 2021)
CRF	1,004	39,545,921	39,388	433	1.24 Gb	2.88 Mb	(Sankararaman et al., 2014)
Sprime (excl. ChrX)	5,369	NA	NA	333	760 Mb	2.28 Mb	(Browning et al., 2018)
IBDMix (excl. Africans)	1,843	99,376,567	53,921	332	820 Mb	2.47 Mb	(Chen et al. 2020)

**Archaic introgression inferred using the HMM outlined in this chapter.*



Figure 4.14, The distribution of archaic deserts and segments throughout the genome in the Irish dataset. The grey plots above the chromosomes represent the number of archaic counts within each window, and the bars below the chromosomes signify the locations of the overlapping deserts from all datasets described in Table 4.4 (purple), and from the Irish dataset alone (blue).

4.3.6.1 Gene Density

We hypothesised that archaic deserts might be gene-dense regions under selective constraint against archaic introgression, as a higher gene density in deserts could indicate that archaic DNA is less tolerable in these regions because they are functionally important regions. The gene list used for this analysis contained 32,092 genes and included noncoding genes in addition to protein-coding genes. The average gene density in the 38 overlapping archaic deserts (0.58837 bases of genic sequence per base of desert) was higher than in non-deserts (0.44491). This difference was significant: the observed gene density was higher than 1,000 randomly generated sets of regions matched for number and size (Figure 4.15), supporting the hypothesis that deserts are functionally important regions where archaic DNA is selected against.

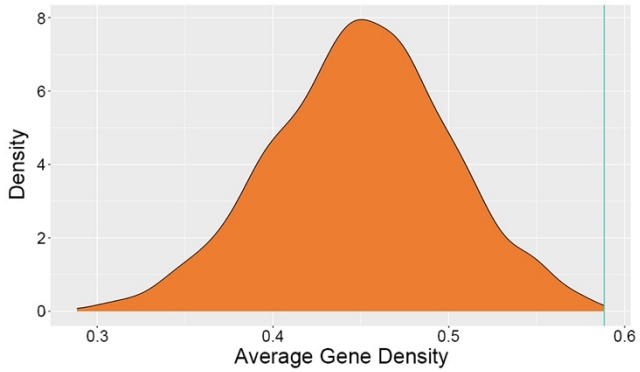


Figure 4.15, Gene density of archaic deserts compared to a random null model. The density plot shows the distribution of average gene density for 1,000 matched random region sets, and the green line indicates the observed gene density for archaic deserts themselves.

4.3.6.2 Functional Constraint

To evaluate the hypothesis that archaic deserts arise from negative selection against introgressed Neanderthal DNA in functionally constrained regions, we compared metrics of functional constraint between archaic deserts, partial deserts, and non-desert regions (Figure 4.16). One-sided Wilcoxon rank sum tests were employed to analyse observed-to-expected (o/e) ratios for three mutation categories: loss-of-function (LoF), missense, and synonymous variants. The results of these tests are presented in Table 4.5. Marginally significant differences (p-value < 0.05) were identified in three comparisons. Specifically, o/e ratios for missense mutations were significantly lower in both deserts and partial deserts compared to non-deserts. The o/e ratio for synonymous mutations was also significantly lower in deserts than in non-deserts.

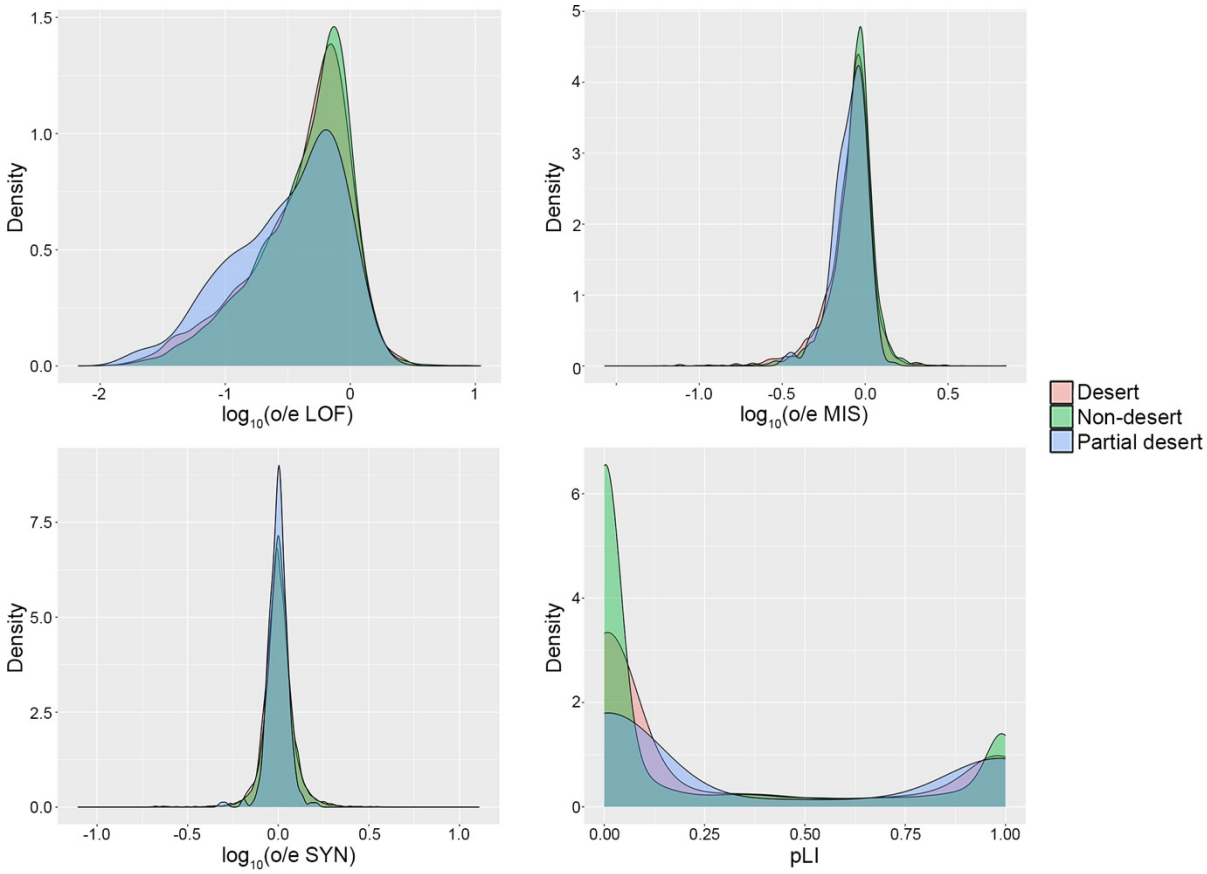


Figure 4.16, Functional constraint metrics for genes stratified by overlap with archaic deserts. Density plots show the distribution of o/e ratios for LoF, missense, and synonymous mutations, and pLI scores. Genes are grouped by their overlap with archaic deserts (fully within, partially within, or non-deserts).

Table 4.5, Wilcoxon rank sum tests for differences in gene constraint metrics between archaic desert categories.

Wilcoxon Test p-value	*O/e LOF*	*o/e MIS*	*o/e SYN*
Desert – Non-desert	0.7386	0.04027	0.0008249
Partial desert – Non-desert	0.1035	0.0415	0.06965
Desert – Partial desert	0.9399	0.8043	0.4766

p-values are reported for comparisons of o/e ratios (LoF, missense (MIS), synonymous (SYN)) between genes in deserts, partial deserts, and non-desert regions.

These findings lack strong evidence for the functional constraint of archaic deserts. The analysis of LoF mutations, which are the most direct indicator of strong purifying selection due to their severe disruptive effects, yielded the highest p-values and were not statistically

significant. Conversely, the most significant difference was observed for synonymous mutations, which are typically considered nearly neutral and are not expected to be under strong functional constraint. Despite this, the most consistent and noteworthy result is that both deserts and partial deserts exhibit significantly reduced o/e ratios for missense mutations compared to non-desert regions. This pattern signals a possibility, albeit not supported by their LoF counterpart, that these regions are under some selective pressure, supporting the hypothesis that purifying selection against deleterious Neanderthal alleles has contributed to the formation of archaic deserts.

4.3.6.3 Stratified LD Score Regression

sLDSC was conducted on the overlapping archaic deserts to investigate whether these regions were enriched or depleted for the heritability of complex traits. The analysis encompassed a diverse set of 196 traits, chosen with particular emphasis on those previously associated with Neanderthal ancestry, such as hair traits, skin pigmentation, immunity, height, and smoking status (Dannemann & Kelso, 2017; Harris & Nielsen, 2016; Rinker et al., 2020; Zeberg & Pääbo, 2021). The enrichment values for all traits are displayed in Figure 4.17. To account for multiple testing, a Bonferroni correction was applied, requiring that results yield a p-value below 0.00026 in order to be considered significant. No traits surpassed this threshold, indicating no significant enrichment or depletion of trait heritability within archaic deserts after multiple-testing correction.

Nevertheless, two prominent traits showed notable associations: body mass index (BMI) and Brussel sprout liking. BMI showed the strongest enrichment, with a 40-fold increase in heritability within archaic deserts (p-value = 0.0466). In contrast, Brussel sprout liking demonstrated the most pronounced depletion, with an enrichment value of 0.0780 (p =

0.0085). While intriguing, these results did not survive multiple testing correction and should be interpreted with caution as potential false positives.

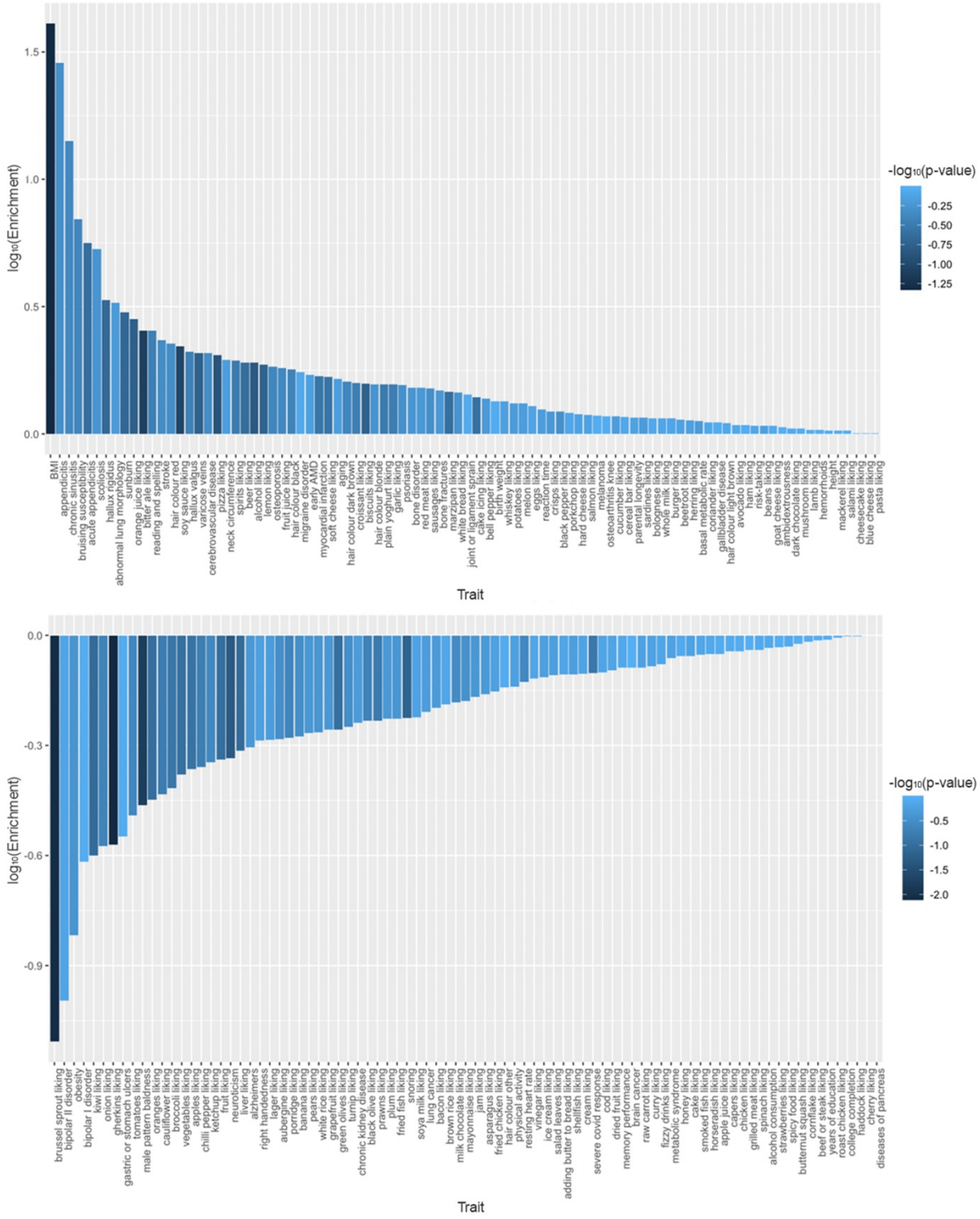


Figure 4.17, Enrichment of trait heritability in archaic deserts. Traits are split into those enriched (top) and depleted (bottom) for heritability within archaic deserts relative to the genomic background. Bar colour denotes the -log10(p-value), with darker blue indicating greater significance. Traits with negative enrichment values are excluded.

4.3.6.4 Archaic Variant Density vs. Archaic Counts

We then sought to determine whether the functional constraint hypothesised to create archaic deserts in the modern human genome is also evident in the archaic genomes themselves. If these regions are under similar selective constraint in Neanderthals and Denisovans, they would exhibit a reduced density of genetic variants. All callable 1 kb windows were classified as either deserts (archaic count = 0) or non-deserts (archaic count ≥1) and the distribution of SNP counts from three archaic genomes (Altai, Vindija, Denisova) were compared between these categories.

Contrary to our hypothesis, the proportion of windows containing a given number of archaic SNPs was nearly identical between desert and non-desert regions for all three archaic genomes (Figure 4.18). This suggests that the variant density in these putative regions of constraint is not significantly different from the rest of the genome in archaic individuals, indicating that the evolutionary forces shaping these regions may not be shared between modern and archaic hominins. However, this does not necessarily indicate that archaic deserts are not functionally constrained. As mentioned previously, Neanderthals had a much smaller effective population size than humans. The absence of functional constraint evidence in their genomes may therefore reflect their reduced efficacy of purifying selection.

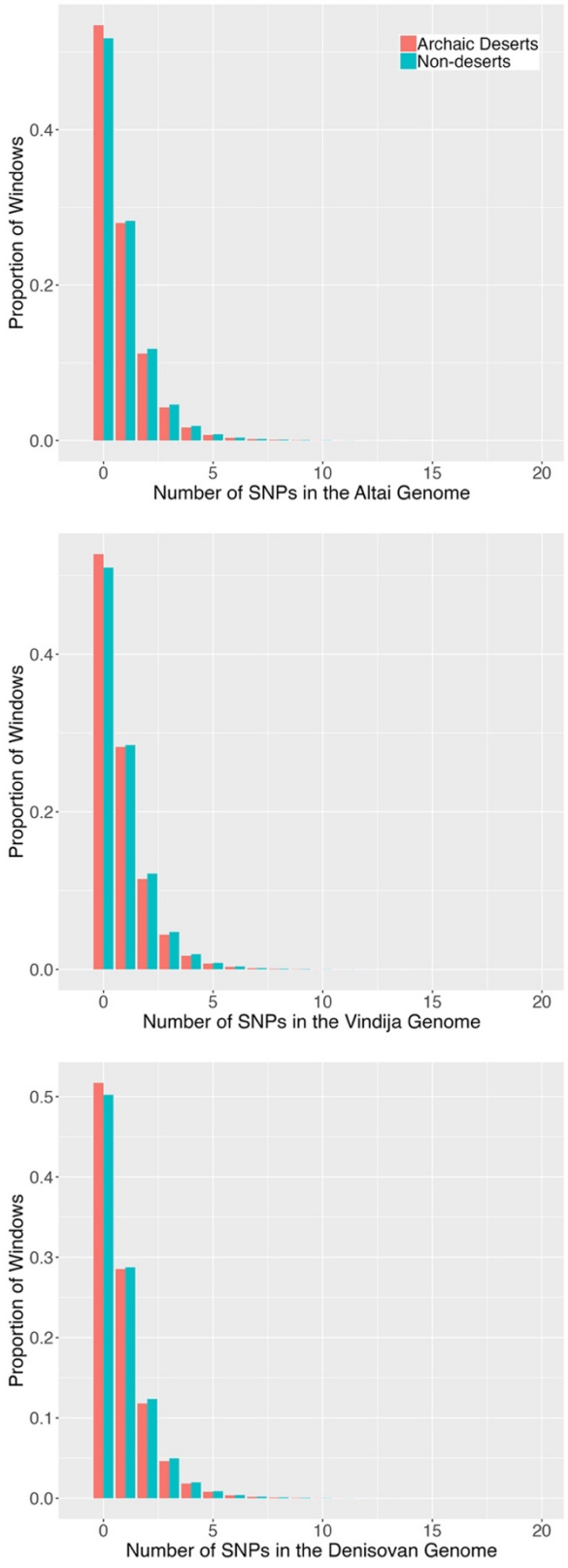


Figure 4.18, Assessment of the depletion of archaic variants in deserts. These plots display the distribution of archaic SNP density in desert vs non-desert windows. The proportion of windows containing a given number of SNPs in the archaic genomes is shown for each category.

We further explored the relationship between archaic SNP density in a window and its archaic count in the Irish dataset (Figure 4.19). Two competing factors could influence this relationship. First, high divergence (many archaic SNPs) could make introgressed segments easier to detect in the modern human genome, creating a positive correlation. However, high divergence could also indicate lower tolerance for introgression, creating a negative correlation. The data show a clustering of points along the y-axis, indicating that windows with a high density of archaic SNPs tend to have low archaic counts in the Irish population, supporting the second scenario of lower tolerance. However, we also observed a surprising excess of points along the x-axis, which represents windows with high archaic counts in the Irish data but very few or no archaic SNPs. This is paradoxical, as the HMM method relies on identifying differences from the African outgroup shared with archaic genomes; regions with no archaic SNPs should not be detectable as introgressed. However, the HMM method concatenates nearby short archaic segments, which could explain the lack of archaic SNPs in certain windows that were inferred as archaic. Alternatively, this could be caused by introgression from an unsampled archaic population whose genome differs from the three reference sequences used here.

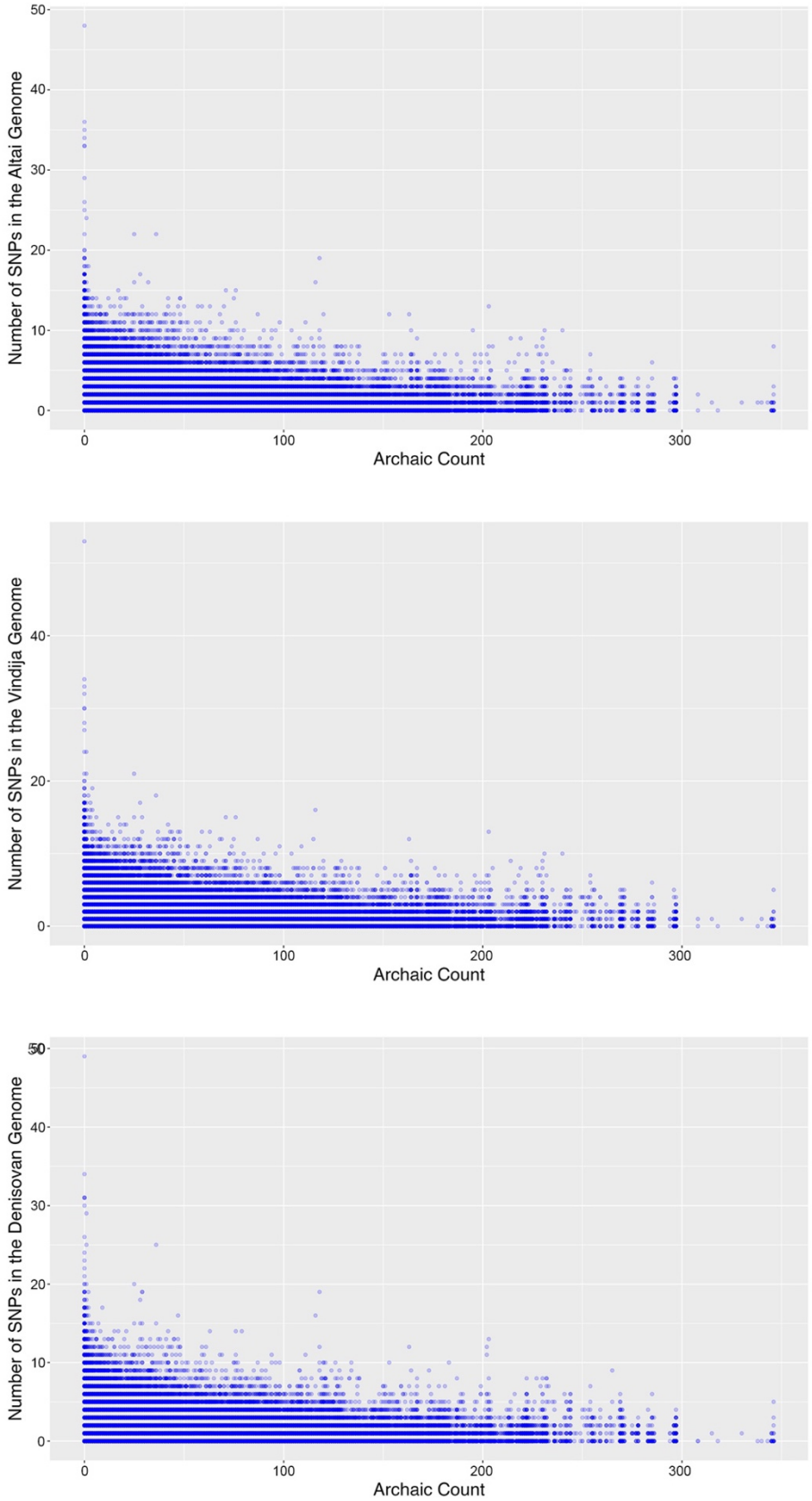


Figure 4.19, Correlation between archaic count and density of archaic SNPs. This plot shows the correlation between archaic count in the Irish dataset and the number of SNPs in the archaic genomes for every callable genomic window. Windows with low archaic coverage (<100 reads) were excluded.

4.3.7 Principal Component Analysis

4.3.7.1 Archaic Overlap

We performed a PCA on the archaic overlap between pairs of individuals by creating a matrix representing the proportion of shared archaic sequence in the genome (excluding the X chromosome) for each pair. The goal was to determine if individuals formed clusters based on their archaic segments, which could potentially reflect historical demographic events such as genetic bottlenecks, migration patterns, or founder effects. Given that it has previously been demonstrated that archaic segments are sufficient to distinguish between populations (Witt et al., 2022), it was particularly intriguing to investigate archaic overlap among Irish individuals and assess whether they form distinct clusters based on their archaic segments. The first four principal components were analysed, although they explained only a very small proportion of the overall variance (PC1 = 0.41%, PC2 = 0.033%, PC3 = 0.028%, and PC4 = 0.025%; Figure 4.20).

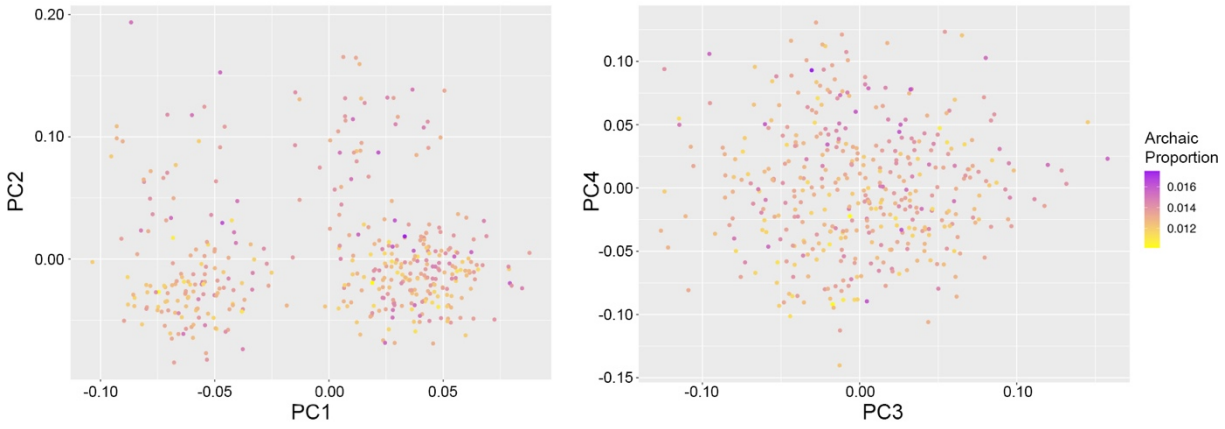


Figure 4.20, Archaic overlap PCA plots. These plots show the distribution of individuals along the first four PCs. Individuals are coloured by their archaic proportion.

Visualisation of PC1 versus PC2 revealed a relatively even distribution along PC2 but suggested some clustering along PC1 (Figure 4.20). We tested for correlations between these PCs and individual attributes (sex, geographic coordinates, archaic fraction, ALS status,

given that the data are a repurposed dataset from an ALS study), the results of which can be found in Appendix B (Figure B5). The archaic fraction per individual has been plotted alongside the PCA plot (Figure 4.20) to facilitate visual interpretation.

We also investigated whether the patterns on PC1 corresponded to previously identified genetic clusters within Ireland (Byrne et al., 2018), which are shown in Figure 4.21. Because these clusters could be distinguished based on genetic similarity (Figure 4.21(B)), we investigated whether they also show distinct patterns based on archaic similarity. However, individuals from different genetic clusters were found to be distributed relatively evenly across the range of PC1 values (Figure 4.22). Although the North Munster (NMN) cluster appeared slightly concentrated at lower PC1 values, a Kruskal-Wallis test confirmed no statistically significant difference in PC1 values among the clusters ($p = 0.2598$).

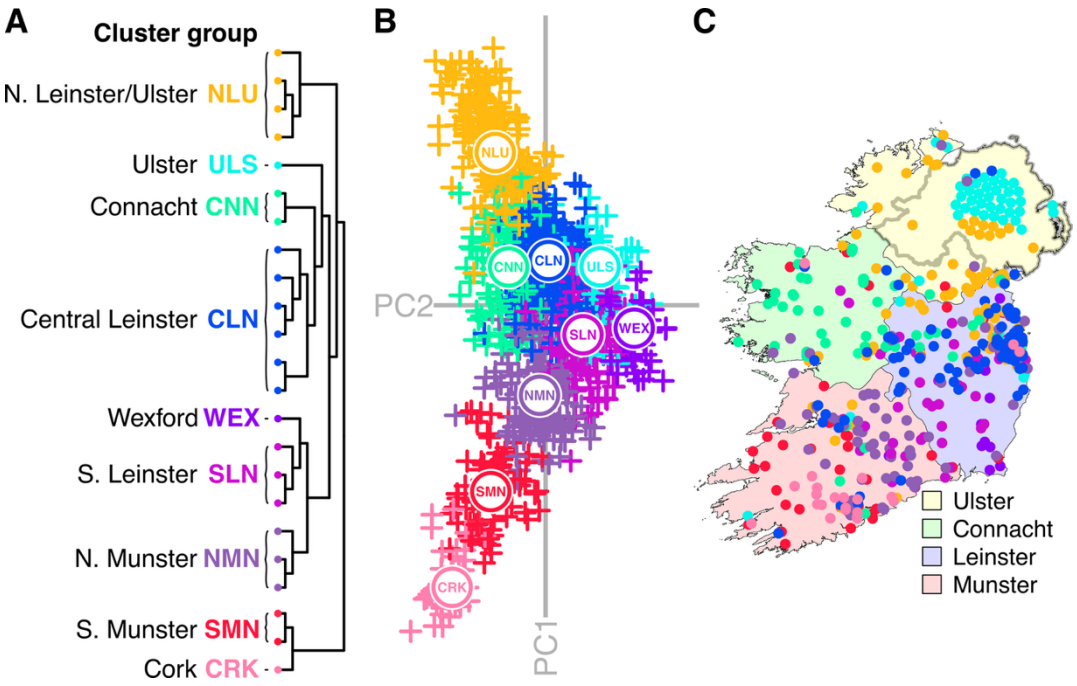


Figure 4.21, Depiction of the genetic clusters within Ireland. A) Clustering dendrogram for 1,035 Irish individuals. B) Principal component analysis of haplotypic similarity (Chromopainter), where points are coloured according to their cluster groups. The median location of each cluster group is plotted. C) Map of Ireland showing the sampling location for a subset of 588 individuals, coloured by cluster group. Precise sampling location for 44 Northern Irish individuals from the People of the British Isles dataset was unknown; these individuals are plotted geometrically in a circle. Figure adapted from Byrne et al., (2018).

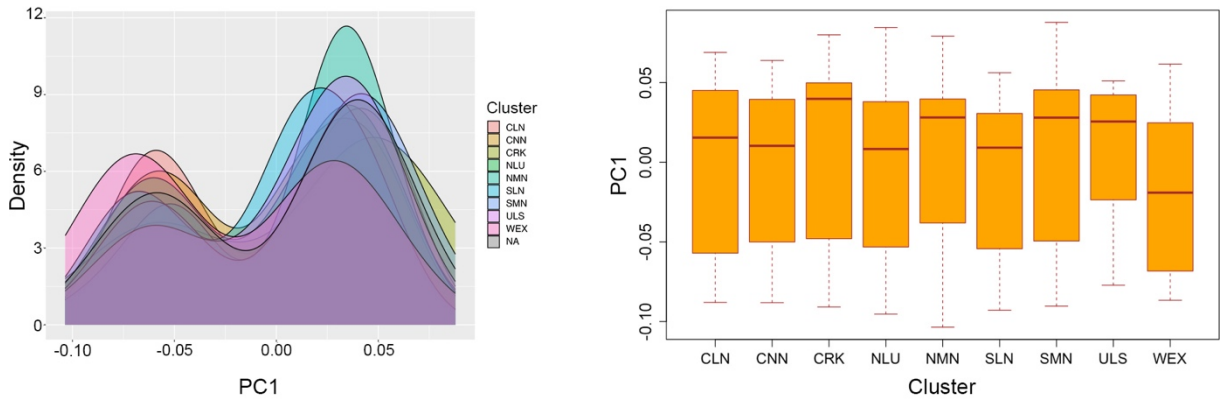


Figure 4.22, Characterisation of Irish genetic clusters along PC1 (Archaic overlap PCA). A) Density plot showing the distribution of individuals from each cluster along the PC1 axis. B) Boxplots summarising the median and variation of PC1 values for each cluster.

4.3.7.2 Private Variants

To obtain a more refined measure of shared archaic ancestry, we performed a second PCA based on a matrix of the fraction of shared private variants within shared archaic segments (excluding the X chromosome). The number of private variants shared between each pair of individuals (in regions that were inferred as archaic in both individuals) was quantified and expressed as a fraction of the total length of shared archaic sequence. This was first conducted using multiple populations from the HGDP dataset. The results, depicted in Figure 4.23, reveal an absence of clustering in all six populations analysed. The PCA plot, however, does tend to cluster based on population (Figure 4.24), suggesting descriptive patterns of archaic introgression unique to different populations.

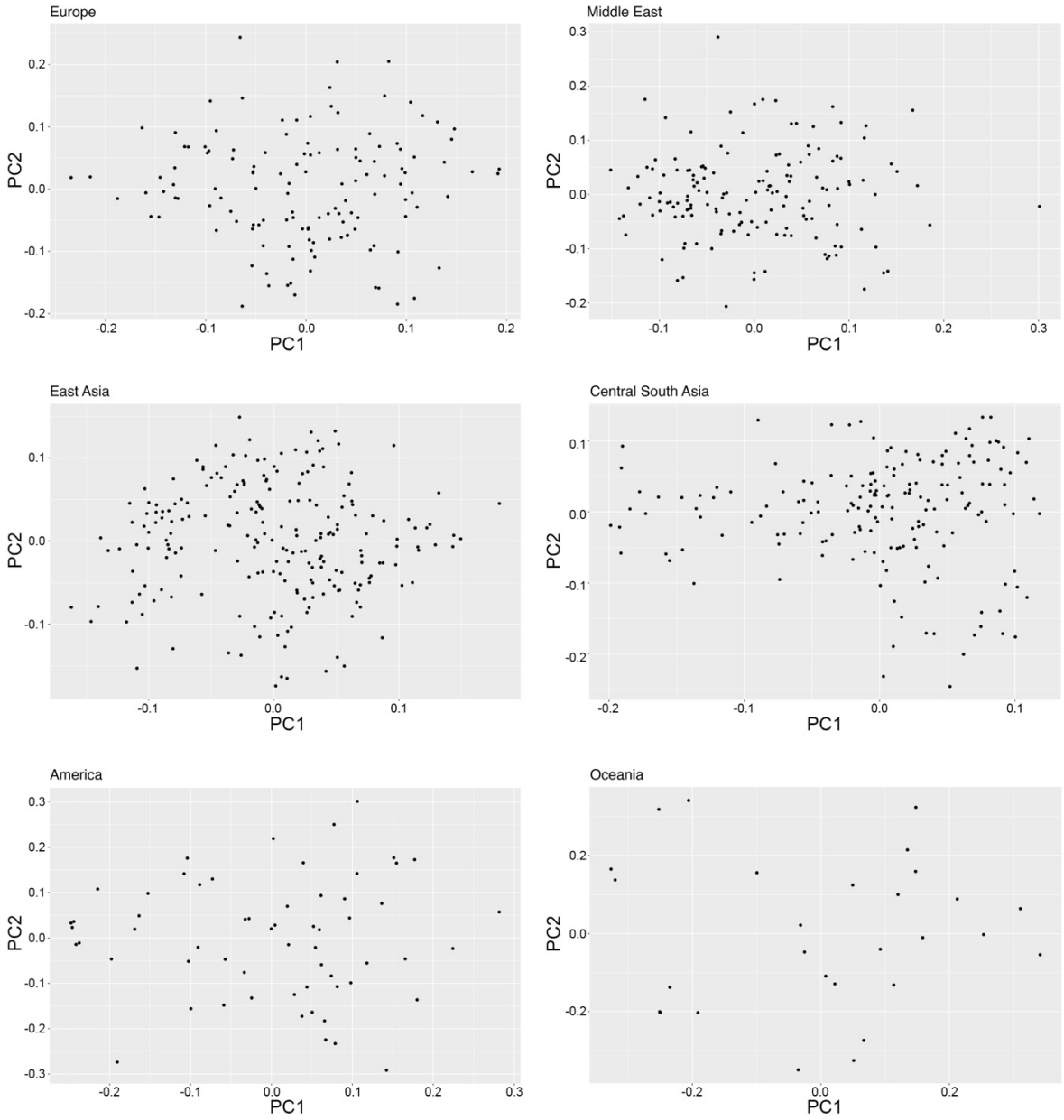


Figure 4.23, HGDP populations private variant PCA plots. These plots display the distributions of individuals along PC1 and PC2 for each HGDP population.

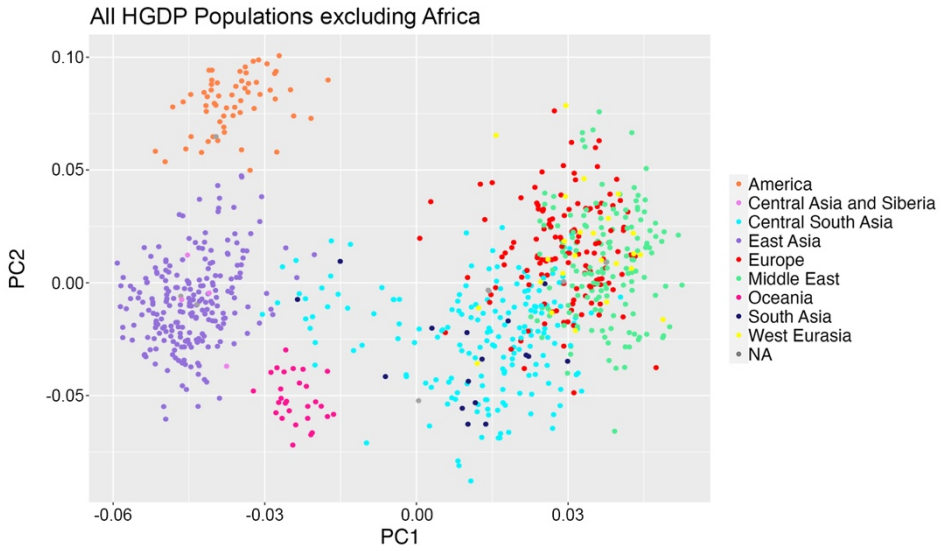


Figure 4.24, HGDP private variant PCA plot. This PCA plot shows the distribution of all individuals in the HGDP dataset along PC1 and PC2, with individuals coloured by population.

This analysis was replicated in the Irish dataset, which revealed a clear separation of individuals into two distinct clusters along PC1 (Figure 4.25). This clustering was not correlated with sex, geography, archaic fraction, or ALS status (Figure B6). The PCA results were also compared to the genetic clusters identified by Byrne et al. (2018), as illustrated in Figure 4.26. While the Wexford cluster appears to exhibit some correlation with PC1, a Kruskal-Wallis test indicates that neither the correlation between genetic clusters and PC1 (cluster–PC1 p-value = 0.8534) nor the correlation between genetic clusters and PC2 (cluster–PC2 p-value = 0.8932) reached statistical significance.

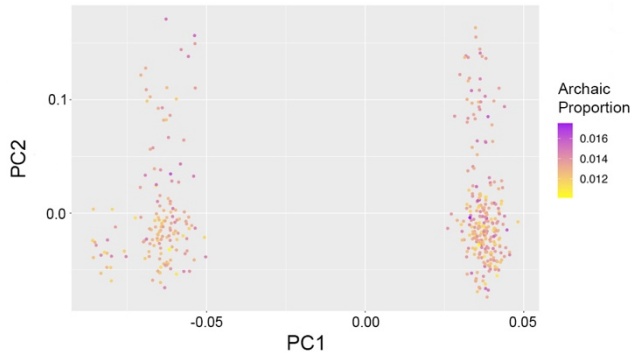


Figure 4.25, Private variant PCA plot. This plot shows the distribution of individuals along PC1 and PC2 for the private variant PCA. Individuals are coloured by their archaic proportion.

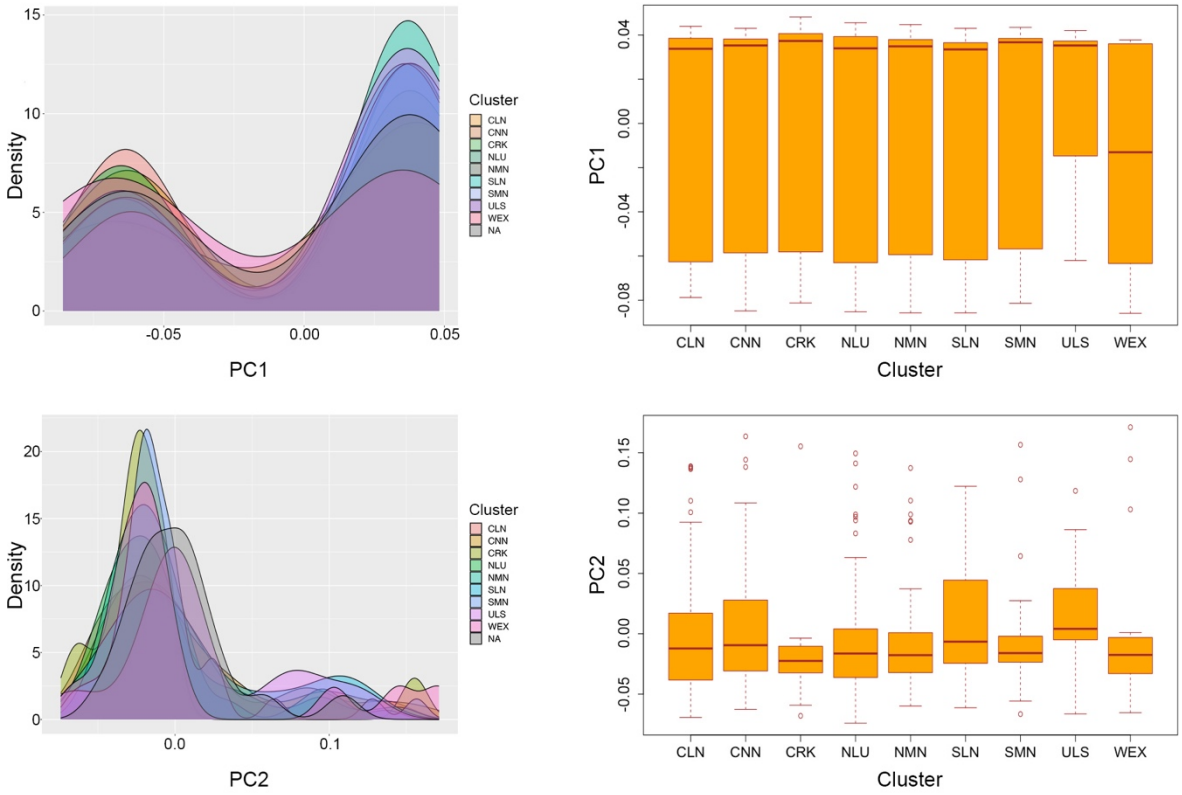


Figure 4.26, Characterisation of Irish genetic clusters along PC1 and PC2 (Private variant PCA). (Left) Density plots showing the distribution of individuals from each cluster along the PC1 and PC2 axis. (Right) Boxplots summarising the median and variation of PC1 and PC2 values for each cluster.

To identify the genomic driver of this clustering in the Irish dataset, PCA was performed on each chromosome separately and clustering was performed using mclust (Scrucca et al., 2016). These clusters were then compared to the whole-genome clusters using the adjusted Rand index (ARI), which measures the similarity between two sets of cluster assignments. The ARI scores for all chromosomal clusters (compared to the genomic cluster) are presented in Table 4.6. Chromosome 17 yielded a significantly higher correlation with the genomic clusters than the other chromosomes. This indicates that the clustering pattern observed in the genomic PCA is largely influenced by the archaic segments present on chromosome 17. The PCA plot of chromosome 17 further supports this finding, as it shows a distinct clustering pattern (Figure 4.27) (the PCA plots for the remaining chromosomes can be found in Appendix B (Figures B7-B17)).

Table 4.6, Concordance between chromosomal and genome-wide clusters.

Adjusted Rand Index Values

Chr1	0.0260	**Chr12**	-0.0102
Chr2	-0.0090	**Chr13**	-0.0061
Chr3	0.0142	**Chr14**	-0.0070
Chr4	0.0101	**Chr15**	0.0092
Chr5	-0.0083	**Chr16**	0.0069
Chr6	-0.0038	**Chr17**	0.5031
Chr7	-0.0019	**Chr18**	0.0369
Chr8	-0.0066	**Chr19**	-0.0145
Chr9	-0.0019	**Chr20**	-0.0015
Chr10	-0.0125	**Chr21**	-0.0044
Chr11	-0.0156	**Chr22**	-0.0086

This table provides adjusted Rand index values quantifying the similarity of clustering patterns inferred from individual chromosomes compared to the clustering based on the full genome.

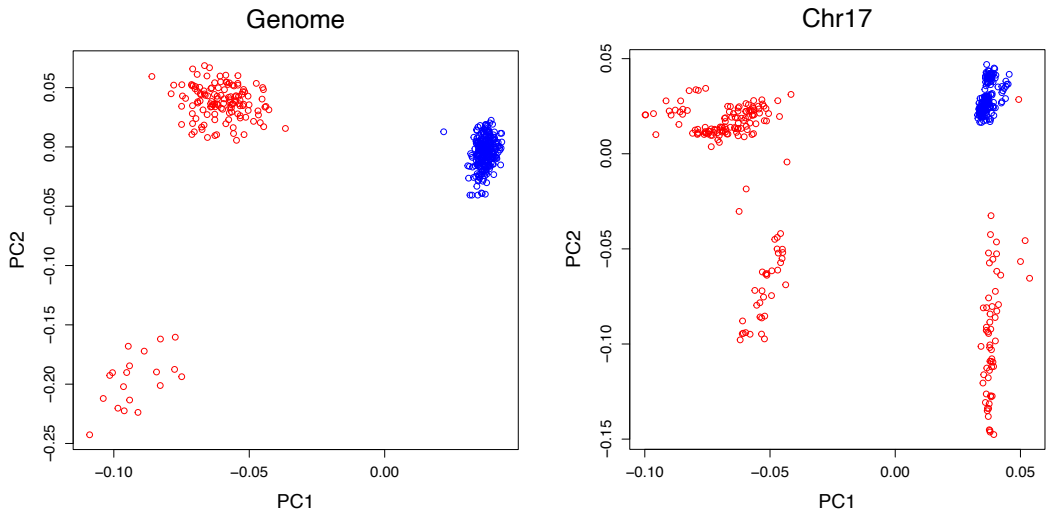


Figure 4.27, Cluster PCA plots. These plots show the distribution of individuals along PC1 and PC2 when the private variant PCA is conducted for the entire genome (left) and for chromosome 17 (right). Individuals are coloured according to their cluster assigned by mclust.

The specific regions on chromosome 17 (and potentially other chromosomes) that contribute to the observed clustering pattern in the PCA were examined. The divergence of the average PC1 value of individuals inferred as archaic from the midpoint of the two

clusters was calculated for each window. Windows with fewer than 20 archaic counts were excluded from the analysis to ensure robust results and avoid potential bias caused by a small number of individuals. The majority of windows showed a deviation towards the right cluster (i.e. greater average PC1 values than the midpoint), reflecting the larger number of individuals assigned to this cluster. However, the analysis of chromosome 17 revealed a 683 kb region with average PC1 values significantly lower than the midpoint (Figure 4.28). This further supports the previous clustering results and indicates that this specific region on chromosome 17 plays a prominent role in the separation of individuals based on PC1 values.

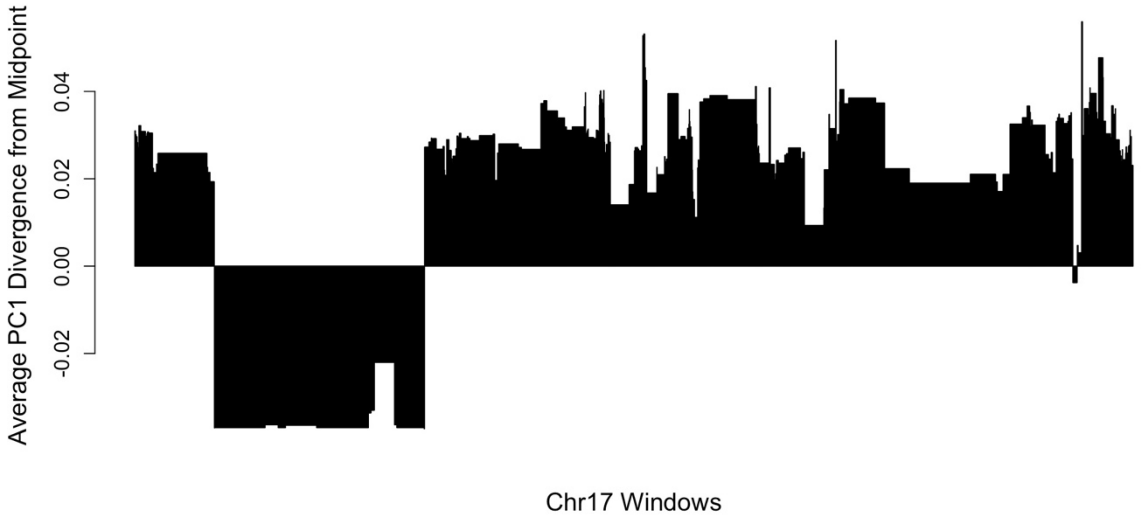


Figure 4.28, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 17 deviates from the central point between the two genomic clusters.

Finally, we conducted QC on this 683 kb region to investigate potential artifacts. No significant differences in GQ scores or inbreeding coefficients were found between this region and the rest of the genome (one-sided Wilcoxon rank sum test p-value = 1 for both tests, Figures B18-B38). However, a highly significant difference in AB was identified (one-sided Wilcoxon rank sum test, $p < 2.2 \times 10^{-16}$; Figure 4.29).

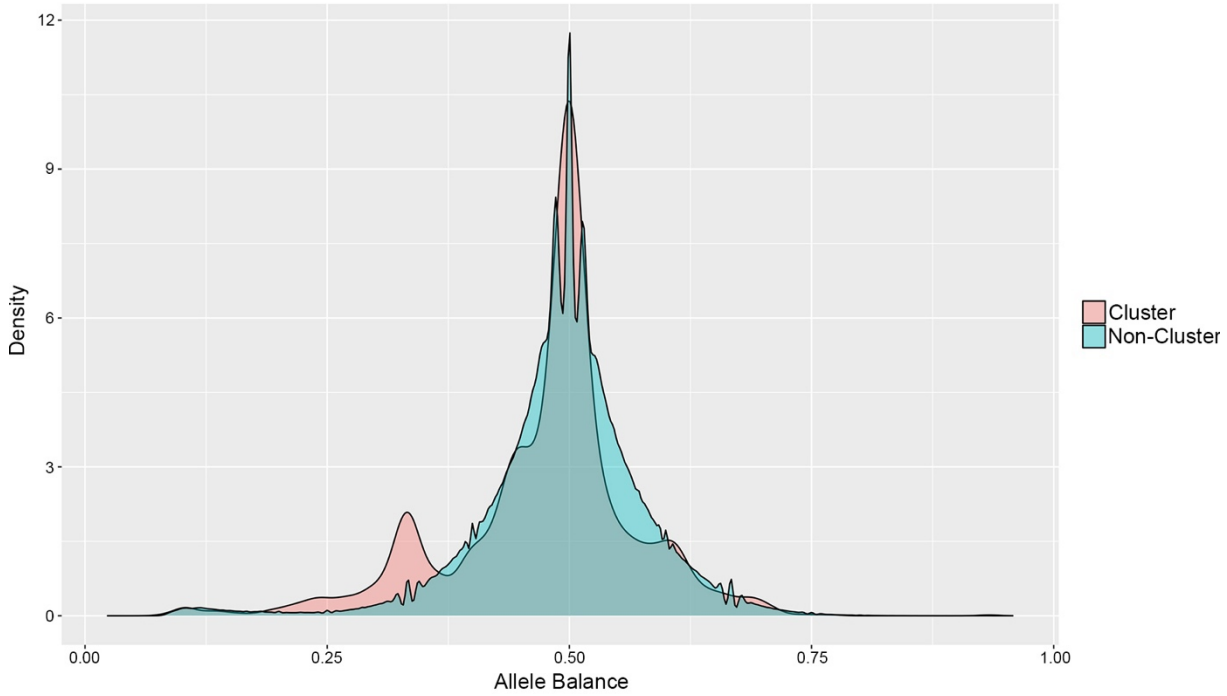


Figure 4.29, Comparison of AB for cluster and non-cluster private variants. The average AB for heterozygous genotypes is compared between private variants within the 683 kb region on chromosome 17 (Cluster) and those elsewhere in the genome (Non-cluster).

Subsequent investigation revealed that the 683 kb region on chromosome 17 (43,670,000-44,353,000) strongly overlaps with the well-studied 17q21.31 inversion polymorphism (approx. 43,703,801-44,212,416) (Stefansson et al., 2005; Steinberg et al., 2012). This locus is known to exist in two major, highly divergent haplotypes: a direct orientation (H1) and an inverted orientation (H2). These haplotypes do not recombine over nearly 2 Mb, creating an extended region of high LD. The near-absence of recombination between H1 and H2 haplotypes means that individuals sharing the same orientation should also share a large block of identical-by-descent alleles. The method used here to infer archaic introgression is based on the density of private variants, and would therefore interpret this large, shared, non-recombining haplotype block as a segment of archaic origin. Consequently, the strong clustering in our PCA is not an artifact of low-quality data but a direct result of this known structural polymorphism. Individuals cluster based on their 17q21.31 haplotype because they share an excess of apparent private variants across the entire region, a signature of the ancient

inversion and its associated LD. This is evident by the absence of clustering after the removal of this 683 kb region (Figure 4.30).

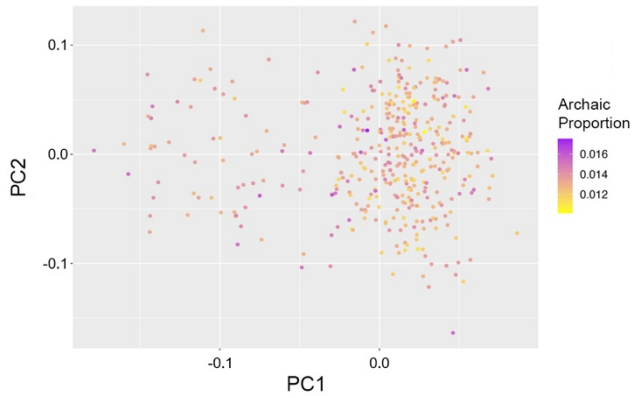


Figure 4.30, Private variant PCA plot excluding the inversion on chromosome 17. This plot shows the distribution of individuals along PC1 and PC2 for the private variant PCA excluding the 683 kb inversion region on chromosome 17. Individuals are coloured by their archaic proportion.

4.4 Discussion

This study undertook a detailed analysis of archaic introgression in a cohort of 408 Irish genomes, with two main focuses. Firstly, we wanted to rigorously validate and apply a HMM-based method for detecting introgressed segments, and secondly, we investigated the genomic landscape of this introgression to understand the evolutionary forces that have shaped it. We successfully implemented a robust pipeline that identifies and corrects for technical artifacts, producing a more reliable map of Neanderthal ancestry in the Irish population. We also identified regions of the genome which are entirely devoid of introgression, termed archaic deserts. However, a multi-faceted analysis of these deserts provided very little support for the predominant hypothesis that they are primarily the result of widespread selective purging of deleterious archaic alleles.

A central finding of this chapter is the profound impact of data quality on the inference of archaic segments. Our initial analysis revealed massive archaic peaks where every individual in the cohort was falsely inferred to carry an archaic segment. The convergence of two distinct methods (HMM and Sprime) on these peaks was a crucial clue that pointed not to a model error, but to a systematic artifact within the data itself. The subsequent investigation uncovered a signature of excess heterozygosity and aberrant quality metrics (low AB, GQ, and inbreeding coefficients) within these peaks, hallmarks of systematic genotyping errors or misalignment. The implementation of stringent filters targeting these specific anomalies successfully eliminated these false positives without removing true archaic signal, as evidenced by the stronger correlation with known archaic genomes in the filtered dataset.

This provides empirical support for a cautionary approach to the analysis of archaic sequence in modern genomes. Searching for archaic introgression often involves looking for rare

signal in a sea of noise, and the signals (regions of high divergence and unusual LD) can be mimicked by technical artifacts. Our results reinforce that rigorous quality control is an important part of the analysis. The improved normality of the archaic proportion distribution post-filtering further suggests that our refined dataset provides a more accurate representation of the true biological variation in archaic ancestry within this population.

Another dramatic finding of this study was the resolution of the strong clustering in our PCA of shared private variants. An apparently major demographic signal, which could potentially indicate subgroups within Ireland with different archaic ancestry histories, was almost entirely driven by a single 683 kb region on chromosome 17. This region overlapped almost perfectly with the 17q21.31 inversion polymorphism, a locus with a deep and complex evolutionary history involving two major haplotypes (H1 and H2) that do not recombine and carry distinct duplication architectures (Stefansson et al., 2005; Steinberg et al., 2012). This acts as a reminder that not all divergent haplotypes are introgressed. Deeply rooted ancestral structural variation, especially inversions that suppress recombination, can create identical patterns of high LD and elevated divergence. This highlights the importance of using a panel of archaic genomes to validate putative introgressed segments as well as the importance of masking known problematic regions like 17q21.31 during genome-wide scans.

A primary goal of this chapter was to assess whether archaic deserts display evidence of negative selection. The prevailing hypothesis states that regions of the genome that are highly constrained and intolerant of variation would have purged deleterious Neanderthal alleles, leaving behind deserts of introgression (Harris & Nielsen, 2016; Sankararaman et al., 2014). This model is supported by the well-documented underrepresentation of archaic ancestry in functional regions, particularly genes highly expressed in testes and on the X

chromosome (the “large X-effect”), which is interpreted as a signature of hybrid male sterility driven by the Dobzhansky-Muller incompatibility model (Sankararaman et al., 2014; Schaefer et al., 2021).

Our analysis of archaic deserts in the Irish genome yielded only limited evidence that purifying selection is the dominant force shaping the landscape of Neanderthal ancestry. The strongest evidence in favour of selection was the significantly higher gene density within the 38 overlapping deserts compared to matched random regions. This aligns with the expectation that functionally important genomic regions are enriched for genes and would be under stronger selective constraint, however there exist other reasons for clustering of gene-rich regions which could lend themselves to alternative explanations for correlation with archaic deserts. Moreover, subsequent analyses failed to build a compelling case for ongoing pervasive selection.

The analysis of functional constraint metrics yielded a paradoxical result. While we observed significantly lower o/e ratios for missense mutations in deserts and partial deserts, which hint at increased constraint, the most statistically significant result was for synonymous mutations. This is biologically difficult to interpret, as synonymous mutations are generally considered to be under only weak (if any) selective pressure. Conversely, the o/e ratios for loss-of-function mutations, the category most directly under intense purifying selection, showed no significant enrichment in deserts. This pattern is inconsistent with a simple model of strong, ongoing purifying selection against archaic haplotypes specifically in these regions. Similarly, the lack of significant enrichment or depletion of trait heritability within archaic deserts from the sLDSC analysis suggests that these deserts are not major reservoirs of common variant heritability for complex traits that were affected by introgression.

Another important result against the selection hypothesis came from comparing variant density in archaic genomes. If deserts are regions where Neanderthal DNA was intolerable to modern humans, it might be expected that these same regions would be more constrained and thus harbour fewer mutations in the Neanderthal genome itself. We found no such difference. The density of archaic SNPs was identical between desert and non-desert regions in the Altai, Vindija, and Denisovan genomes. This suggests that the evolutionary forces that created these deserts are specific to the modern human genetic background or to the process of introgression itself, rather than representing a universal and intrinsic property of these genomic regions that was shared with Neanderthals.

This leads directly to the premise of the next chapter. We will explore an alternative hypothesis: that the depletion of Neanderthal ancestry was driven by meiotic incompatibilities in archaic-human hybrids. Specifically, we will investigate whether archaic-derived alleles could have altered the landscape of meiotic recombination, promoting genomic rearrangements like deletions and duplications that would eventually purge archaic segments from the modern human genome. This process would create deserts of introgression not because the archaic segments themselves were deleterious, but because the genomic context they created was unstable during meiosis in a hybrid background.

5 Archaic SNPs as Drivers of NAHR in the Modern Human Genome

5.1 Introduction

5.1.1 Hybridisation

Hybridisation has increasingly been recognised as a pervasive evolutionary force, which contributes not only to the generation of genetic diversity but also to reproductive isolation and speciation (Mallet, 2005; Seehausen, 2004). Hybridisation introduces new alleles into a population's gene pool. This can facilitate adaptation when introgressed alleles provide beneficial traits, such as pesticide resistance in *Heliconius* butterflies (Dasmahapatra et al., 2007), altitude adaptation in Tibetans via Denisovan introgression (Huerta-Sánchez et al., 2014), and immune gene variation in humans through Neanderthal and Denisovan introgression (Quach et al., 2016). On the other hand, hybridisation can also introduce deleterious alleles or generate incompatibilities, which can lead to reduced hybrid fitness (Maheshwari & Barbash, 2011; Presgraves, 2010).

A central theory within the context of hybridisation is the Bateson-Dobzhansky-Muller incompatibility (BDMI) model, whereby independent substitutions in diverging lineages interact negatively when brought together in hybrids (Bateson, 1909; Dobzhansky, 1934; Muller, 1942). Such incompatibilities can result in hybrid sterility, inviability, or reduced fertility (Sweigart & Willis, 2012). In mammals, one of the clearest examples is hybrid male sterility in crosses between house mouse subspecies (*Mus musculus domesticus* and *M. m.*

musculus), which is often linked to disruptions in meiotic chromosome pairing and recombination (Larson et al., 2018; Mihola et al., 2009).

Hybridisation load is another important concept in hybrids, where selection acts to reduce ancestry of the parental species with the lower effective population size in hybrids, due to their reduced fitness (Juric et al., 2016a; Schumer et al., 2018). Across diverse taxa, hybrid genomes bear this signature of selection, with ancestry from the minor parent species often reduced near functionally important elements (Hvala et al., 2018; Martin & Jiggins, 2017; Schumer et al., 2018). In swordtail fish hybrids, Schumer et al. (2018) demonstrated that ancestry from the minor parental species persists disproportionately in regions of high recombination, where deleterious alleles can be more rapidly uncoupled from neutral or beneficial neighbours.

5.1.2 Recombination in Hybrids

While the mechanisms of meiosis and recombination in uniancestral populations are well-established, hybridisation between divergent populations or species introduces special challenges. Hybrid genomes may carry divergent binding motifs, mismatches in regulatory machinery, chromosomal rearrangements, or structural divergence that affect how recombination works. The initial stage of meiosis is the alignment of homologous chromosomes, which can be compromised by genetic divergence, leading to widespread failure.

Hybrid meiocytes face a fundamental challenge in distinguishing true homologs (from the same parental species) and homoeologs (similar chromosomes from different species). The machinery for homology search involves proteins like the HORMA-domain protein HOP1

in mammals and is sensitive to even modest levels of sequence divergence. Studies in *Mus musculus x M. spretus* hybrids have shown that sequence divergence of ~6-8% is sufficient to cause significant delays, errors in pairing, and extensive asynapsis (failure to pair) or non-homologous synapsis (Bhattacharyya et al., 2013; Davies et al., 2016; Gregorova et al., 2018). Unsynapsed chromosomes are recognised by the pachytene checkpoint, a surveillance mechanism that detects aberrations and induces apoptosis of the aberrant meiocyte. This checkpoint-mediated cell death is a direct cause of hybrid sterility, as it drastically reduces the number of viable gametes produced (Bhattacharyya et al., 2013). The strength of this response varies between sexes, often explaining Haldane's rule where the heterogametic sex is more likely to be sterile in hybrids (Coyne, 1985; Schilthuizen et al., 2011). In XY males, for example, the divergence of the X and Y chromosomes and the challenges of X-chromosome inactivation and reactivation during meiosis make the process particularly vulnerable in hybrids. Furthermore, the X chromosome often shows disproportionately high levels of divergence and gene regulation issues in a hybrid background, thereby exacerbating meiotic defects (Lifschytz & Lindsley, 1972; Presgraves, 2008).

PRDM9 is one of the few well-characterised speciation genes in vertebrates. Mihola et al. (2009) first identified *PRDM9* as a hybrid sterility gene, showing that particular alleles are capable of influencing fertility in male hybrids of mice. The rapid evolution of *PRDM9*'s zinc finger domain and the erosion of its binding motifs mean that in an F1 hybrid, the *PRDM9* allele from one parent may not recognise its preferred binding sites in the divergent genome of the other parent. Symmetric binding of PRDM9 motifs (i.e. both parental chromosomes share matching binding motifs so that hotspots are equally recognised) is important in hybrids. Davies et al. (2016) humanised one allele of *PRDM9* in a mouse so

that the degree of symmetric binding increased, which restored fertility in otherwise sterile hybrid males. This demonstrates that fertility largely depends on whether hotspots are recognised symmetrically across homologs.

In rice (subspecies *indica* vs *japonica*), a comparative study showed that ~80% of crossover hotspots are not shared between subspecies despite their relatively recent divergence. This indicates rapid hotspot turnover and suggests a role for crossover hotspots as potentially powerful drivers of sequence and subspecies evolution. In ruminants, *PRDM9* shows multiple duplications and strong positive selection in its zinc finger arrays, reflecting rapid evolution of binding specificity. Variation in *PRDM9* among lineages suggests that divergence among *PRDM9* alleles predates speciation events and can potentially contribute to hybrid incompatibilities if alleles are mismatched in hybrids (Padhi et al., 2017). Furthermore, AbuAlia et al. (2024) identified many *PRDM9* alleles in wild mouse subspecies and showed that certain alleles associated with sterility in lab hybrids also operate in wild contexts, with variable fertility depending on the degree of divergence.

The modern human and Neanderthal lineages accumulated differences in both the *PRDM9* gene itself and the genomic landscape of its binding motifs. Sequencing of archaic genomes has revealed that Neanderthals and Denisovans carried distinct *PRDM9* alleles different from those found in modern humans (Schwartz et al., 2014). The divergence of the subspecies also allowed for the independent erosion of PRDM9 binding motifs in each lineage; motifs recognised by one lineage's PRDM9 variant were likely degraded by genetic drift in the other, and vice versa (Lesecque et al., 2014). This would have created an asymmetry, such that the PRDM9 protein from a modern human parent in a hybrid would bind poorly to the

Neanderthal-derived chromosome, which would have a different set of intact motifs, and vice versa (Davies et al., 2016).

5.1.3 Implications for Human-Neanderthal Hybrids

While hybridisation between humans and archaic hominins such as Neanderthals and Denisovans was clearly viable, as evidenced by the legacy of their DNA in modern human populations, the meiotic consequences of merging these long-diverged genomes may have imposed a fitness cost. The ~500,000 years of separation between the modern human and Neanderthal lineages (Prüfer et al., 2014) would have offered an opportunity for the accumulation of genetic variants that may be neutral within their respective lineages but would become deleterious in a hybrid genomic context. This could include a vast number of mutations within PRDM9 binding motifs that would be functionally classified as gain- or loss-causing SNPs from the perspective of the modern human genome and PRDM9 protein. This mismatch in motif composition between archaic and modern human haplotypes could have increased the frequency of asymmetric and mismatched motif pairs during meiosis in hybrids.

We propose here that human-archaic hybrids experienced an elevated burden of *de novo* CNVs, and that archaic introgressed segments may still promote NAHR in the modern human genome today (Figure 5.1). Archaic segments are potential NAHR substrates as their homology to the human genome is sufficient for the recombination machinery to engage, yet potentially divergent enough to promote mispairing with paralogous modern human sequences. This model of persistent meiotic instability could provide an explanation for the well-documented depletion of Neanderthal ancestry in the modern human genome (Jégou et al., 2017; Sankararaman et al., 2014, 2016; Wei et al., 2022). Even if a Neanderthal

haplotype contained no deleterious alleles in isolation, its mere presence could be detrimental if it increased the likelihood of a NAHR event during meiosis. The recurrence of NAHR would act as a potent force for the reduction and eventual removal of Neanderthal segments from the population gene pool.

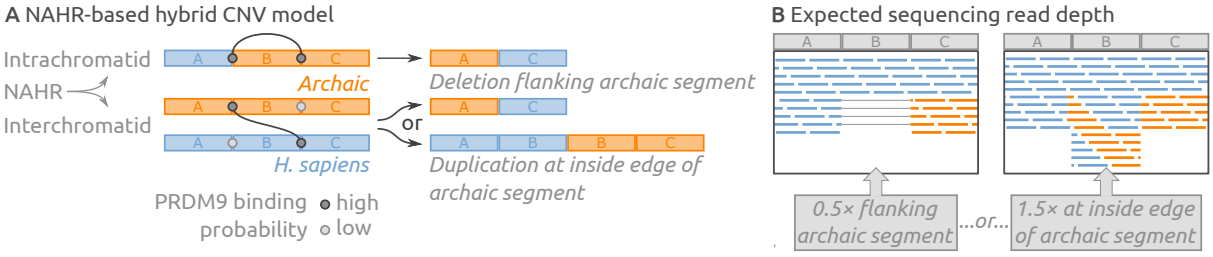


Figure 5.1, NAHR-based hybrid CNV model. A) Schematic representation of the hypothesised mechanism. Divergence between modern human and archaic Neanderthal/Denisovan genomes alters PRDM9 binding motifs and their affinity. This creates a scenario where a genomic region may contain high-affinity PRDM9 binding sites in the archaic haplotype (orange) but not in the modern human haplotype (blue), or vice versa. During meiosis in a hybrid individual, this asymmetry can facilitate NAHR between mispaired high-affinity sites. Intra-chromatid NAHR results in a deletion, while inter-chromatid NAHR produces both a deletion and a reciprocal duplication. B) Expected genomic signature of the proposed NAHR event. This is predicted to generate a haplotype where the introgressed archaic segment is precisely flanked by a deletion. In the case of inter-chromatid NAHR, this deletion is coupled with a duplication at the internal junction of the remaining archaic segment.

NAHR can occur via three pathways: inter-chromosomal (between homologous chromosomes), inter-chromatid (between sister chromatids), and intra-chromatid (within a single chromatid) (Figure 5.1). While inter-chromosomal and inter-chromatid NAHR can generate both deletions and duplications, intra-chromatid NAHR exclusively produces deletions (Liu et al., 2012). This mechanism involves the looping out and excision of a sequence between two direct repeats on the same chromatid. Additionally, Turner et al. (2008) observed that deletions are generated at a higher rate than their reciprocal duplications in the male germline. Both of these factors may establish a systematic bias towards deletion events. Over time, this bias would result in a slow, steady erosion of introgressed archaic sequence, as deletion predominates over its duplication counterpart. This bias towards loss presents a potential explanation for the depletion of Neanderthal ancestry, which likely acts

together with classical negative selection to remove introgressed segments from the modern human genome.

5.1.4 Research Aims

This chapter investigates the hypothesis that the divergent genomic landscapes of modern humans and Neanderthals, particularly at PRDM9 binding motifs, created a sustained risk of NAHR in hybrid offspring. We propose that this mechanism not only contributed to a fitness cost in ancient hybrids but continues to drive SV and the erosion of archaic ancestry in contemporary human populations. Using a combination of population genomics datasets, we test for several genomic signatures of this process, including an enrichment of CNV within archaic segments, a propensity for deletions to occur specifically at the boundaries of archaic segments, a population-level depletion of archaic ancestry in regions of high meiotic recombination, and evidence for the ongoing loss of archaic sequence over a single generation. Together, these analyses aim to establish NAHR as an underappreciated force shaping the fate of introgressed sequences in the modern human genome.

5.2 Materials and Methods

5.2.1 Private Variant Analysis

Archaic segments for all individuals from the 1000G Trio, HGDP, and HGDP+1000G datasets were inferred according to the method described in section 4.2.1. All private variants, defined here as a variant not present in the outgroup, from each individual were categorised based on their genomic context: those located within an inferred archaic segment of the individual (PVAS) versus those in non-archaic regions (PV). For each variant, two key metrics were calculated: AB, representing the ratio of alternate to total reads, and relative RD, normalised against individual-specific baselines derived from common SNPs represented on the Infinium Global Screening Array (Illumina, Inc.). This normalisation accounted for technical variation in sequencing depth across samples.

A 50x50 matrix structure was employed where each bin represented the two-dimensional distribution space of AB versus relative RD. For each bin, the fractional difference between PV and PVAS distributions was computed (i.e. the proportion of PVs minus the proportion of PVASs occupying that particular AB-RD combination). This allowed the detection of systematic differences in variant distributions between archaic and non-archaic segments. An average matrix was created of all individual matrices in each dataset, which was then visualised as a heatmap. This analysis was performed separately on the HGDP, HGDP+1000G, 1000G Trio, and Irish datasets.

5.2.2 Deletions Flanking Archaic Segments

5.2.2.1 Relative Read Depth

To test whether deletions are preferentially found flanking archaic segments in the modern human genome (consistent with the model in Figure 5.1), the edges of all archaic segments were examined. Because the precise start and end coordinates of introgressed segments are difficult to define with high accuracy, we extracted 10 kb regions immediately upstream (10 kb before the start) and downstream (10 kb after the end) of each archaic segment. For each individual, an average RD was calculated from 10,000 randomly selected biallelic SNPs from callable genomic regions. This average was used to calculate the relative RD of all biallelic SNPs in each 10 kb region. The average relative RD was then calculated for each region.

To generate a null expectation, the same number of random callable 10 kb windows was sampled from each individual and their relative RD also calculated. This analysis was conducted separately in the Irish, HGDP, and 1000G Trio datasets. To assess robustness to archaic boundary uncertainty, the analysis was repeated with flanking windows defined at increasing distances away from the archaic segment edges (e.g. 10 kb windows starting 5 kb or 10 kb away from the segment boundary).

5.2.2.2 Deletions

To further investigate whether regions immediately adjacent to archaic haplotypes are enriched for deletions, we analysed the overlap between genomic regions flanking archaic segments and individual-specific deletions. This analysis was conducted using the 1000G Trio dataset due to the availability of a confident, curated CNV callset (section 2). The 10

kb windows flanking archaic segments and the corresponding 10 kb random regions from section 5.2.2.1 were used. For each individual, the proportion of their archaic-flanking 10 kb windows that overlapped at least one of their called deletions was calculated and compared to the proportion of their random 10 kb windows that overlapped one of their deletions. Statistical significance of the within-individual differences was evaluated using the Wilcoxon signed-rank test, which is appropriate for paired comparisons of distributions that are not necessarily normally distributed.

5.2.2.3 Deletion Breakpoints vs Archaic Proportion

The individual-specific deletions from the 1000G Trio dataset used in section 5.2.2.2 were also utilised for this analysis. For every called deletion per individual, a 2 kb region (comprising 1 kb upstream and 1 kb downstream of each breakpoint) was defined. The total proportion of these breakpoint-associated regions that overlapped with any of the individual's own archaic segments was computed. This metric represents the observed propensity for deletions to occur near archaic segments. The baseline expectation was defined as the individual's total archaic proportion, calculated as the sum of the lengths of all their archaic segments divided by their total callable genome length. This represents the proportion of the genome where an overlap would be expected by random chance. The statistical significance of the difference between the breakpoint overlap proportion and the archaic proportion for each individual was assessed using the Wilcoxon signed-rank test. This paired non-parametric test evaluates whether the breakpoint overlap proportion is significantly different than the genomic background expectation across the entire dataset.

5.2.2.4 Read Depth Profile across Archaic Segment Boundaries

This analysis utilised the 1000G Trio and the Irish dataset. For each dataset, all archaic segments from all individuals were first collected. For every segment, the set of variants within a ±20 kb window from both the start and end positions of the segment was retrieved. These SNPs were annotated with their distance to the segment boundary. For the start position of an archaic segment, variants downstream (inside the archaic segment) were given negative distances, while those upstream (outside) were given positive distances. For the end position, variants upstream (inside the segment) were given negative distances, and downstream variants (outside) were given positive distances. This ensured that negative distances always pointed into the archaic segment, and positive distances always pointed away from it.

For each SNP within the ±20 kb windows, its relative RD was calculated as the ratio of the SNP's own RD to the average RD of all SNPs within the associated archaic segment itself. This controlled for local variation and ensured that comparisons focused on relative differences near segment boundaries. To visualise trends, the relative RD values were aggregated across all archaic segments and plotted as a rolling average with a window size of 1 kb and a step size of 1 bp. This smoothing allowed us to observe subtle patterns at segment boundaries. To statistically assess the difference in RD patterns on either side of the boundary, a non-parametric Wilcoxon rank sum test was performed to compare the distribution of relative RD values of SNPs with negative distances (toward the segment) against those with positive distances (away from the segment).

5.2.3 Archaic Segment Depletion in the Modern Human Genome

5.2.3.1 Archaic Deserts

To test whether regions with high recombination rate are depleted of archaic introgressed segments, we analysed the overlap between archaic deserts and DSB hotspots. These analyses were conducted separately for three datasets: HGDP+1000G, Irish and Icelandic cohorts. For each dataset, all archaic deserts were intersected with the full set of DSB hotspots in the matching genome build, and the total overlap was recorded. To assess significance, a null distribution was generated by producing 1,000 random sets of regions matched in length to the DSB hotspots. These random sets were generated separately for GRCh37 and GRCh38. The overlap between these random sets and archaic deserts was calculated in the same way as for the real DSB hotspots. To account for mapping and callability biases, the analysis was then refined to include only DSB hotspots that fell entirely within callable regions of the genome. In these refined analyses, the random regions were also restricted to callable regions.

5.2.3.2 Recombination Rate

To investigate the relationship between meiotic recombination activity (independent of whether a region is defined as a hotspot) and the persistence of archaic genomic segments, paternal crossover rates were obtained from Halldorsson et al. (2019) and analysed. The genome was separated into 1 kb windows and the mean recombination rate (in centiMorgans (cM) per kb) of each window was calculated. This was then compared to the archaic counts within each dataset (HGDP+1000G, Irish, and Icelandic datasets). Because the recombination rate map was originally in GRCh38, the coordinates were converted to

GRCh37 for the Irish dataset using the UCSC Genome Browser LiftOver tool. The association between recombination rate and archaic count was assessed for each dataset by calculating Spearman's rank correlation coefficient. This non-parametric test was chosen to account for the highly skewed distributions of both recombination rates and archaic counts.

Given the very large number of windows genome-wide and the broad distribution of recombination rates and archaic counts, the windows were separated into two groups: low recombination windows (< 0.1 cM/kb) and high recombination windows (≥ 0.1 cM/kb). The distributions of archaic counts between these two groups were compared statistically using the Wilcoxon rank sum test.

5.2.3.3 Archaic Proportion Depletion

The proportion of archaic sequence in each individual's autosomes was estimated for the 1000G Trio dataset. For each individual, total archaic sequence length was divided by the summed length of the callable autosomes (1.997 Gb) to produce the autosomal archaic proportion. For each trio, the mean autosomal archaic proportion of both parents was calculated and compared to that of the child. This enabled detection of potential changes in archaic sequence content, which could indicate ongoing loss of archaic sequence from the modern human genome. Differences between child and parental averages were assessed using a paired t-test, which evaluated whether observed differences deviated significantly from zero.

5.3 Results

5.3.1 Private Variant Analysis

Given that tens of thousands of private variants were identified per individual, with only a subset actually becoming incorporated into archaic segments, all private variants were separated into two groups per individual: those in archaic segments (PVAS) and those not in archaic segments (PV). The latter group served as a control, establishing baseline patterns for private variants, while deviations observed in PVAS could reflect unique characteristics associated with archaic introgression. AB and relative RD were used as primary metrics to determine whether PVAS exhibit signatures of CNVs more frequently than PVs, with AB ~0.33 and relative RD ~1.5 serving as indicative thresholds for duplications.

For each individual, a 50x50 matrix was constructed based on these two metrics, with each bin representing the fractional difference between PV and PVAS distributions (i.e. the proportion of PVs in a given bin minus the proportion of PVASs in the same bin). These matrices were combined across all individuals and the average was visualised as a heatmap (Figure 5.2 and Figure 5.3, Irish dataset), which revealed an excess of PVAS in the duplication-associated region (AB ~0.33 and relative RD ~1.5). Notably, this excess extended to another region with AB ~ 0.5 and relative RD ~2, which may represent segments duplicated twice. Both of these observations indicate an excess of duplications within archaic segments compared to non-archaic regions.



Figure 5.2, Duplication-screening heatmap of private variants in the Irish individuals. Averaged across all Irish individuals, this heatmap shows the differences between PV and PVAS, where red represents an excess of PV and blue represents an excess of PVAS. Only bins with an AB between 0.1 and 0.9 were considered to focus on heterozygous genotypes.

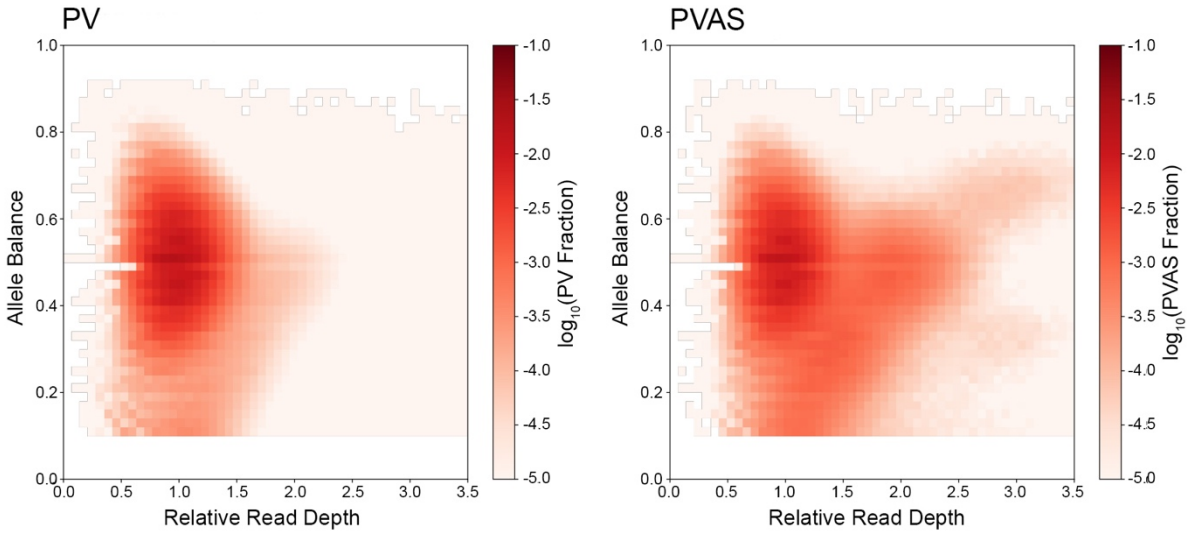


Figure 5.3, Duplication-screening heatmaps for PV and PVAS. These heatmaps display the distribution of variants in the Irish dataset along relative RD and AB for the two categories (PV (left) and PVAS (right)).

This analysis was then carried out on three additional datasets. Initial examination of the HGDP and HGDP+1000G datasets showed that their utility for this analysis was compromised due to prior filtering steps. The HGDP dataset had undergone relative RD-based filtering (section 2) while HGDP+1000G had been processed with AB filtering (section 2). Although these preprocessing steps likely confounded the analysis, these results have been included in Appendix C (Figures C1 and C2).

The 1000G Trio dataset, which had undergone less restrictive variant filtration, produced similar results to the Irish dataset (Figure 5.4). An excess of PVAS in the duplication-associated region (characterised by AB ~0.33 and relative RD ~1.5) was observed, though this feature was much more subtle. This persistent signal across independent datasets strengthens the hypothesis that archaic segments may be associated with CNV formation. The heatmap revealed an additional distinctive pattern: PVASs demonstrated lower relative RD compared to matched PVs. This finding may represent mapping challenges arising from sequence divergence between archaic and modern human references.

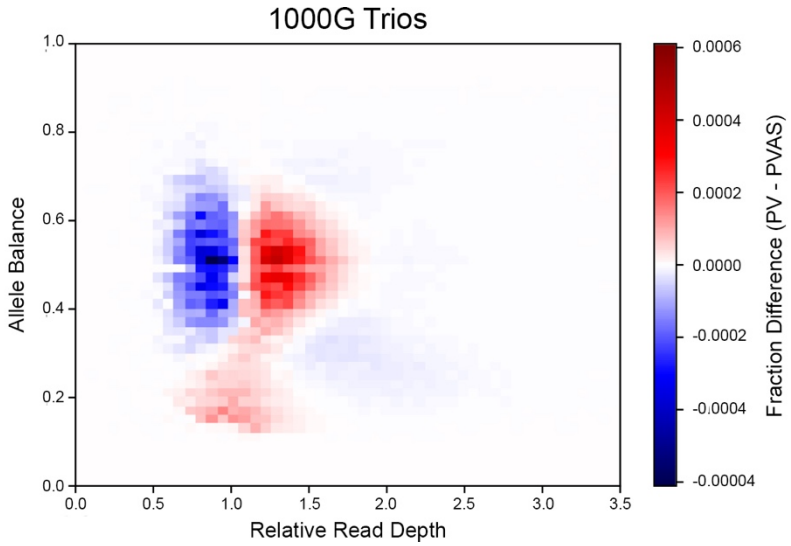


Figure 5.4, Duplication-screening heatmap of private variants in the 1000G Trio dataset. Averaged across all parents in the 1000G Trio dataset, this heatmap shows the differences between PV and PVAS, where red represents an excess of PV and blue represents an excess of PVAS. Only bins with an AB between 0.1 and 0.9 were considered to focus on heterozygous genotypes.

5.3.2 Deletions Flanking Archaic Segments

5.3.2.1 Relative Read Depth

Analysis of the Irish dataset revealed a significant excess of 10 kb windows flanking archaic segments with reduced relative RD (~0.5) compared to matched random regions (Figure 5.5). This enrichment supports the hypothesis that deletions are more likely to occur at the boundaries of archaic segments than expected by chance. Such a pattern is consistent with the hypothesis that divergent PRDM9 binding landscapes between humans and Neanderthals can create risk diplotypes when combined and thus facilitate NAHR, leading to an excess of deletions flanking archaic segments in the modern human genome.

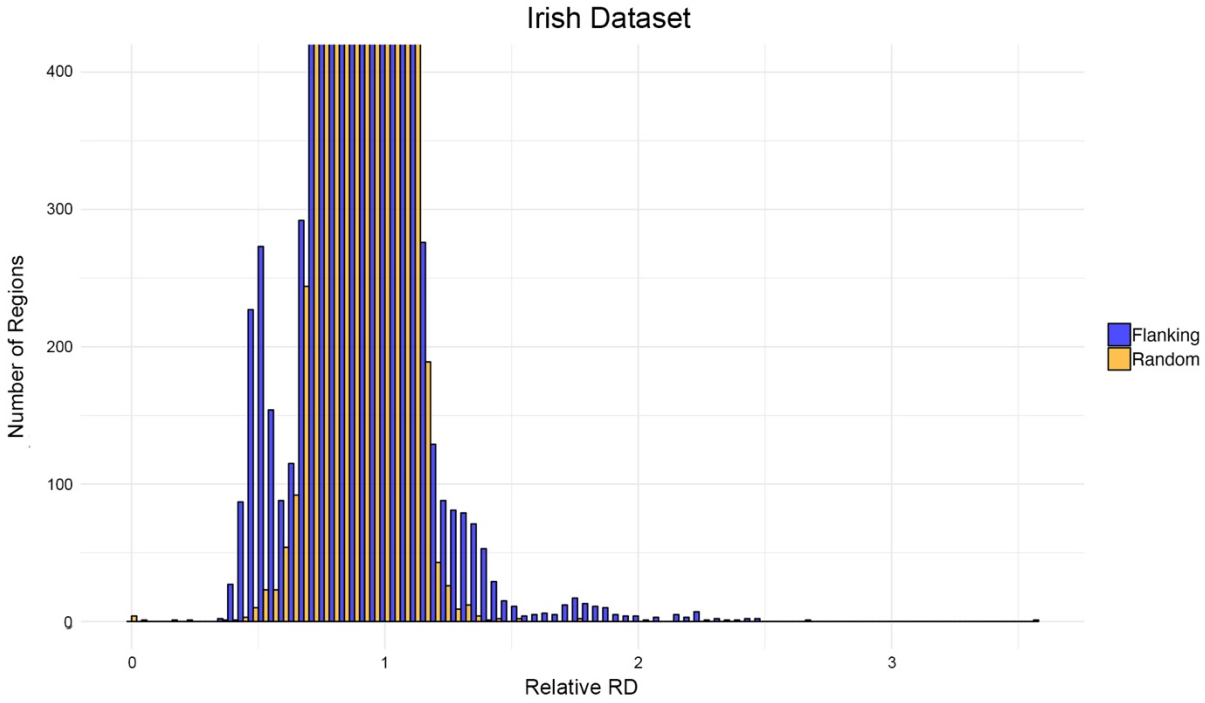


Figure 5.5, Deletion signatures flanking archaic segments. The distribution of relative RD in 10 kb regions flanking archaic segments is compared to matched random control regions across all Irish individuals.

However, this signal was not uniform across all datasets. While the 1000G Trio dataset also displayed a mild excess of deletion-like RD signals relative to random controls (Figure 5.6), the HGDP dataset exhibited no significant difference between flanking and random regions

(Figure C3). There are several factors that could explain this discrepancy. As outlined in section 2, each of the datasets underwent their own unique filtering pipeline. While the filtering steps of the Irish and 1000G Trio datasets were almost identical, the HGDP dataset had been filtered previously. This, along with population-specific differences, may have caused the different patterns observed. In particular, the Irish population's historical bottlenecks and prolonged geographical isolation may have permitted the persistence of rare SVs through genetic drift as well as archaic SNPs that drive NAHR by modifying the PRDM9 landscape.

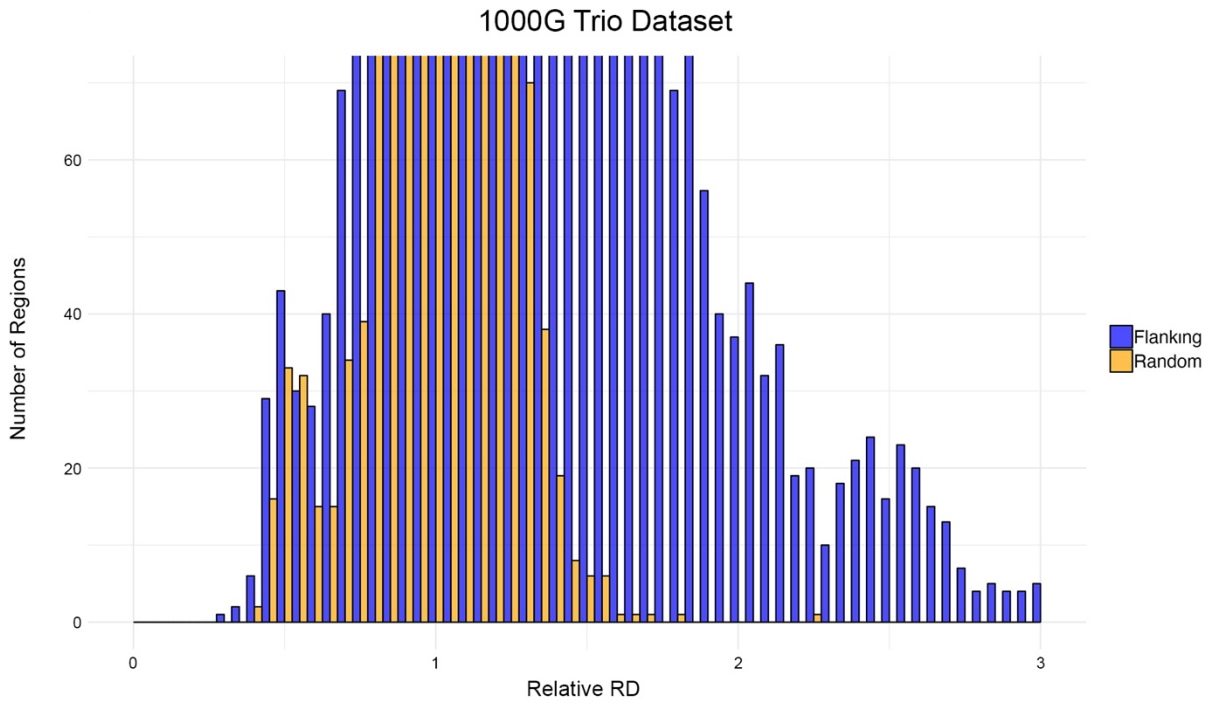


Figure 5.6, Deletion signatures flanking archaic segments. The distribution of relative RD in 10 kb regions flanking archaic segments is compared to matched random control regions across all parents in the 1000G Trio dataset.

In addition to the depletion signal, an excess of flanking regions with elevated relative RD (≥1.5) was observed in the Irish and 1000G Trio datasets, indicative of apparent duplications. This may reflect inaccuracies in archaic segment boundary definition, as true segment edges may extend beyond the defined coordinates. If true, these elevated RD regions may partially capture the previously reported enrichment of duplications within archaic segments.

However, increasing the distance between the segment edges and the flanking regions did not strengthen the deletion signal (Figure C4-C7).

These results provide partial support for the role of PRDM9 mismatch in mediating structural variation during human-Neanderthal hybridisation. Although they do not provide conclusive evidence, the localised depletion of RD at segment boundaries is consistent with NAHR-driven deletions arising from ectopic recombination between divergent PRDM9 binding sites. The inconsistency across datasets highlights the challenge of detecting ancient SVs and highlights the need for long-read sequencing and improved annotation of introgressed segments.

5.3.2.2 Deletions

Across the 200 1000G Trio individuals analysed, the majority displayed an enrichment of deletions in archaic-flanking regions relative to random controls ($p = 4.58 \times 10^{-20}$, binomial test). Specifically, 163 out of the 200 individuals had more of their 10 kb flanking windows overlapping deletions compared to their corresponding random windows. Among the remainder, 11 individuals showed an equal number of overlaps in flanking and random sets, while 26 had more random regions overlapping deletions than flanking regions.

On average, the number of archaic-flanking regions overlapping deletions per individual was 42.99, compared to 35.32 for random regions. This difference is visualised in Figure 5.7, which shows the proportion of overlaps per individual across both categories. A Wilcoxon signed-rank test confirmed that the number of flanking regions overlapping deletions was significantly greater than expected by chance ($p = 5.0 \times 10^{-26}$).

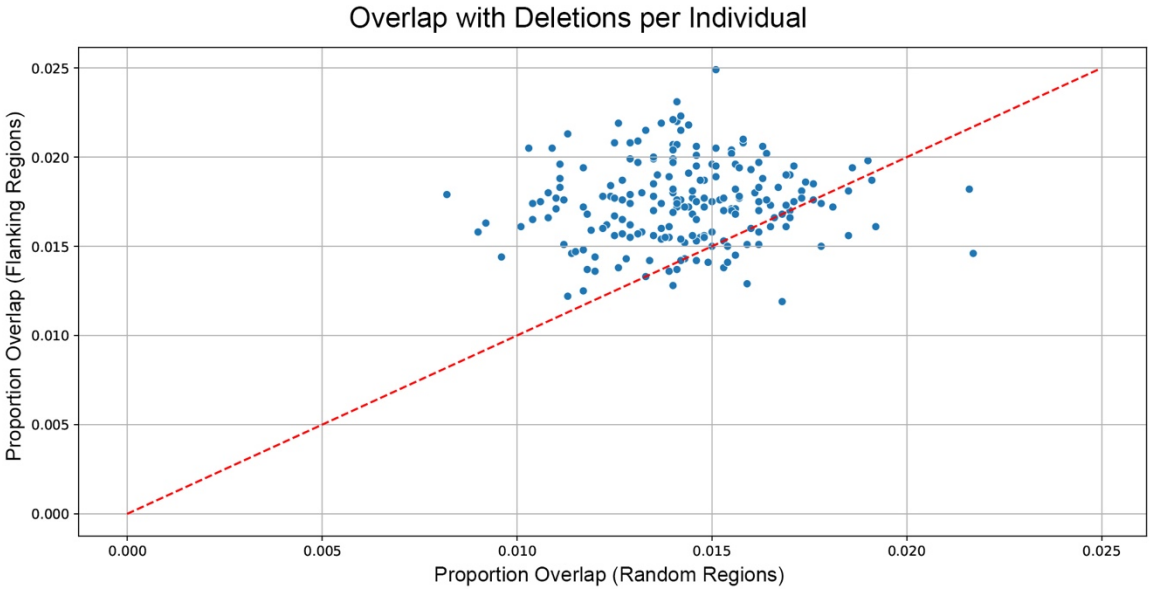


Figure 5.7, Enrichment of deletions flanking archaic segments. Each point represents an individual (parents in the 1000G Trio dataset), comparing the proportion of deletions in their archaic-flanking regions (y-axis) to matched random regions (x-axis). The dashed 1:1 line indicates equal proportion in both sets.

These findings reinforce the idea that archaic introgressed haplotypes predispose neighbouring sequence to deletions. Not only did flanking archaic windows exhibit lower relative RD (as described in section 5.3.2.1), but they also harboured more deletions than expected from randomly sampled regions. This pattern is consistent with the hypothesis that archaic sequence promotes NAHR in surrounding genomic regions, thereby generating deletions adjacent to the introgressed haplotype.

5.3.2.3 Deletion Breakpoints vs Archaic Proportion

This analysis revealed a significant enrichment of archaic segments in the immediate vicinity of deletion breakpoints. In the majority of individuals (174 out of 200), the proportion of deletion breakpoint regions (±1 kb) overlapping with an archaic segment was higher than the individual's overall genomic proportion of archaic ancestry ($p = 4.44 \times 10^{-28}$, binomial test; Figure 5.8). This was confirmed at the population level by a Wilcoxon signed-rank test, which showed that the breakpoint overlap proportion was significantly greater than the

genomic archaic proportion across the cohort ($p < 2.2 \times 10^{-16}$). This finding compliments the previous analyses, providing evidence that deletions are not randomly distributed but are often associated with the boundaries of archaic segments. Although alternative explanations may exist to explain the observed enrichment, this pattern is precisely the expected genomic signature of NAHR driven by the divergent PRDM9 landscapes of humans and Neanderthals, as outlined in Figure 5.1.

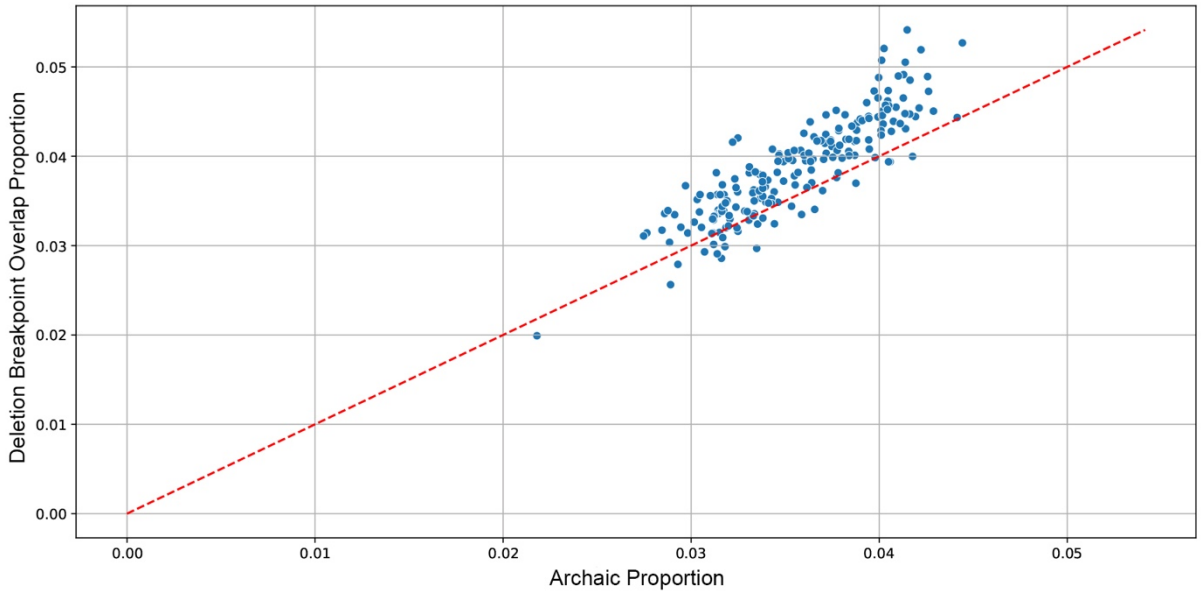


Figure 5.8, Archaic ancestry correlation with deletion breakpoint location. For each parent in the 1000G Trio dataset, the proportion of their deletion breakpoint regions (±1 kb) overlapping an archaic segment is plotted against their total proportion of archaic ancestry. The dashed 1:1 line indicates equal proportion in both sets.

5.3.2.4 Read Depth Profile across Archaic Segment Boundaries

To investigate fluctuations in RD in regions flanking archaic segments, we first established a standardised definition of relative RD. Initial investigations revealed that the RD of private variants within archaic segments was generally higher than that of a common reference SNP set (Infinium Global Screening Array, Illumina, Inc.). However, analysis of alignment depth from CRAM files showed that coverage in archaic segments was slightly lower than in

callable regions overall (Figures C8 and C9). These results highlighted the need to focus on relative RD normalised by local segment averages, which was calculated as the ratio of a SNP's RD to the mean RD of the archaic segment it flanks.

This method was applied to all SNPs within ±20 kb of archaic segment boundaries in the 1000G Trio dataset, which revealed several significant patterns (Figure 5.9). Within the archaic segment, approaching its edge, there was an increase in average relative RD. This suggests that duplications are enriched within archaic segments, which is consistent with the proposed hypothesis. The SNPs closest to the boundary are also the most likely to fall within archaic segments. For segments shorter than 20 kb, SNPs far from the boundary (e.g. -20 kb) are already outside the segment. Thus, the strong increase near zero reflects the higher duplication signal specifically inside archaic sequence. Another striking finding was the significant drop in relative RD at the outside edge of the boundary. This indicates an excess of deletions in the flanking regions, potentially reflecting NAHR events between mismatched PRDM9 landscapes in archaic versus modern haplotypes.

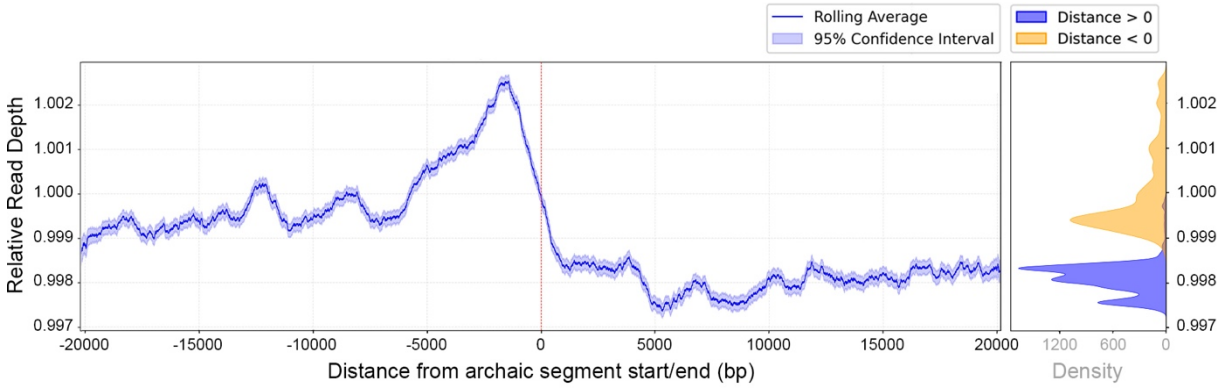


Figure 5.9, Relative RD profile across archaic segment boundaries. This plot shows the rolling average (1 kb window) of relative RD calculated for variants within ±20 kb of all archaic segment start and end positions from the 1000G Trio dataset. Distance 0 represents the segment boundary; negative distances are inside the segment, positive distances are outside. The density plot shows the distribution of relative RD for variants inside (distance < 0) and outside (distance > 0) the segments.

Together, these patterns create a distinct profile. They show that relative RD rises within archaic sequence, then falls sharply outside it. This signal provides genome-wide evidence that archaic haplotypes are associated with duplications within the introgressed regions, as well as with deletions in their immediate flanking regions. This pattern was also replicated in the Irish dataset (Figure 5.10), and a Wilcoxon rank sum test confirmed that SNPs inside archaic segments (negative distances) had significantly higher relative RD compared to those outside archaic segments (positive distances) in both datasets (1000G Trio: $p < 2.2 \times 10^{-16}$; Irish: $p < 2.2 \times 10^{-16}$).

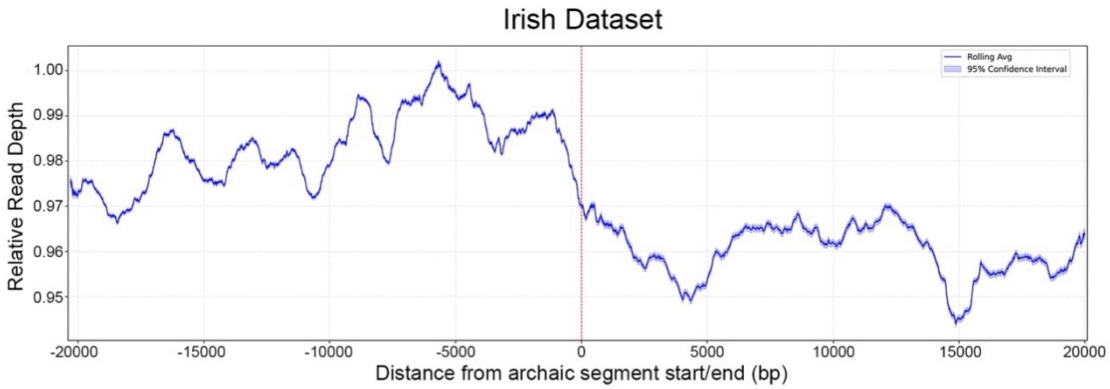


Figure 5.10, Relative RD profile across archaic segment boundaries. This plot shows the rolling average (1 kb window) of relative RD calculated for variants within ±20 kb of all archaic segment start and end positions from the Irish dataset. Distance 0 represents the segment boundary; negative distances are inside the segment, positive distances are outside.

5.3.3 Archaic Segment Depletion in the Modern Human Genome

5.3.3.1 Archaic Deserts

Analysis of the overlap between archaic deserts and DSB hotspots revealed a significant enrichment in the HGDP+1000G dataset compared to random sets of regions (p-value = 0.000999) (Figure 5.11). This pattern indicates that regions completely lacking archaic introgression occur more frequently in recombinationally active regions than expected by

chance. When the analysis was restricted to fully callable regions, the enrichment was less pronounced, but the overlap between deserts and DSB hotspots still exceeded that of all 1,000 random sets (Figure 5.12). These results are consistent with the hypothesis that recombination-related processes, potentially including recurrent NAHR, may contribute to the progressive loss of introgressed archaic segments in genomic regions with high recombination rates.

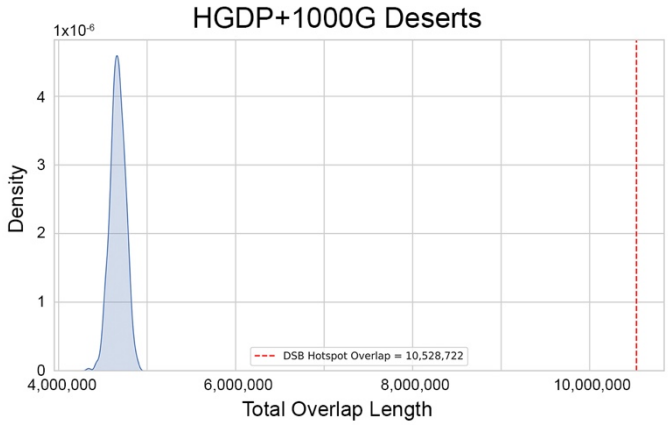


Figure 5.11, Overlap between HGDP+1000G archaic deserts and DSB hotspots. The density plot shows the distribution of overlapping bases with DSB hotspots for 1,000 matched random region sets, and the dashed line indicates the observed overlap for archaic deserts themselves.

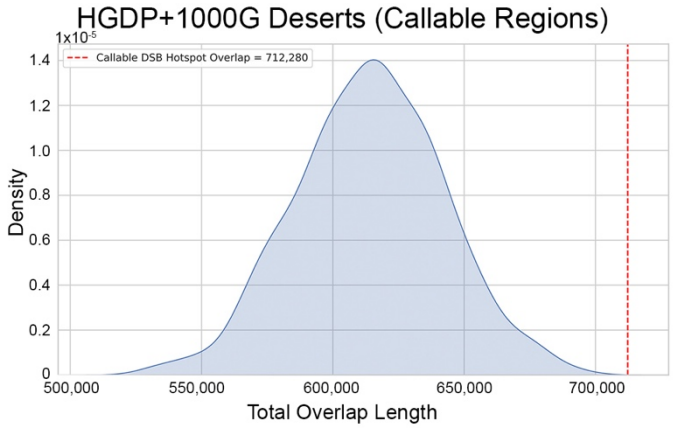


Figure 5.12, Overlap between callable HGDP+1000G archaic deserts and DSB hotspots. The density plot shows the distribution of overlapping bases with DSB hotspots for 1,000 callable matched random region sets, and the dashed line indicates the observed overlap for fully callable archaic deserts.

In contrast, the Irish and Icelandic datasets yielded different results. In the initial analyses (without callability control), both populations also appeared to show an excess of archaic desert overlap with DSB hotspots compared to their random sets (Figure 5.13 and Figure

5.15). However, this signal disappeared once the analysis was restricted to callable regions. In this case, both of the observed overlaps were actually on the lower end of the random distribution (Figure 5.14 and Figure 5.16). This shift indicates that the apparent enrichment in the unfiltered analysis may have been an artefact of callability differences. Because archaic deserts require at least 10% callability to be defined, while random sets could be drawn from poorly callable regions, the random sets may have been less likely to overlap deserts, which could have inflated the apparent signal. DSB hotspots, by definition, occur in well-mapped and better-characterised regions with a higher probability of being callable, making them more likely to overlap with deserts.

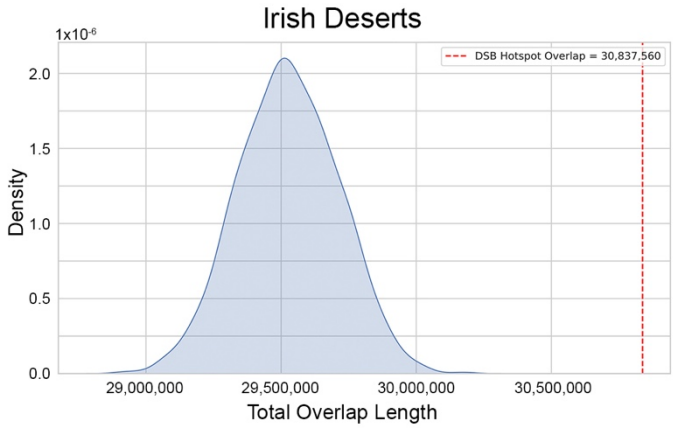


Figure 5.13, Overlap between Irish archaic deserts and DSB hotspots. The density plot shows the distribution of overlapping bases with DSB hotspots for 1,000 matched random region sets, and the dashed line indicates the observed overlap for archaic deserts.

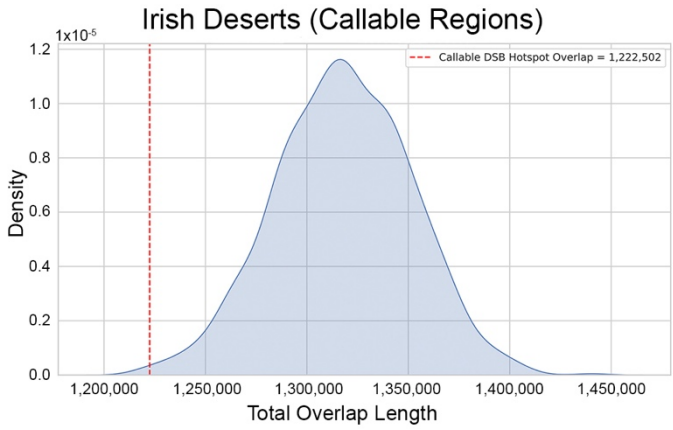


Figure 5.14, Overlap between callable Irish archaic deserts and DSB hotspots. The density plot shows the distribution of overlapping bases with DSB hotspots for 1,000 callable matched random region sets, and the dashed line indicates the observed overlap for fully callable archaic deserts.

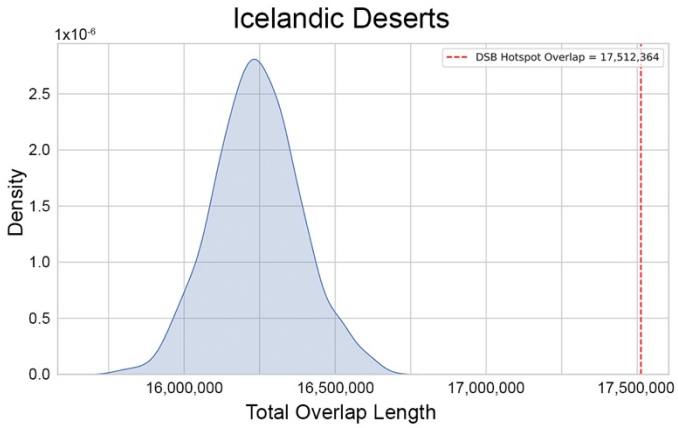


Figure 5.15, Overlap between Icelandic archaic deserts and DSB hotspots. The density plot shows the distribution of overlapping bases with DSB hotspots for 1,000 matched random region sets, and the dashed line indicates the observed overlap for archaic deserts.

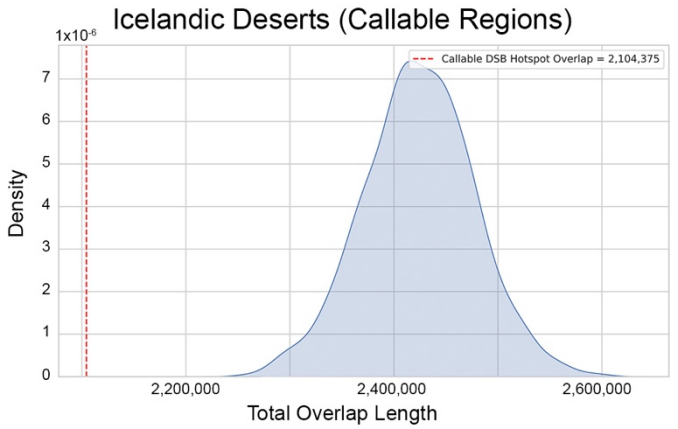


Figure 5.16, Overlap between callable Icelandic archaic deserts and DSB hotspots. The density plot shows the distribution of overlapping bases with DSB hotspots for 1,000 callable matched random region sets, and the dashed line indicates the observed overlap for fully callable archaic deserts.

The divergence in results between datasets may reflect different factors. The HGDP+1000G dataset integrates multiple different populations, meaning that the deserts identified are more robust representations of true deserts, i.e. regions where archaic sequences have been actively purged from the human genome. In contrast, the Irish and Icelandic datasets each represent single, relatively isolated populations that have undergone bottlenecks. Their deserts are therefore more likely to be a mixture of true deserts and artefactual deserts generated by founder effects of genetic drift, thereby diluting any signal of recombination-driven loss.

Another important consideration is the source of the DSB hotspot map itself. These hotspots were derived from five male individuals, one of which is Asian and the other four are Caucasian. The hotspot map is more likely to represent the recombination landscape of the diverse populations included in the HGDP+1000G dataset than that of the more isolated Irish or Icelandic cohorts. This may also help explain why the signal of enrichment was observed most clearly in HGDP+1000G and not in the single-population datasets.

5.3.3.2 Recombination Rate

Consistent with the NAHR-mediated purging hypothesis, a negative correlation was observed between local recombination rate and archaic count in all three datasets (HGDP+1000G: Spearman's $\rho = -0.0362$; Irish: $\rho = -0.0283$; Icelandic: $\rho = -0.0125$). Although these correlations are relatively weak in magnitude, the directionality is consistent across datasets, suggesting that regions with higher recombination activity possess fewer archaic sequences. This pattern is consistent with the possibility that recombination-related processes may contribute to the reduction of archaic sequence in these regions. A strong overall negative correlation would not be expected, as many additional forces (e.g. selection, drift, mutation rate) also shape the distribution of archaic introgression. Instead, the key prediction is a selective depletion of archaic sequence specifically in high recombination regions, rather than a uniform relationship across all recombination rates.

To test this more explicitly, windows were divided into low recombination (≤ 0.1 cM/kb) and high recombination (> 0.1 cM/kb) groups. In the HGDP+1000G dataset, archaic counts were significantly lower in high recombination windows compared to low recombination windows (Wilcoxon rank sum test, $p < 2.2 \times 10^{-16}$; Figure 5.17), with a similar significant reduction observed in the Irish dataset ($p < 2.2 \times 10^{-16}$; Figure 5.17). The Icelandic dataset

showed the same directional trend but the difference was not statistically significant ($p = 0.1152$; Figure 5.17). These results indicate that the most recombinationally active regions of the genome are significantly depleted of introgressed archaic segments, which aligns with the prediction that NAHR efficiently deletes these segments over generations. The absence of statistical significance in the Icelandic dataset may reflect its smaller effective population size and history of bottlenecks, possibly reducing power to detect a significant difference.

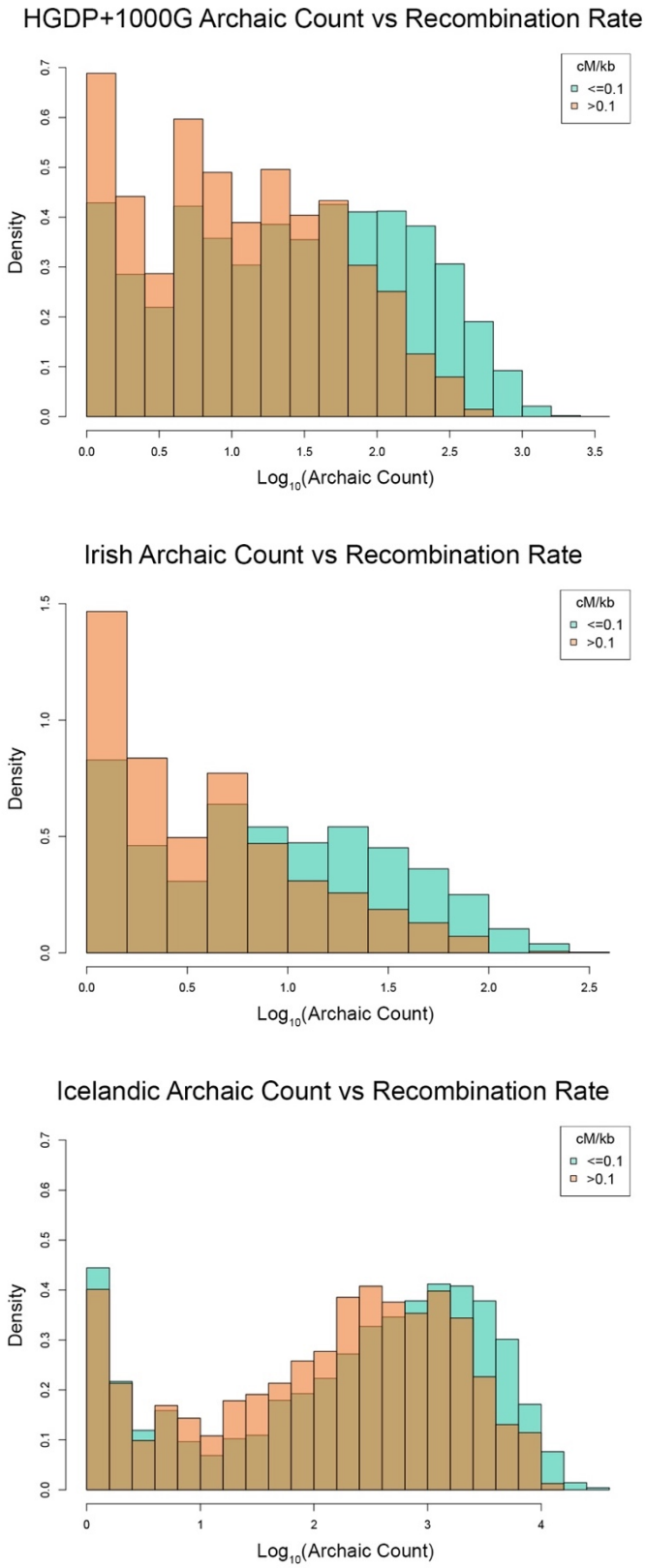


Figure 5.17, Relationship between archaic introgression and recombination rate. Distributions of archaic count are compared between genomic regions with low recombination rate (≤0.1 cM/kb) and high recombination rate (>0.1 cM/kb) across three datasets.

While these data are consistent with the NAHR mechanism, an alternative explanation may also exist. Haplotypes in regions of high recombination are likely more subject to purifying selection, since recombination reduces LD, increases haplotype diversity and may increase the efficiency with which deleterious alleles can be removed (Presgraves, 2005; Preska Steinberg et al., 2022). It is therefore possible that archaic segments were purged more efficiently from these regions, causing the observed inverse relationship between archaic count and recombination rate.

5.3.3.3 Archaic Proportion Depletion

The autosomal archaic proportion of each child in the 1000G Trio dataset was calculated and compared to the mean proportion of their two parents (Figure 5.18). We hypothesised that if recombination between modern human and archaic-derived haplotypes at PRDM9 binding sites generates NAHR-mediated deletions, then humans could progressively lose archaic sequence over successive meiosis, which may be visible over a single generation. The different PRDM9 binding site motif landscapes of modern humans and Neanderthals may enable such events, which could potentially accelerate the removal of archaic segments from the genome.

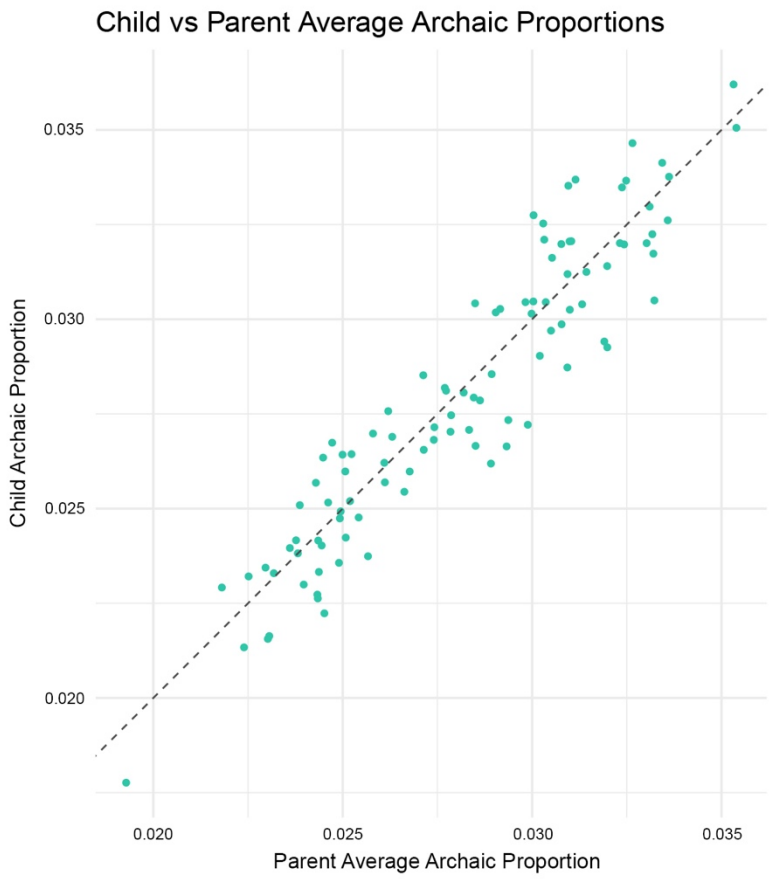


Figure 5.18, Heritability of archaic ancestry in parent-offspring trios. Each point represents a child, plotting their observed archaic proportion against the average of their parents' proportions. The dashed 1:1 line indicates equal proportion in both sets.

Across all 100 trios analysed, children inherited on average 250.5 kb less archaic sequence than would be expected based on the mid-parent value. Although this trend is consistent with the hypothesised loss of archaic sequence, it was not statistically significant ($p = 0.391$; Wilcoxon rank sum test). The lack of significance may reflect the limited sample size, as analysing only 100 trios provides limited power to detect such subtle effects.

When maternal and paternal transmissions were considered separately, the effect appeared more pronounced in maternally transmitted haplotypes. On average, children inherited 459.7 kb less archaic sequence than expected from maternal haplotypes, compared to only 41.2 kb less from paternal haplotypes. This difference is consistent with the higher recombination rate in oogenesis compared to spermatogenesis (Halldorsson et al., 2019; Kong et al., 2010),

which could lead to more frequent NAHR events involving hybrid archaic-modern PRDM9 binding site pairs during female meiosis. Although speculative, such a mechanism could preferentially eliminate archaic segments through maternal transmission, contributing to the long-term decay of archaic ancestry in modern humans. However, given the limited sample size and the lack of statistical significance in the overall trend, this pattern should be interpreted cautiously.

5.4 Discussion

Our findings in this chapter are consistent with the possibility that the novel alleles introduced into modern human populations by archaic introgressed haplotypes may have created substrates for structural instability through NAHR. The evolutionary divergence between modern humans and Neanderthals likely allowed the accumulation of mutations in PRDM9 binding motifs. When these divergent genomes were brought together through hybridisation, the resulting asymmetry in meiotic recombination landscapes could have created hotspots for genomic rearrangement. Our results suggest that such mismatched haplotypes may still exhibit mutagenic effects today, even tens of thousands of years after introgression.

The initial analysis of private variants revealed that variants located within archaic segments (PVAS) exhibit distinctive signatures relative to non-archaic private variants. PVAS were disproportionately enriched in the duplication-associated region of the AB versus RD space, with the strongest signal observed in the Irish dataset and replicated, albeit more subtly, in the 1000G Trio dataset. This enrichment suggests that CNV events are more frequently associated with archaic segments than with non-archaic sequence. This pattern is consistent with our hypothesis that archaic haplotypes introduced divergence in PRDM9 binding sites, thereby promoting ectopic recombination events and generating duplications and deletions. The additional enrichment of PVAS in regions suggestive of double duplications (AB ~0.5, relative RD ~2) is also compatible with the hypothesis that archaic sequence predisposes to recurrent CNV formation.

The phenotypic consequences of *de novo* CNVs are variable and context-dependent. While many CNVs appear to be effectively neutral (Conrad et al., 2010; Sudmant et al., 2015),

others are known to be deleterious, such as those associated with neurodevelopmental disorders, congenital anomalies, and reduced fitness (Cooper et al., 2011; Itsara et al., 2010). This NAHR-based model could therefore strengthen the purifying selection against archaic segments. The direct and immediate negative selection acting on deleterious introgressed alleles could be combined with an indirect, mutagenic effect of the archaic haplotype that increases the risk of NAHR. This therefore represents a separate mechanism that could contribute to reduced hybrid fitness. However, regardless of the immediate fitness impact of any single NAHR event, the recurrent generation of deletions would act as a strong force for the gradual erosion and depletion of introgressed archaic sequence over generations.

The analysis of RD in flanking regions revealed that archaic segment edges are particularly prone to deletion-like signals. In the Irish dataset, flanking windows displayed significantly reduced RD compared to matched random regions, while the 1000G Trio dataset showed a milder but consistent effect. The enrichment of deletions immediately adjacent to archaic sequence boundaries was further supported by direct CNV calls. Across the 200 1000G Trio individuals, flanking regions overlapped significantly more deletions than random genomic controls. This signal was then further refined by the breakpoint analysis, which demonstrated that deletion breakpoints (±1 kb) are significantly enriched for archaic ancestry compared to the genomic background. This finding is consistent with the possibility that archaic segments may act as direct catalysts for these deletion events, potentially by facilitating mispairing during meiosis.

If NAHR between divergent haplotypes is a persistent force, we would expect its effects to accumulate over time, shaping the genomic distribution of archaic ancestry at a population level. Our analyses of archaic deserts and recombination rate support this prediction. The

significant enrichment of archaic deserts in DSB hotspots in the HGDP+1000G dataset indicates that the most recombinationally active regions of the genome have been depleted of archaic segments. These are the regions where the PRDM9-directed recombination machinery is most active, and therefore where the risk of NAHR due to motif mismatches is highest. The weak but consistent negative correlation between recombination rate and archaic count across all three datasets, and the significant depletion of archaic sequence in high-recombination windows, add further weight to this conclusion.

However, there are two important factors that cannot be ruled out as alternative explanations for the observed negative correlation between recombination rate and archaic count. Firstly, it may be more challenging to detect archaic segments in regions of high recombination, as segments may have been fragmented into shorter tracts due to more frequent recombination. The HMM used here to identify archaic introgression cannot reliably identify these short archaic fragments (Skov et al., 2020). Secondly, recombination may increase the efficiency of purifying selection (Presgraves, 2005; Preska Steinberg et al., 2022). It is a possibility that disadvantageous archaic segments can be more easily removed from these regions, which would also result in the observed depletion in recombinationally active regions. Indeed, the role of purifying selection in removing deleterious archaic alleles has been established (Harris & Nielsen, 2016; Juric et al., 2016a; Petr et al., 2019). Nevertheless, the observation (albeit not statistically supported due to low power) from analysis of parent-offspring trios – that children inherit on average 250.5 kb less archaic sequence than expected from their parents – provides a baseline to model the depletion of archaic sequence over evolutionary time. This, if extended in future work, can be incorporated into population genetic models representing human-Neanderthal hybridisation where selection need not play as major a role as previously thought.

Finally, the segment-wise flanking deletion analysis efficiently ties together the signals of duplication and deletion into a single genomic profile around archaic segment boundaries. The observed pattern – a rise in relative RD (suggesting duplications) within the archaic segment, followed by a sharp drop (suggesting deletions) immediately outside its boundaries – provides a striking summation of the features associated with the proposed mechanism. This is the exact structural signature expected of genomic regions that have been shaped by recurrent NAHR, where the archaic segment itself can be duplicated and deletions can be observed in the modern human flanking sequence.

While our results are consistent with a model of NAHR driven by divergent PRDM9 landscapes, several limitations and alternative explanations must be considered. First, the identification of archaic segments itself is challenging and prone to error. False-positive archaic calls, often in regions of low complexity or high mutation rate, could coincidentally overlap with CNV-prone regions, inflating our signals. Similarly, the precise boundaries of archaic segments are difficult to define, and this could blur the distinct duplication-within and deletion-flanking patterns we observe. Our reliance on short-read sequencing data also heavily limits the resolution of CNV and archaic segment calls. Many of the deletions and duplications predicted to occur may be small, complex, or located in repetitive regions that are poorly identified by these technologies. The development of long-read phased genomes with comprehensive archaic ancestry maps and accurately-resolved CNV breakpoints will be crucial for validating these findings and pinpointing the exact NAHR events relative to PRDM9 motifs and archaic segment boundaries. Finally, the population-specific differences we observed highlight the impact of demographic history. Bottlenecks and genetic drift can allow rare, large CNVs to rise in frequency, which could be amplifying the signal of NAHR

in isolated populations. This does not invalidate the mechanism but suggests its genomic footprint may be variable and highly dependent on a population's unique history.

6 Discussion

6.1 Summary of Findings and their Significance

The human genome is dynamic, being sculpted over evolutionary and generational timescales by mutation, recombination, demography, and admixture. This thesis has explored some of the genomic features of these processes, focusing specifically on how changes in PRDM9 binding sites may have shaped structural variation and long-term genome evolution. This relationship has been explored from two different perspectives. First, the contemporary, within-species context of modern human populations was explored, establishing that common genetic variation has the potential to alter the meiotic recombination landscape by creating or disrupting PRDM9 binding motifs, and that these risk diplotypes are associated with CNV formation and are therefore subject to purifying selection. Second, the deep evolutionary context of hybridisation between *Homo sapiens* and archaic hominins was analysed, beginning with the application of a robust pipeline to map the legacy of archaic introgression in the modern human genome. Chapter 5 then directly tested the novel hypothesis that the divergent genomic landscapes of modern humans and Neanderthals created a sustained risk of NAHR in hybrid offspring, and that this mechanism may have contributed significantly to the depletion of Neanderthal ancestry in the modern human genome.

Chapter 3 demonstrated that SNPs are capable of reconfiguring the meiotic recombination landscape by altering PRDM9 binding affinity. The erosion of PRDM9 motifs through biased gene conversion and mutational turnover is well documented (Boulton et al., 1997; Myers et al., 2010), but far less appreciated is the contribution of polymorphism to their continual

gain and loss. The signature of purifying selection acting on loss-causing SNPs indicates that the potential to disrupt the meiotic landscape and predispose to NAHR may carry a fitness cost. The stronger selection signal of loss variants compared to gain variants also aligns with the expectation that disrupting an existing, epigenetically primed recombination hotspot is likely to have more immediate consequences than creating a novel motif that may not be located in a chromatin context permissive for recombination initiation. These findings carry significant implications. Recombination hotspots are not fixed features of the genome but instead are highly dynamic and can be reshaped both by changes in PRDM9 alleles and by local sequence evolution. Within populations, this creates the possibility for lineage-specific recombination landscapes, which could alter haplotype structures and potentially contribute to population differentiation. At a broader level, this establishes a mechanism through which ordinary SNP variation can combine with recombination machinery to produce large-scale structural consequences.

Chapter 4 served as a critical pivot, which applied a rigorously validated pipeline to map the legacy of archaic introgression in multiple datasets. This involved a deep investigation into the impact of data quality on introgression inference, and the subsequent implementation of quality control filters to eliminate false positives without ablating true archaic signal. Apart from the stringent quality control measures, this chapter also investigated the forces shaping archaic deserts. While these regions of complete introgression depletion were found to be more gene-dense, a direct analysis of functional constraint provided little support for the prevailing model of pervasive negative selection. Additionally, we found a similar density of SNPs in Neanderthal genomes in desert and non-desert regions, indicating that archaic deserts do not appear to be inherently more constrained in the Neanderthal genome itself.

Chapter 5 proposed a more dramatic application of the principle established in chapter 3 by considering the different PRDM9 landscapes in humans and Neanderthals. Their divergence would have allowed for the accumulation of a large number of mutations that would functionally be classified as gain- or loss-causing SNPs. The distinct PRDM9 alleles possessed by humans, Neanderthals, and Denisovans (Schwartz et al., 2014) would have allowed such mutations to be effectively neutral within their own lineage, only possessing a functional consequence in the context of the other lineage's PRDM9 protein. Furthermore, the phenomenon of hotspot erosion, whereby PRDM9's own activity continuously mutates its binding sites, would have proceeded independently in each lineage (Lesecque et al., 2014). This would contribute to the acceleration of the divergence of their recombination landscapes. The collision of these landscapes during hybridisation would have created a genome-wide abundance of mismatched motif pairs. A hybrid meiocyte containing a chromosomal pair comprising one modern human and one Neanderthal haplotype with different motif sets would experience a fundamental asymmetry. In such a context, the recombination machinery may be more prone to engage in ectopic recombination between paralogous sequences.

The analyses in chapter 5 provided multiple lines of evidence supporting this model. We found a significant enrichment of deletion signals flanking archaic segments along with an enrichment of duplication signals within archaic segments, a strong association between deletion breakpoints and archaic ancestry, and a distinct RD profile around segment boundaries that is consistent with recurrent NAHR. This profile, which showed a rise in relative RD (suggestive of duplication) within the archaic segment and a sharp drop (suggestive of deletion) immediately outside its boundaries, is a striking pattern suggestive of the proposed mechanism. A significant depletion of archaic segments in regions of high

meiotic recombination activity in the HGDP+1000G dataset was also demonstrated. This is another key prediction of the model, as these are the regions where PRDM9-directed recombination is most active and thus where the risk of NAHR due to motif mismatches is highest.

6.2 The Depletion of Neanderthal Ancestry in the Modern Human Genome

It has been well established that following the interbreeding between humans and Neanderthals ~47,000 years ago (Iasi et al., 2024), Neanderthal haplotypes were rapidly purged from the human genome (Green et al., 2010; Harris & Nielsen, 2016; Pfennig & Lachance, 2024; Quilodrán et al., 2023). This large-scale removal of introgressed material over a relatively short evolutionary timeframe requires a powerful explanation. For over a decade, the predominant theory to explain this depletion has centred around negative selection. However, this narrative may be incomplete. The findings here suggest that diverging PRDM9 landscapes leading to NAHR-driven genomic instability may be a contributing force to the observed depletion in the modern human genome.

Neanderthals are believed to have had a higher genetic load than humans due to their smaller long-term effective population size (Harris & Nielsen, 2016; Prüfer et al., 2014). Harris & Nielsen (2016) conducted simulations based on the distribution of fitness effects and estimated that Neanderthal alleles reduced the fitness of hybrid offspring by at least 0.5%. When these deleterious alleles were introduced into the larger, outbred modern human population, natural selection could act with more efficiency to purge them. This, along with the general underrepresentation of archaic ancestry in functional genomic regions

(Sankararaman et al., 2014; Vernot & Akey, 2014), helps to explain the widespread reduction in Neanderthal ancestry in modern human genomes. The fact that many Neanderthal alleles were removed within just a few generations after introgression (Iasi et al., 2024; Sümer et al., 2025) is also consistent with strong selection against highly deleterious variants.

The Neanderthal alleles that were purged from the modern human genome were not necessarily deleterious in their own population, but may have been involved in negative epistatic interactions when combined with alleles from the modern human genome (Bateson, 1909; Dobzhansky, 1934; Muller, 1942). The most prominent genomic signature of such incompatibilities is the pronounced depletion of Neanderthal ancestry on the X chromosome (Sankararaman et al., 2014). This 'large X-effect' is a classic hallmark of hybrid problems, as recessive incompatible alleles on the X chromosome are exposed in hemizygous males, which can often lead to sterility or inviability (Coyne, 1985). This significant depletion is also seen in genes specifically expressed in the testes (Sankararaman et al., 2014, 2016), which suggests that hybrid sterility may have been a significant barrier. However, Juric et al. (2016) showed that the majority of selection against Neanderthal ancestry in the modern human genome was not caused by strong purifying selection, but rather by weakly deleterious alleles that may have been effectively neutral in the smaller Neanderthal population.

As detailed in chapter 4, our analysis of archaic deserts also provided little support for strong negative selection as the sole driver. Thus, the model of selection may not provide a complete explanation for the depletion of Neanderthal DNA in the modern human genome. We propose that NAHR-driven genomic instability represents a complementary force that contributed to the depletion of Neanderthal ancestry. This model is built on the principle

established in chapter 3, that divergence in PRDM9 binding motifs predisposes to meiotic errors. When scaled to the ~500,000 years of divergence between humans and Neanderthals, this predicts a genome-wide substrate for instability.

The results from chapter 5 reveal a consistent enrichment of duplications within archaic segments as well as an enrichment of deletions flanking archaic segments. These genomic signatures are consistent with our model. As noted in chapter 5, NAHR has a documented bias towards generating deletions over duplications (Turner et al., 2008). This would result in a slow and steady erosion of introgressed sequence over generations, therefore acting as a persistent removal force even on haplotypes that were neutral or only weakly deleterious. This mechanism can help to explain the formation of archaic deserts without requiring that the entire landscape of deserts are functionally constrained. Thus, the persistence of alleles in the Neanderthal genome that were quickly purged after introgression into the modern human genome may not be solely attributable to a reduced efficacy of purifying selection due to their smaller effective population size. Instead, the genomic incompatibility of a portion of these alleles may have only arisen in the hybrid state, predisposing these regions to meiotic errors and subsequent deletion.

This model is particularly compelling because it can operate in concert with the effects of negative selection. For example, a Neanderthal haplotype harbouring a weakly deleterious allele might persist for many generations under the weak selective pressures estimated by Juric et al. (2016). However, if that same haplotype also resides in a genomic region prone to NAHR due to motif asymmetry, the resulting SV could delete the haplotype in a single generation. Thus, NAHR could act as an accelerator of purifying selection, effectively clearing the most NAHR-prone haplotypes from the genome much faster than selection

alone could. This has the potential to explain the speed at which initial ancestry was depleted following admixture. Although this thesis did not attempt to estimate the relative extent to which individual loci are prone to NAHR due to the abundance and effect size of gain/loss SNPs, this could be a productive future direction in work aimed at further delineating the role of NAHR in archaic sequence erosion.

6.3 Limitations and Future Directions

A previously discussed limitation of this thesis is the reliance on short-read sequencing data. The ~150 bp sequence read length underpinning all of the data analysed creates significant challenges for accurately detecting SVs, which are central to the NAHR hypothesis. For example, the identification of CNVs in the 1000G Trio dataset relied on indirect signals from discordant read pairs, split reads, and RD anomalies (Rausch et al., 2012). Many reported events in this case showed no apparent supporting RD evidence, suggesting a high false positive rate. This greatly impaired the ability to confidently connect a parental risk diplotype to a specific *de novo* NAHR event in the offspring, as the CNV call itself was likely often spurious. The unphased nature of the genotype data used in this analysis also necessitated conservative filters, such as excluding breakpoint regions with more than one heterozygous SNP. This approach likely reduced false positives but also could have discarded genuine cases where multiple SNPs cooperatively alter PRDM9 binding affinity or where the phase relationship between a motif-altering SNP and a CNV breakpoint is critical for establishing causality. Genome sequence data derived from long-read sequencing technologies would be more appropriate for addressing these questions. Long reads can span entire CNVs and their breakpoints, providing base pair resolution for NAHR events. This would allow for the examination of whether breakpoints occur within or near diverged PRDM9 motifs, offering direct evidence for the mechanism.

The HMM-based method used here to detect archaic introgression also possesses many limitations. Aside from the potential false positives and false negatives, the imprecise definition of segment boundaries is a major issue with relevance to the NAHR hypothesis. The edges of archaic segments are difficult to pinpoint because they are defined by the locations of private variants, which are not necessarily located at the edge of a haplotype. The method used involves smoothing and probabilistic inference, meaning that the precise start and end coordinates of a segment are uncertain. This directly impacts the analyses in chapter 5 that looked for deletions or duplications specifically flanking or within archaic segments. A deletion that appears to flank a segment might, in reality, be located just inside an inaccurately defined boundary, or vice versa. This imprecision adds noise and potentially obscures the signal we seek to detect. To enhance the accuracy of introgression maps, future work should integrate diverse methods to create a high-confidence set of introgressed segments. Regions identified by multiple, independent methods with different statistical principles (e.g. Sprime (Browning et al., 2018), which uses LD; IBDMix (Chen et al., 2020), which uses identity by descent; and machine learning approaches such as ArchIE (Durvasula & Sankararaman, 2019)) might represent a more unambiguous set of genuine archaic segments.

A significant simplification in this thesis is the treatment of all predicted PRDM9 binding motifs as being functionally equivalent. The analysis assumed that a sequence match to a PRDM9 PWM is a reliable proxy for an active recombination hotspot, which is not necessarily the case. The binding of PRDM9 and the subsequent initiation of a DSB are dependent on the local chromatin environment. Factors such as chromatin accessibility, nucleosome positioning, and histone modifications are essential for determining whether a

potential binding site will actually be utilised during meiosis. Thus, many of the millions of gains, losses, and existing PRDM9 binding sites identified in the genome likely reside in chromatin that is closed or otherwise non-permissive for recombination initiation. Consequently, our analyses likely include a high background of non-functional motifs, thereby diluting the signal from truly active sites. To counter this effect, we restricted many of our analyses to PRDM9 binding sites found within experimentally validated DSB hotspots, though these may not be representative of the entire datasets we investigated.

Another simplification in this thesis was the utilisation of one set of consensus PRDM9 motifs, primarily representing the common *PRDM9-A* and *PRDM9-B* alleles, to scan the genome for potential binding sites. However, the zinc finger domain of PRDM9 is one of the most rapidly evolving regions in the human genome, leading to substantial variation in binding specificities (Chen et al., 2020). For example, the *PRDM9-C* allele, which is common in parts of Asia, recognises a distinctly different set of motifs than *PRDM9-A* or *PRDM9-B*. Therefore, the functional impact of a given motif-altering SNP is not definite but is entirely dependent on the *PRDM9* genotype of the individual. This presents a major challenge for population-level analyses. The classification of a SNP as a universal gain or loss is an oversimplification; it is only a gain/loss for individuals carrying the relevant *PRDM9* allele. This is further complicated by the divergence of the Neanderthal and Denisovan *PRDM9* alleles, which are closely related to modern-day African populations (Schwartz et al., 2014), thereby adding more diversity to the binding specificities studied here. This complexity likely added noise to our analyses and may have diluted the association between risk diplotypes and CNVs. For more specific analyses, individual-specific PWMs could be generated for each individual's *PRDM9* allele, which would allow a more accurate identification of true gains and losses.

Many of our analyses focused on risk diplotypes, i.e. pairs of PRDM9 motifs within 200 kb that could theoretically mispair. However, an important mechanistic requirement for NAHR was not incorporated into the research: the necessity for sequence homology surrounding the PRDM9 motifs. For NAHR to occur, the recombination machinery requires substantial homologous sequence (typically hundreds of base pairs with >95% identity (Parks et al., 2015; Startek et al., 2015)). The analyses detailed here considered only the motif pair itself, without requiring that the surrounding genomic sequence has sufficient homology to actually support ectopic recombination. Therefore, the majority of the risk diplotypes identified are likely non-functional because they are not located within larger homologous sequences. The decision in this thesis to focus only on identifiable PRDM9 motifs was one made out of analytical parsimony, but future work should develop computational pipelines that, for a given motif pair, perform local pairwise alignment of the flanking sequences to calculate the degree of homology. Although this would be computationally intensive, it would allow for the identification of a refined set of high-probability NAHR-prone loci. Alternatively, a more targeted approach could be used where known LCRs are analysed. One could investigate whether SNPs that alter PRDM9 motifs within these known homologous regions are especially enriched for CNV breakpoints and under strong purifying selection.

This thesis relied solely on present-day genomes in an attempt to understand processes that have been unfolding for tens of thousands of years. The depletion of Neanderthal ancestry and the accumulation of SVs may be more apparent in DNA from early modern humans, as the current distribution of Neanderthal ancestry is the product of not only selection and instability but also complex demographic events such as population bottlenecks, expansions, and replacements. By analysing genomes from different time periods (e.g. the ~40,000 year

old Oase 1 from Romania (Fu et al., 2015), the ~45,000 year old Ust'-Ishim individual from Siberia (Fu et al., 2014), and later Neolithic and Bronze Age individuals), we could track how the proportion and genomic distribution of Neanderthal ancestry changed. The detection of SVs in ancient genomes could also provide direct evidence for past NAHR events. Identifying a deletion, for example, that precisely flanks an introgressed archaic segment in an ancient individual would be powerful support for the model. This is currently challenging, however, as the difficulties in detecting SVs would be amplified in these genomes due to the preexisting challenges in the analysis of ancient DNA (Dalal et al., 2023; Prüfer et al., 2010), including short fragment length.

Finally, quantitative models could be built to test whether the proposed role for NAHR in ancestry depletion could explain the observed patterns of persistence and absence of introgressed sequence. Population genetic simulations that explicitly incorporate the NAHR mechanism could be used to test this. Simulations have previously been used in evolutionary genetics to demonstrate that the patterns of genetic diversity are consistent with a serial founder effect out of Africa (DeGiorgio et al., 2009), to estimate the strength of selection against Neanderthal ancestry (Juric et al., 2016), and to test different models of archaic admixture timing and number of pulses (Vernot & Akey, 2014). Similarly, we could construct a temporospatial simulation framework anchored by high coverage Neanderthal and Denisovan genomes, as well as other sequenced ancient genomes that could act as intermediate checkpoints to validate the model. Such a simulation could incorporate parameters for demographic history, purifying selection, variable rates of migration and expansion, hybrid infertility, assortative mating and breeding structure, and combine these with a NAHR risk function defined based on motif divergence, flanking homology, and recombination rate. Such simulations could be compared on various levels to identify the

parameters required to result in the current observed level of Neanderthal ancestry and distribution of surviving archaic segments. By comparing simulations that include or exclude the NAHR mechanism, we could quantify the relative contribution of genomic instability to the depletion of Neanderthal ancestry over evolutionary time. If the NAHR model produces a genomic landscape that is a significantly better fit to the empirical data, this would provide powerful support for the hypothesis that the divergence in meiotic landscapes was a key driver of human evolutionary history.

6.4 Conclusion

This thesis aimed to explore the consequences of variation in the genomic landscapes that guide meiosis. We found that meiotic recombination is not only a driver of genetic diversity, but is also itself susceptible to modulation by common genetic variation. Alterations to PRDM9 binding sites may predispose genomes to structural instability through NAHR, a phenomenon which is intensified by the deep evolutionary divergence between subspecies.

The initial investigation established that SNPs altering PRDM9 binding sites have the potential to reconfigure the meiotic landscape, thereby creating risk diplotypes that are associated with SV formation and are subject to purifying selection. This principle was then expanded when applied to the context of human-Neanderthal hybridisation. The evolutionary divergence between these lineages resulted in fundamentally different PRDM9 binding site motifs and, thus, different recombination landscapes, and the collision of these landscapes in hybrid individuals created a substrate for genomic instability. We have presented observations consistent with the idea that this instability manifested as an elevated

risk of NAHR, generating a characteristic signature of duplications within and deletions flanking introgressed archaic segments.

This mechanism provides a compelling alternative explanation for the depletion of Neanderthal ancestry. While the prevailing theory of negative selection acts on the functional content of alleles, the NAHR mechanism proposed here acts on the genomic context of the haplotype itself. A neutral archaic segment could be removed from the modern human genome not because it is selected against, but because its presence increases the probability of a meiotic error resulting in a deletion. This model works in concert with negative selection to explain the purging of Neanderthal segments from the modern human genome.

Ultimately, this thesis reframes our understanding of hybridisation's legacy. Analysis of archaic introgression in the modern human genome reveals not only the functional consequences of admixed alleles but also the structural challenges of integrating divergent genomic architectures. The deserts of Neanderthal ancestry may not exclusively be signals of purifying selection, they may also signify fundamental incompatibilities that took place in the germlines of our hybrid ancestors. By demonstrating this connection between meiotic processes and evolutionary outcomes, this work offers a new perspective on the complex processes that have made us who we are today.

Appendix A

NAHR risk diplotype distributions

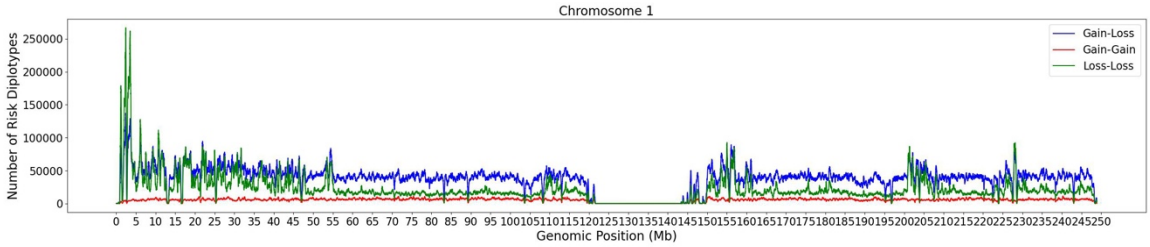


Figure A1, NAHR risk diplotype distributions along chromosome 1. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 1, with a window size of 100 kb and a step size of 1 bp.

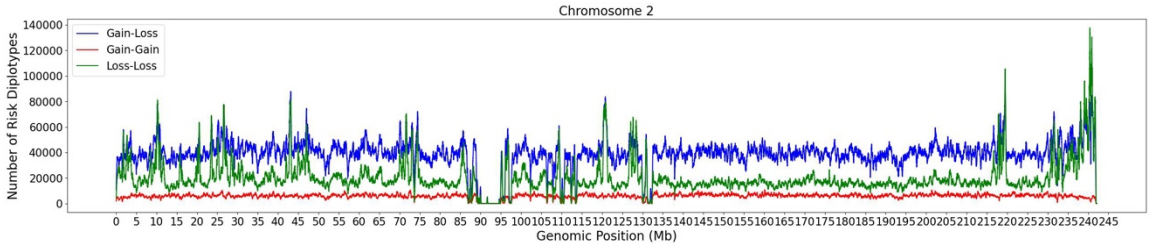


Figure A2, NAHR risk diplotype distributions along chromosome 2. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 2, with a window size of 100 kb and a step size of 1 bp.

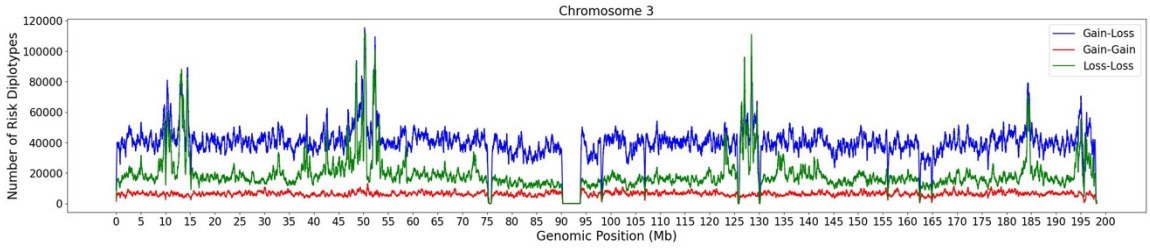


Figure A3, NAHR risk diplotype distributions along chromosome 3. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 3, with a window size of 100 kb and a step size of 1 bp.

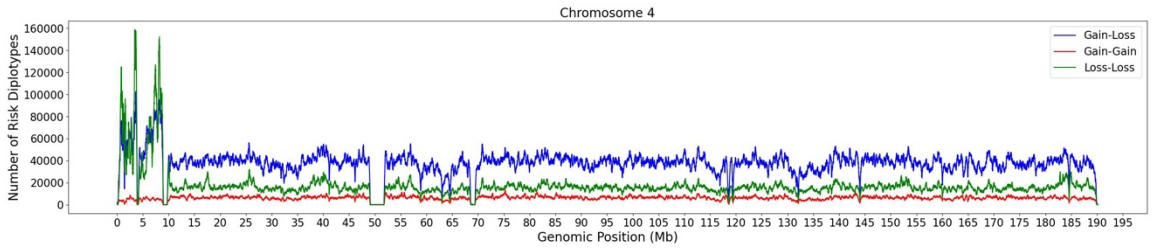


Figure A4, NAHR risk diplotype distributions along chromosome 4. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 4, with a window size of 100 kb and a step size of 1 bp.

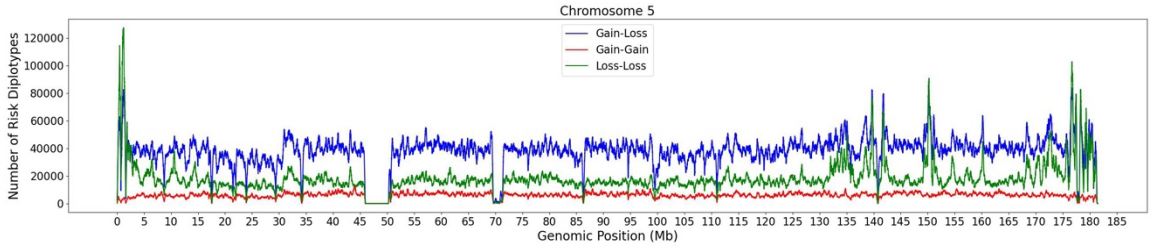


Figure A5, NAHR risk diplotype distributions along chromosome 5. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 5, with a window size of 100 kb and a step size of 1 bp.

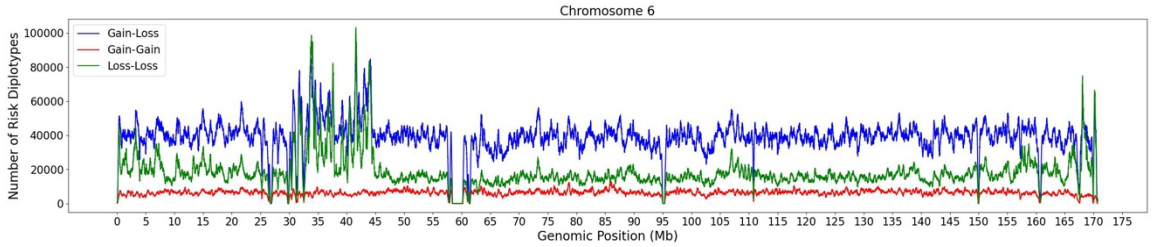


Figure A6, NAHR risk diplotype distributions along chromosome 6. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 6, with a window size of 100 kb and a step size of 1 bp.

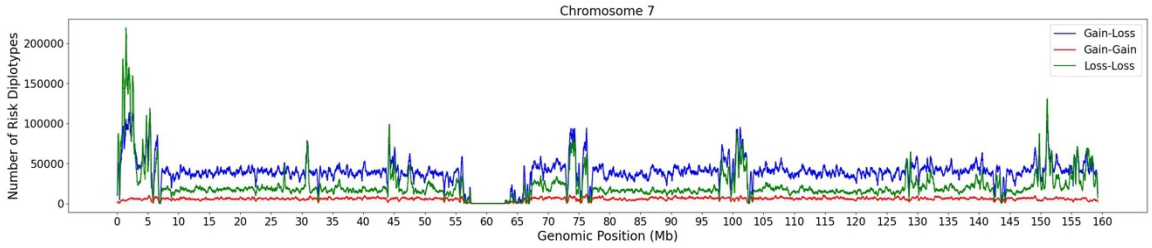


Figure A7, NAHR risk diplotype distributions along chromosome 7. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 7, with a window size of 100 kb and a step size of 1 bp.

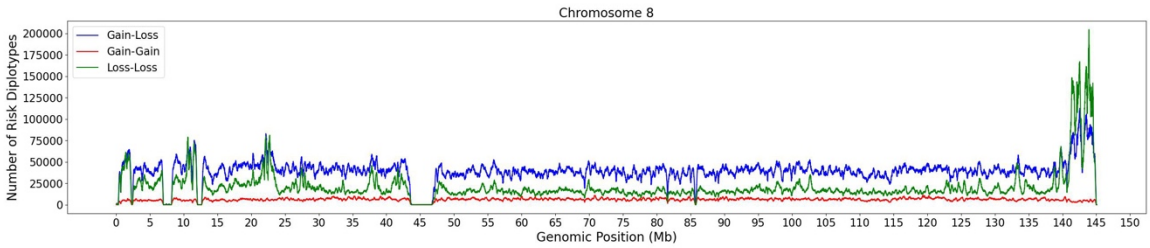


Figure A8, NAHR risk diplotype distributions along chromosome 8. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 8, with a window size of 100 kb and a step size of 1 bp.

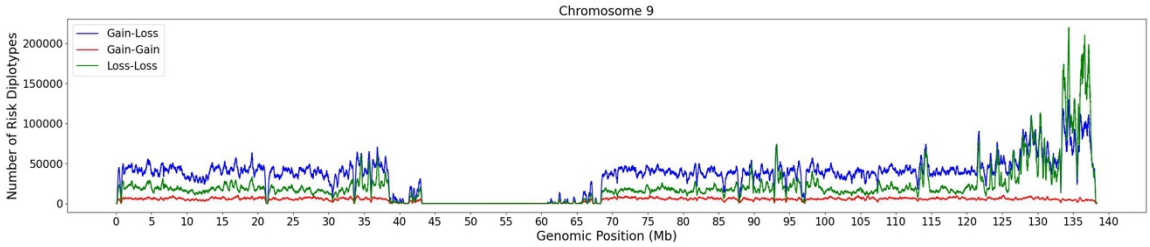


Figure A9, NAHR risk diplotype distributions along chromosome 9. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 9, with a window size of 100 kb and a step size of 1 bp.

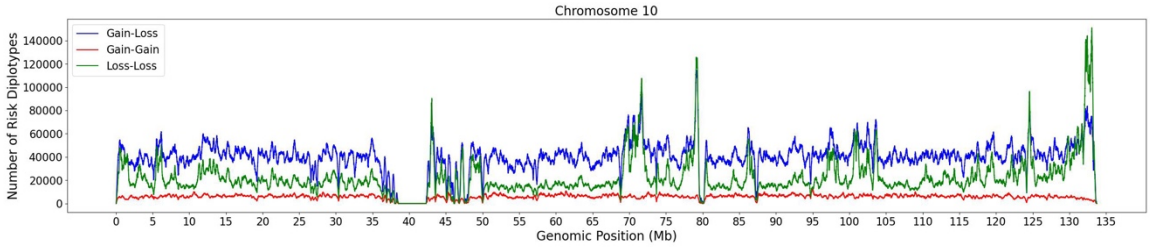


Figure A10, NAHR risk diplotype distributions along chromosome 10. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 10, with a window size of 100 kb and a step size of 1 bp.

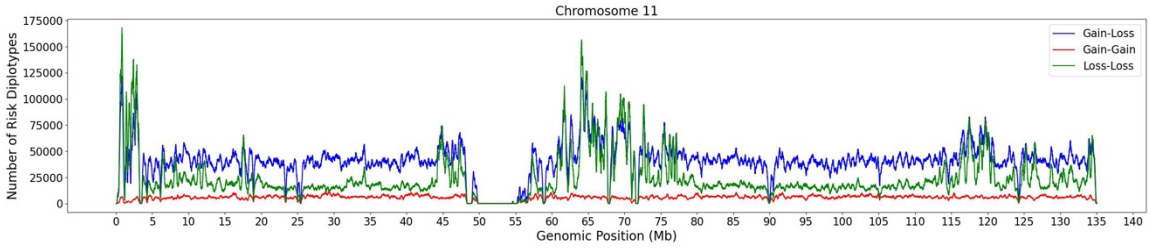


Figure A11, NAHR risk diplotype distributions along chromosome 11. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 11, with a window size of 100 kb and a step size of 1 bp.

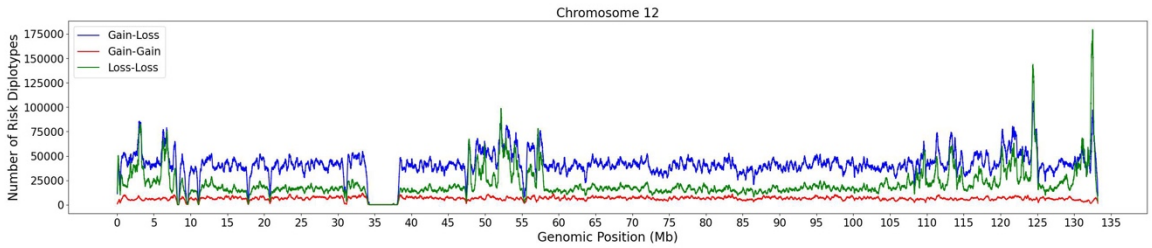


Figure A12, NAHR risk diplotype distributions along chromosome 12. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 12, with a window size of 100 kb and a step size of 1 bp.

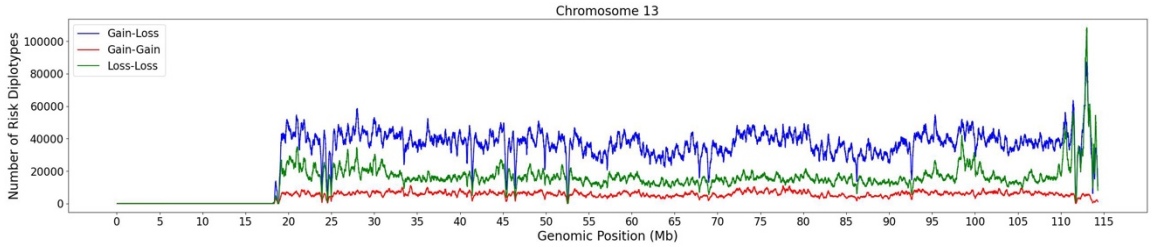


Figure A13, NAHR risk diplotype distributions along chromosome 13. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 13, with a window size of 100 kb and a step size of 1 bp.

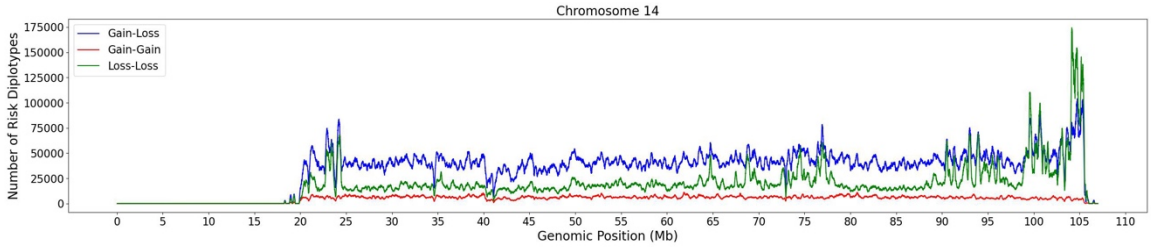


Figure A14, NAHR risk diplotype distributions along chromosome 14. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 14, with a window size of 100 kb and a step size of 1 bp.

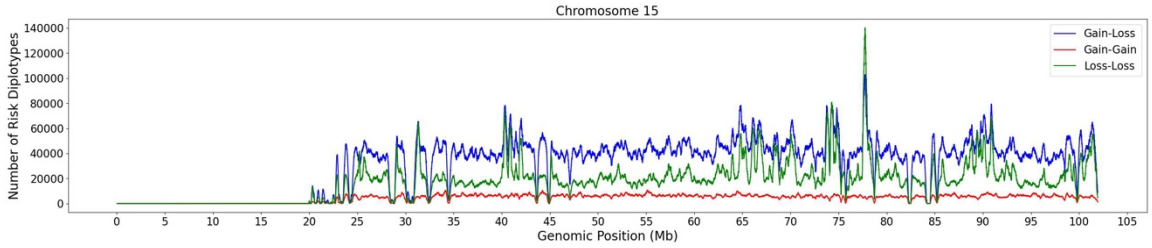


Figure A15, NAHR risk diplotype distributions along chromosome 15. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 15, with a window size of 100 kb and a step size of 1 bp.

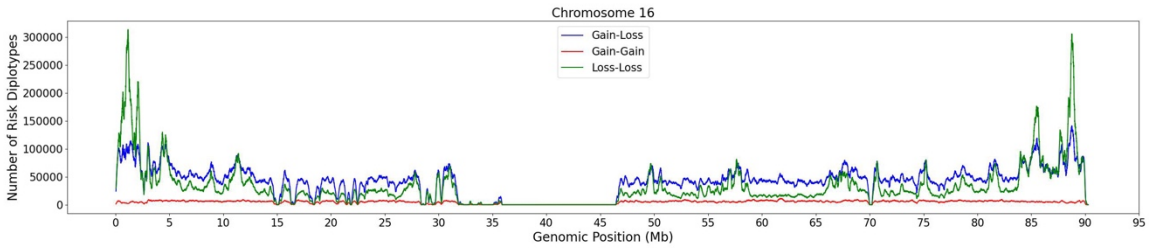


Figure A16, NAHR risk diplotype distributions along chromosome 16. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 16, with a window size of 100 kb and a step size of 1 bp.

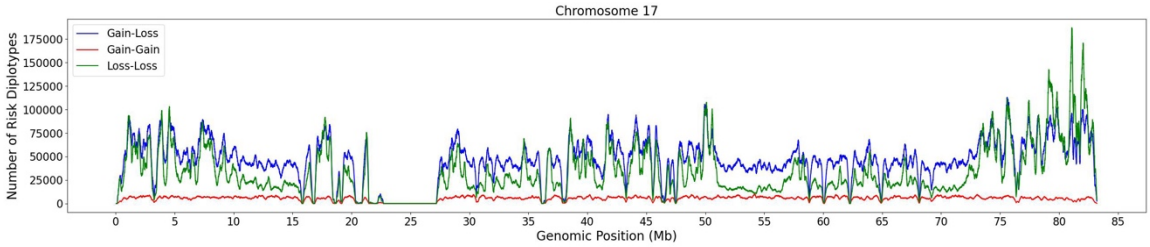


Figure A17, NAHR risk diplotype distributions along chromosome 17. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 17, with a window size of 100 kb and a step size of 1 bp.

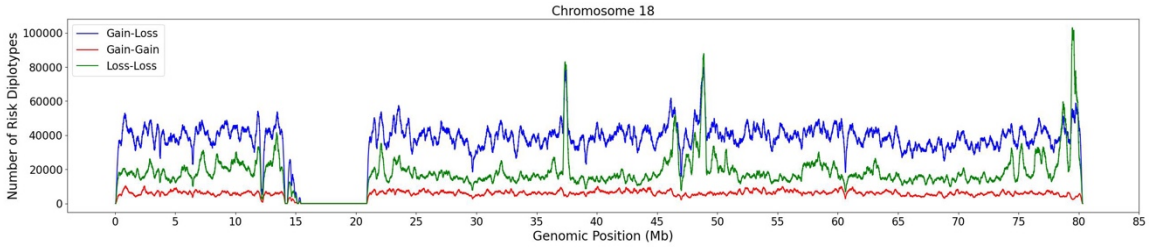


Figure A18, NAHR risk diplotype distributions along chromosome 18. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 18, with a window size of 100 kb and a step size of 1 bp.

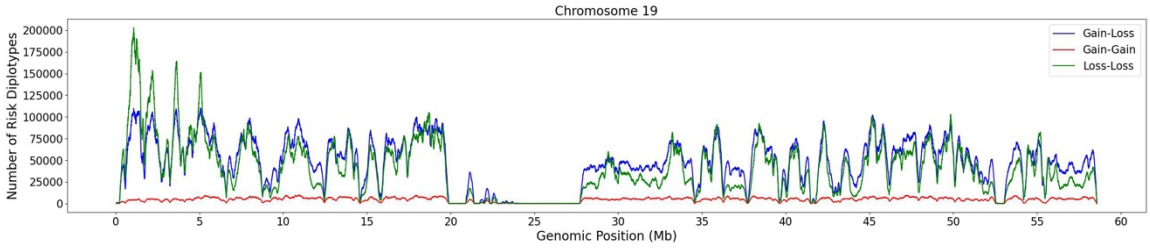


Figure A19, NAHR risk diplotype distributions along chromosome 19. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 19, with a window size of 100 kb and a step size of 1 bp.

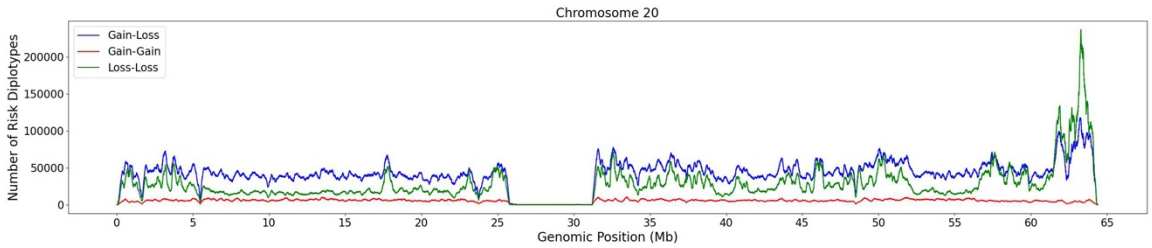


Figure A20, NAHR risk diplotype distributions along chromosome 20. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 20, with a window size of 100 kb and a step size of 1 bp.

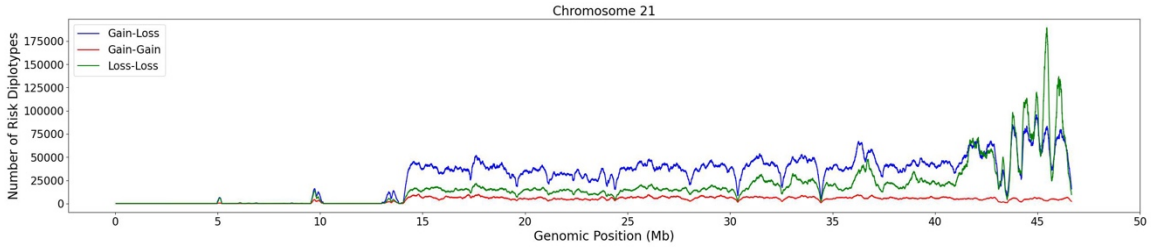


Figure A21, NAHR risk diplotype distributions along chromosome 21. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 21, with a window size of 100 kb and a step size of 1 bp.

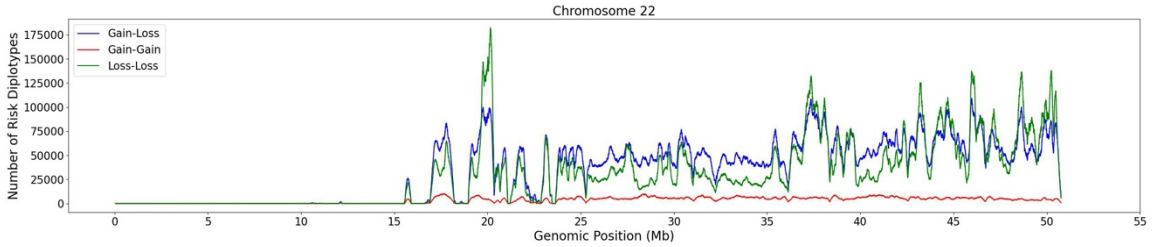


Figure A22, NAHR risk diplotype distributions along chromosome 22. This plot displays a rolling average of the number of gain-loss, gain-gain and loss-loss combinations throughout chromosome 22, with a window size of 100 kb and a step size of 1 bp.

De novo CNV plots

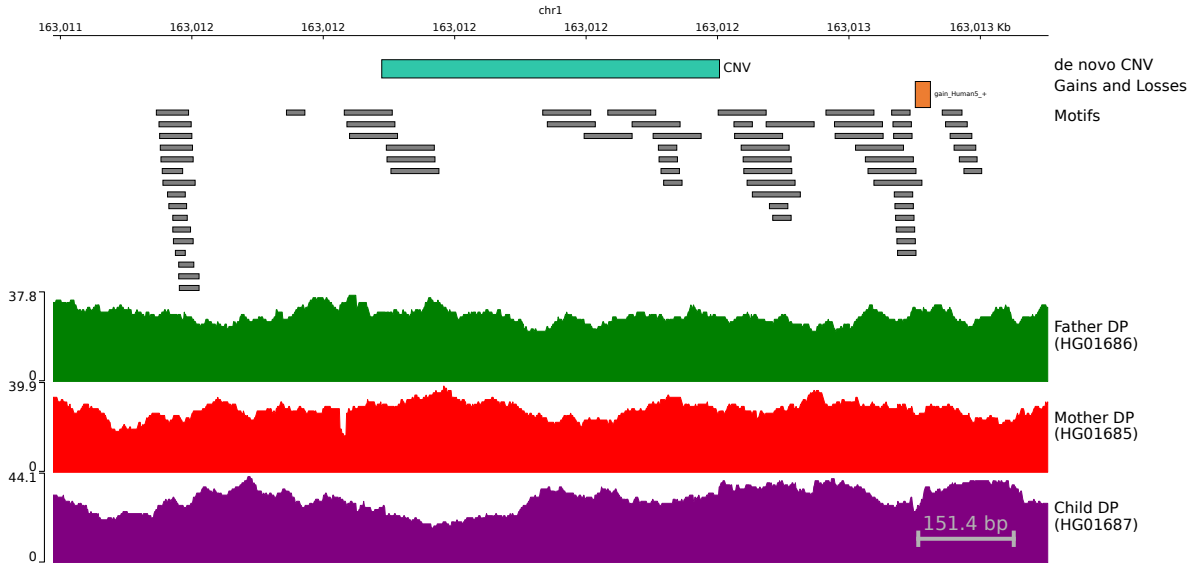


Figure A23, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

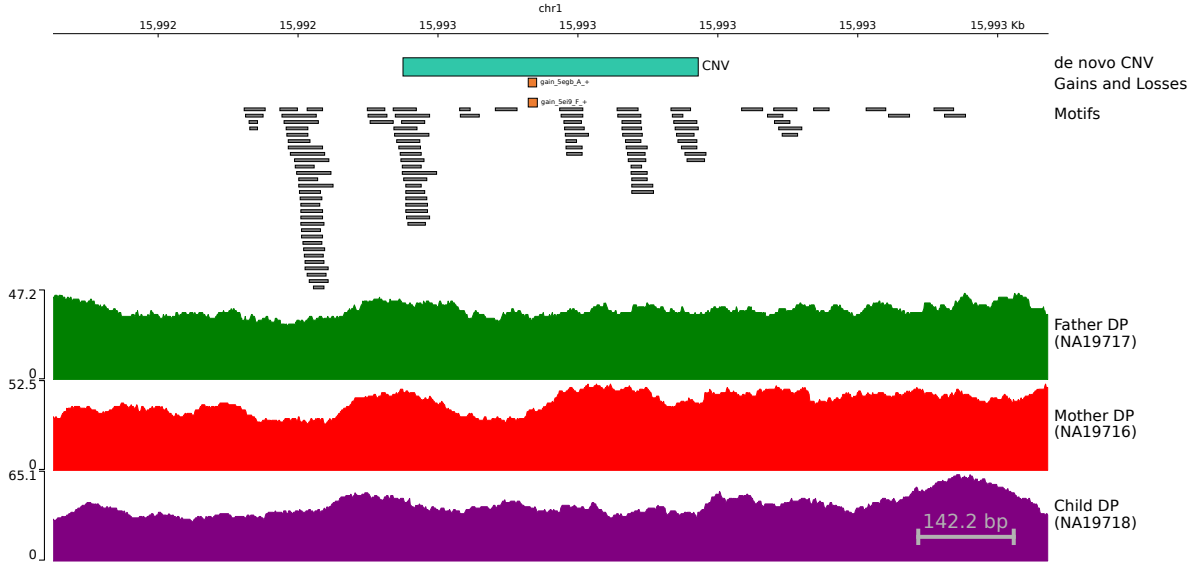


Figure A24, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

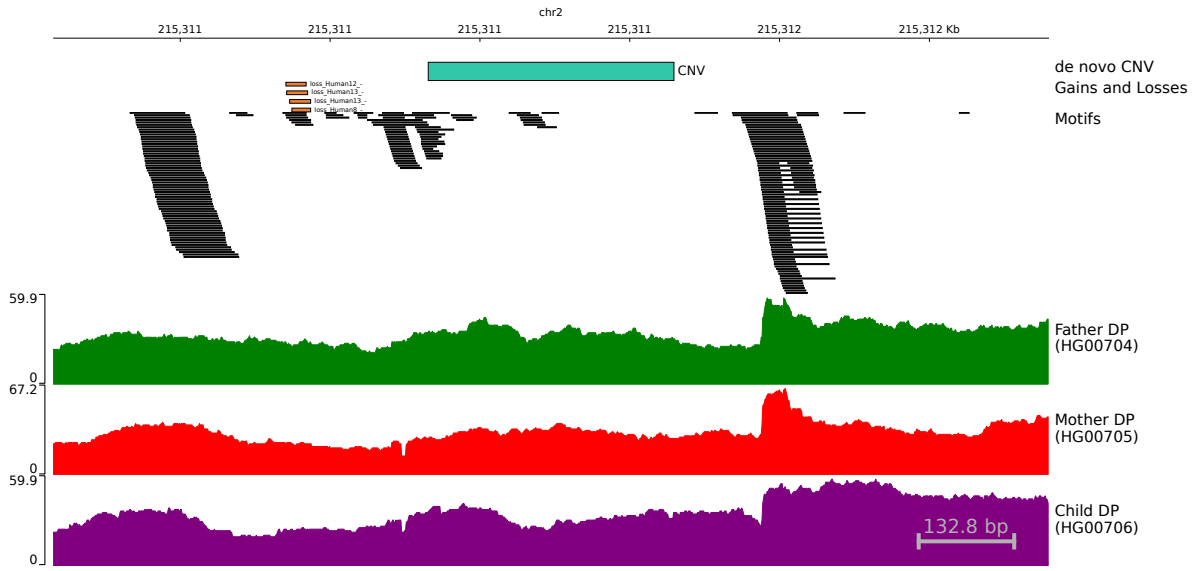


Figure A25, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

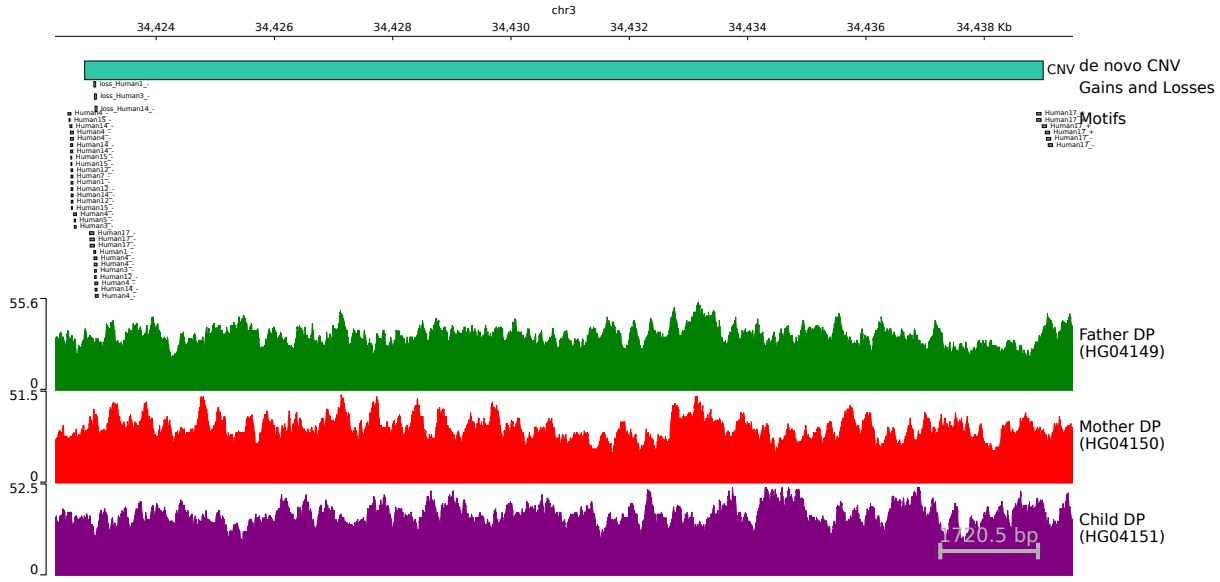


Figure A26, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

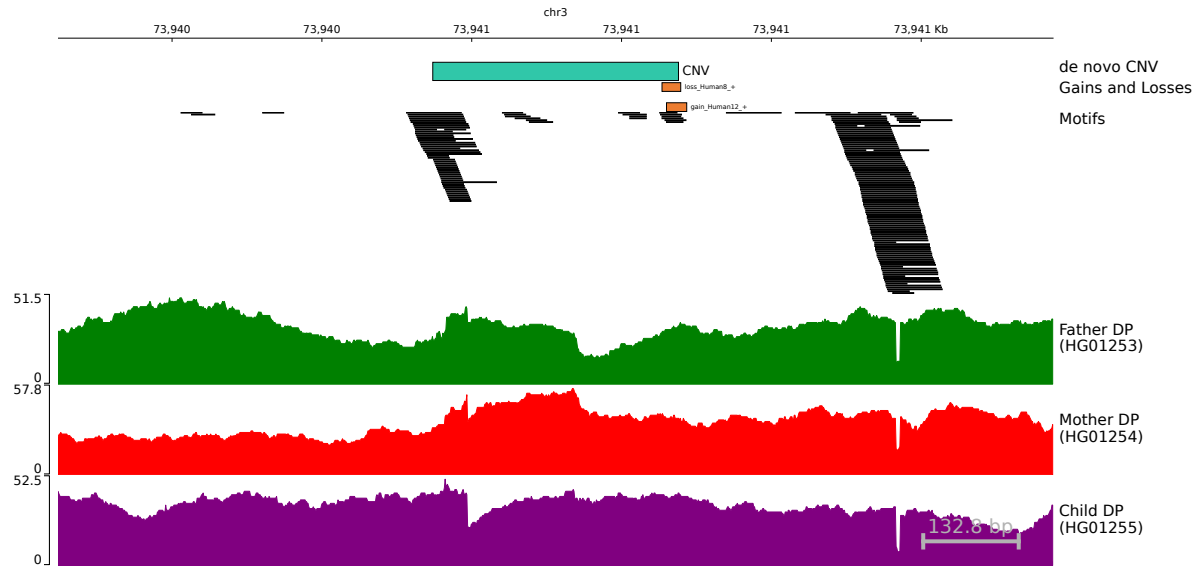


Figure A27, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

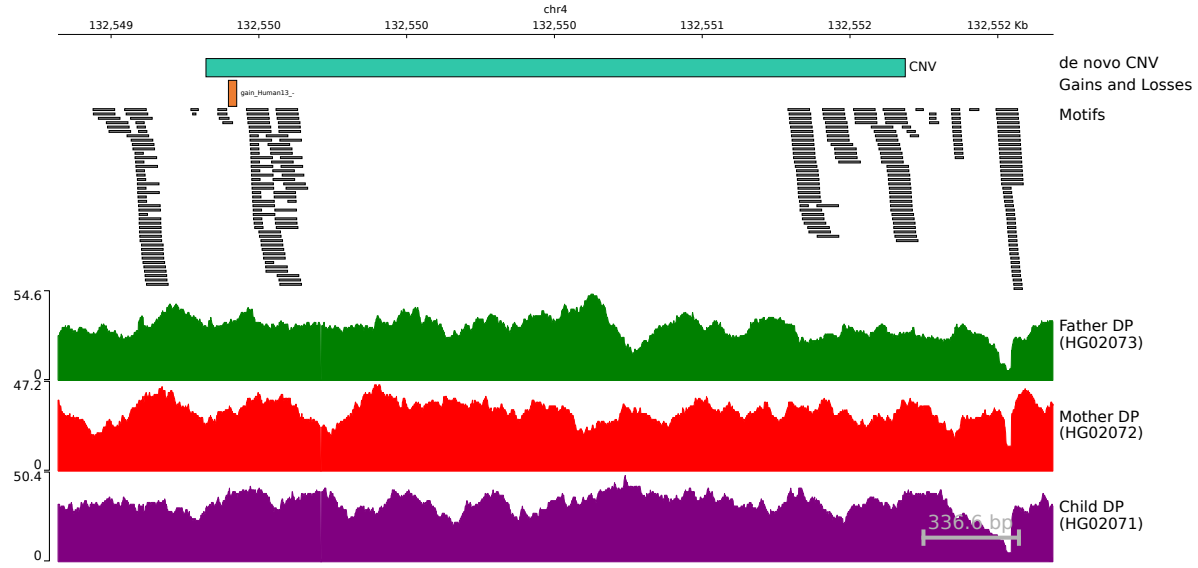


Figure A28, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

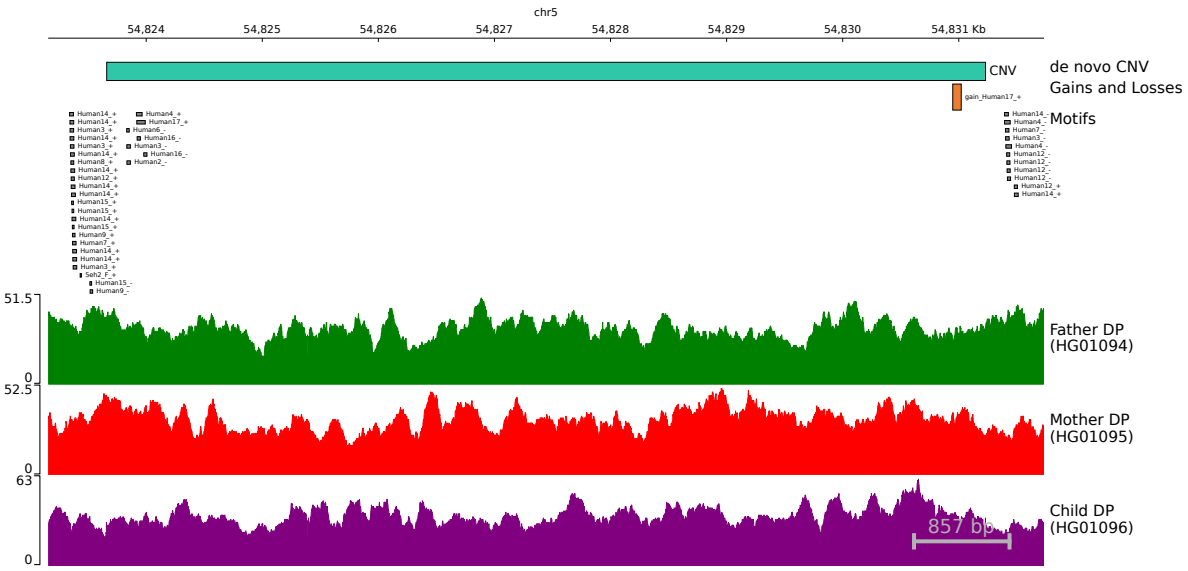


Figure A29, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

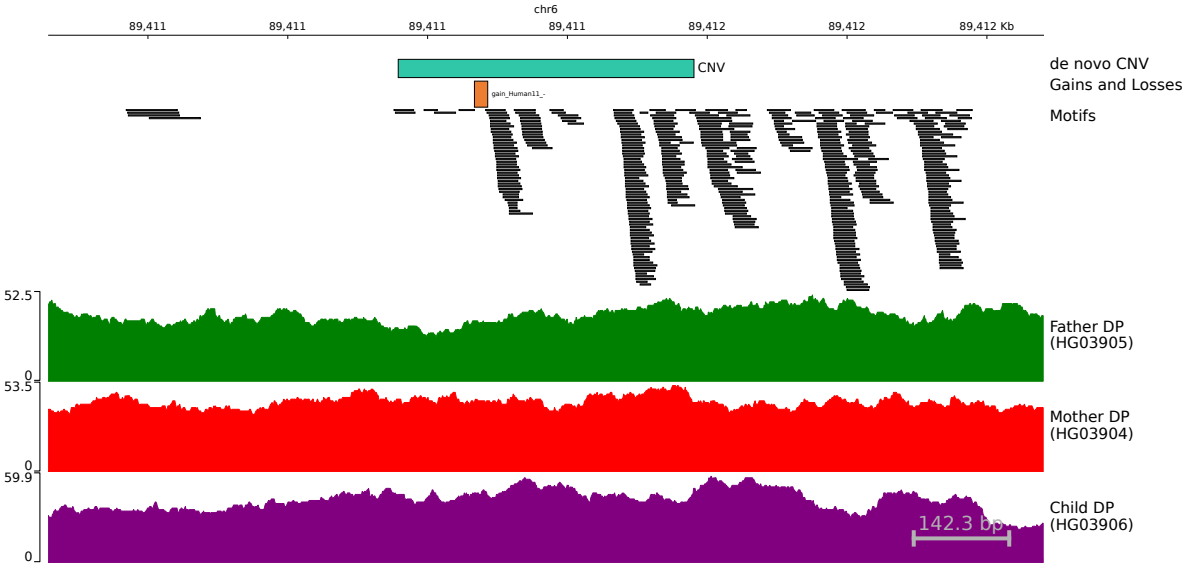


Figure A30, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

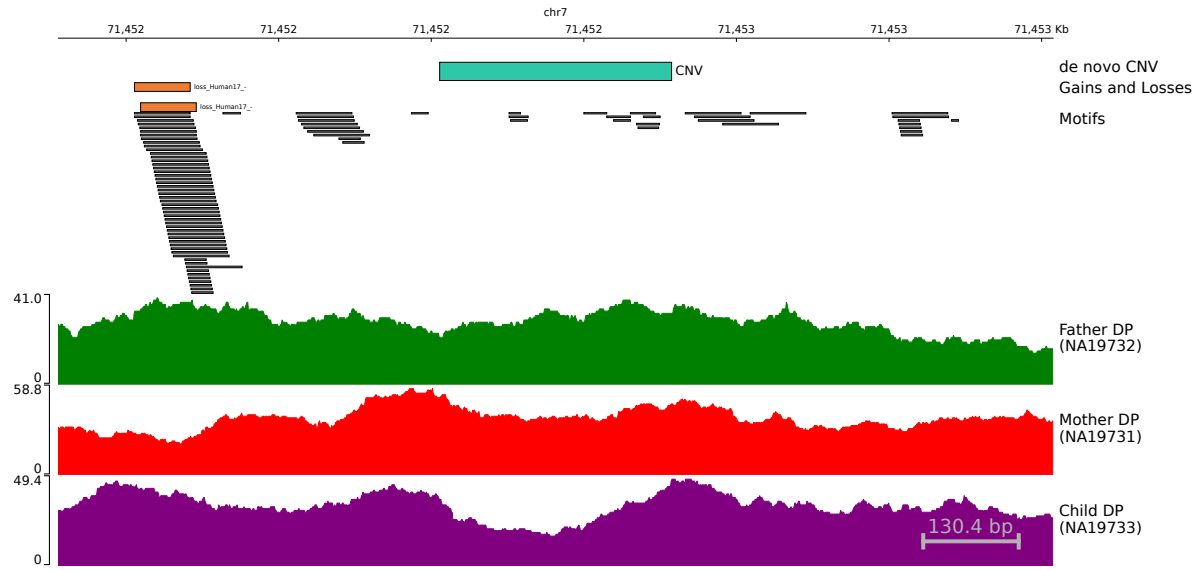


Figure A31, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

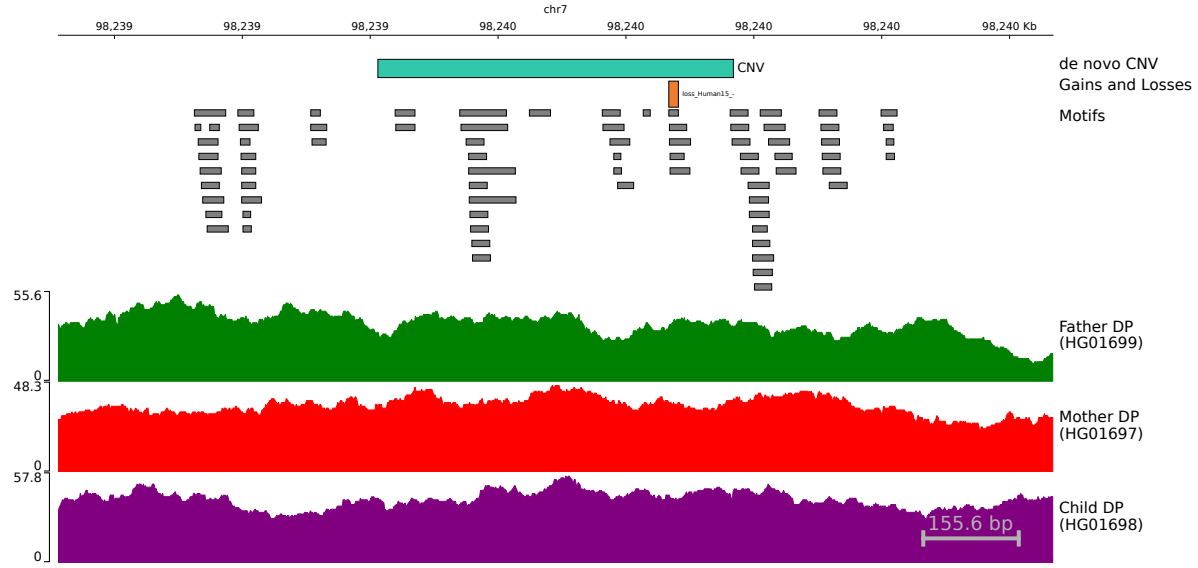


Figure A32, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

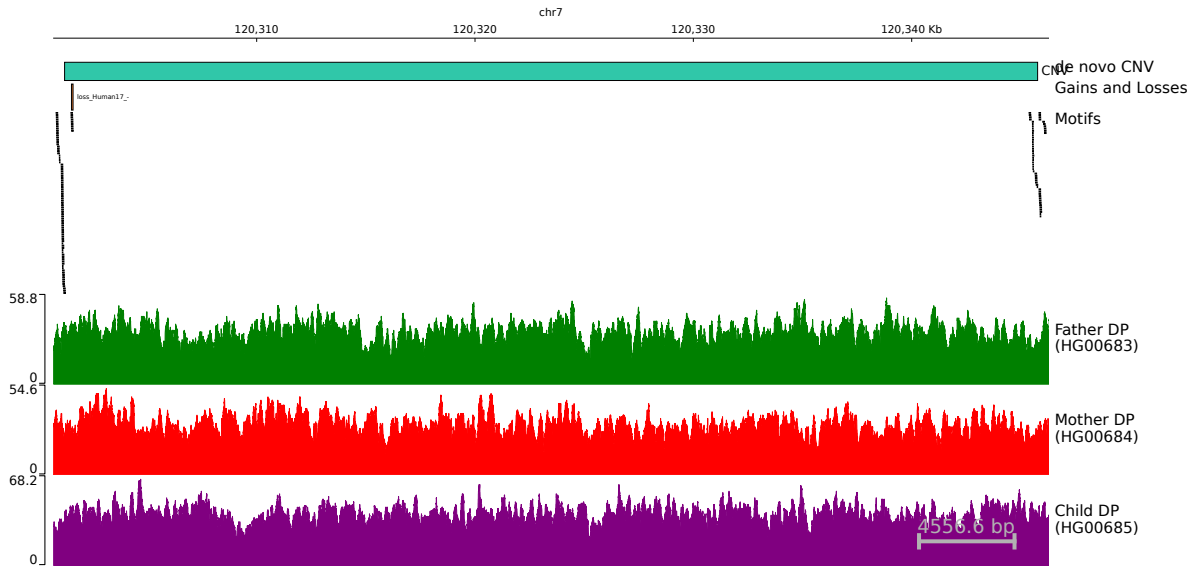


Figure A33, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

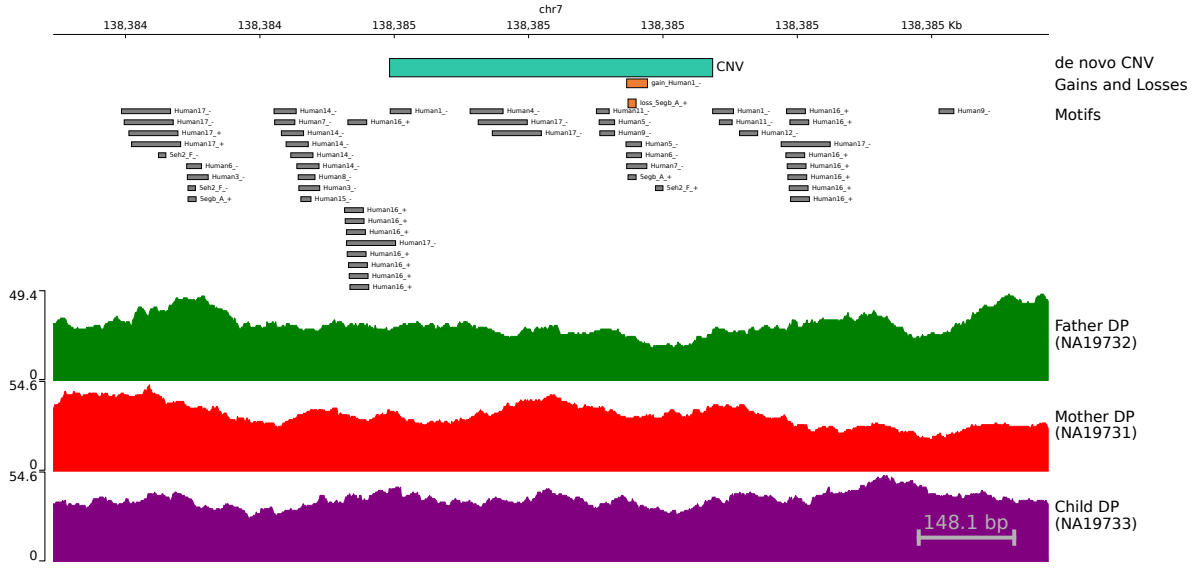


Figure A34, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

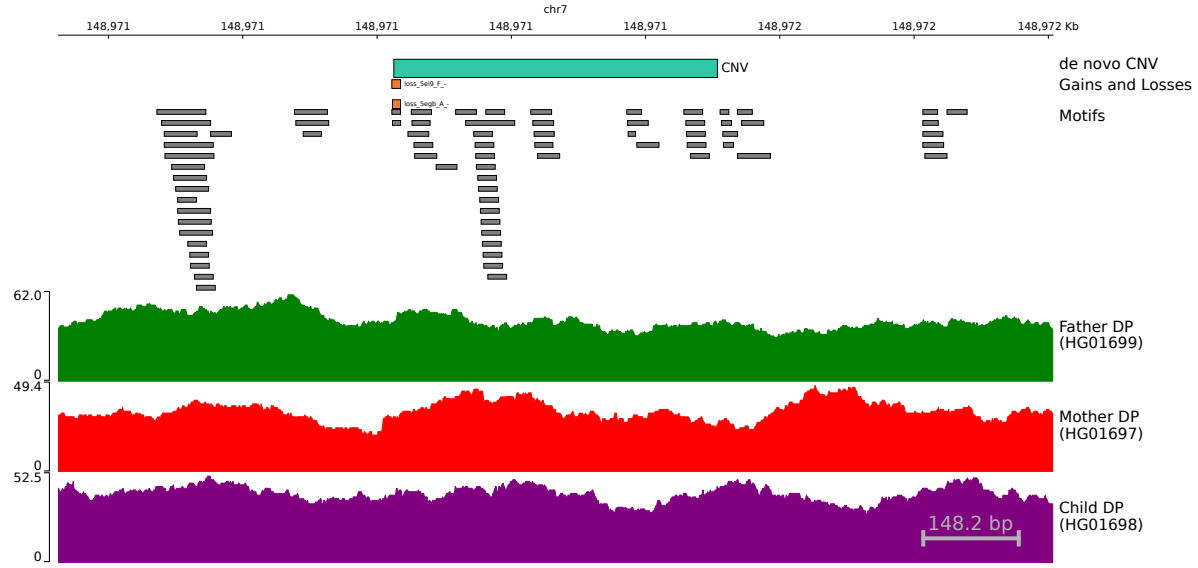


Figure A35, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

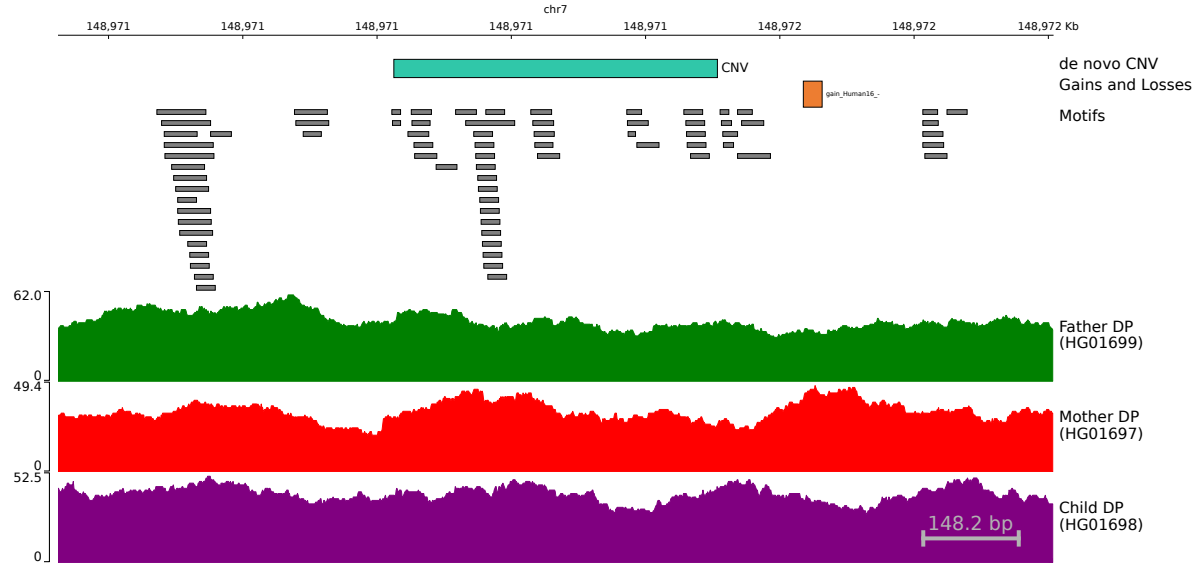


Figure A36, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

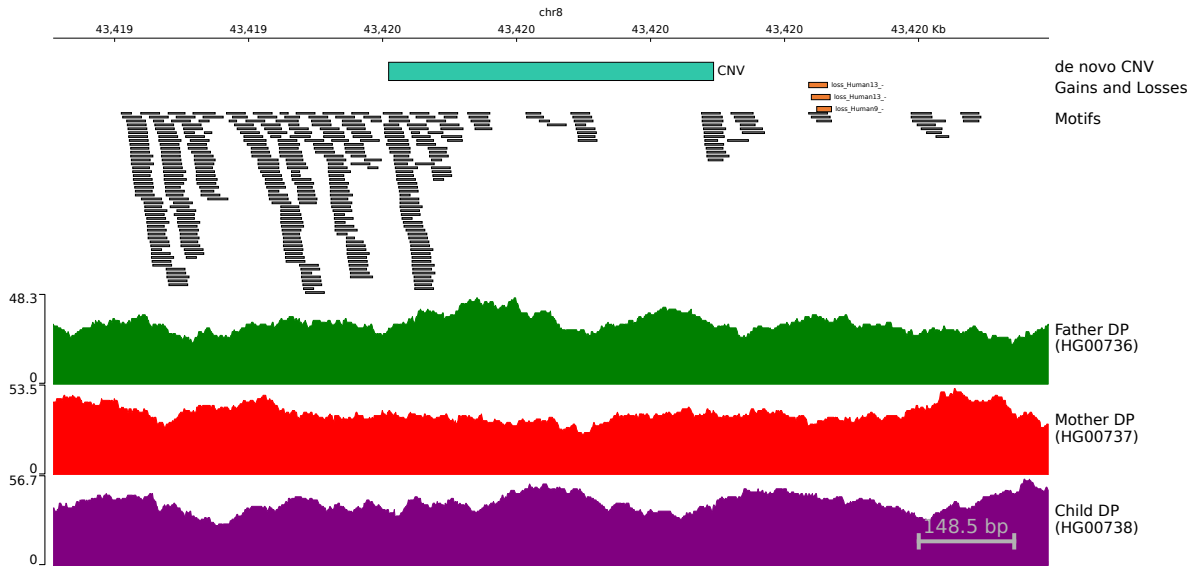


Figure A37, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

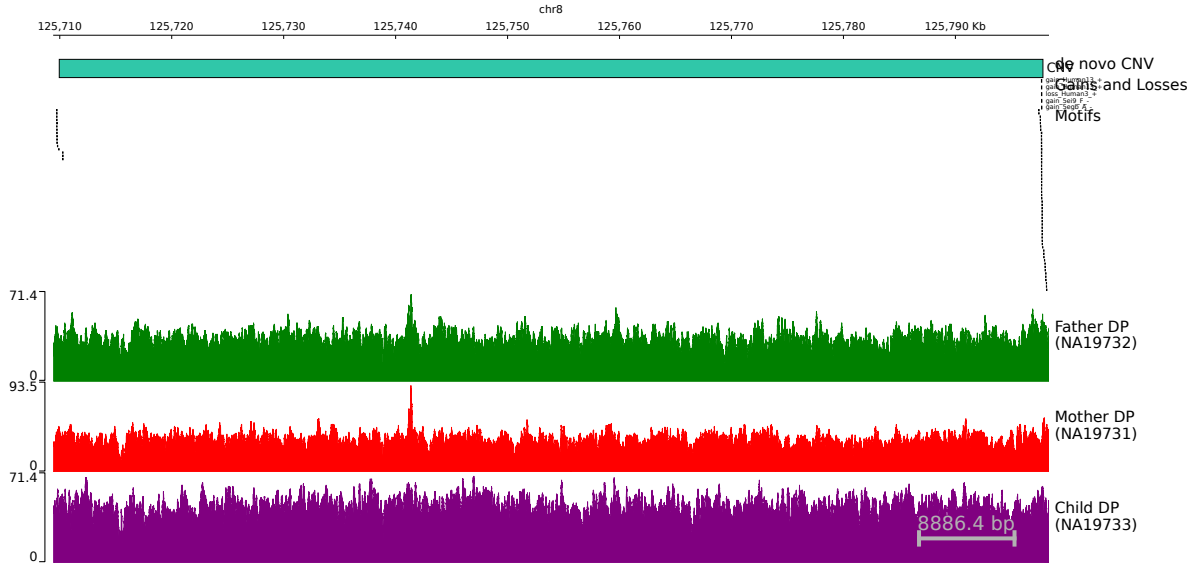


Figure A38, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

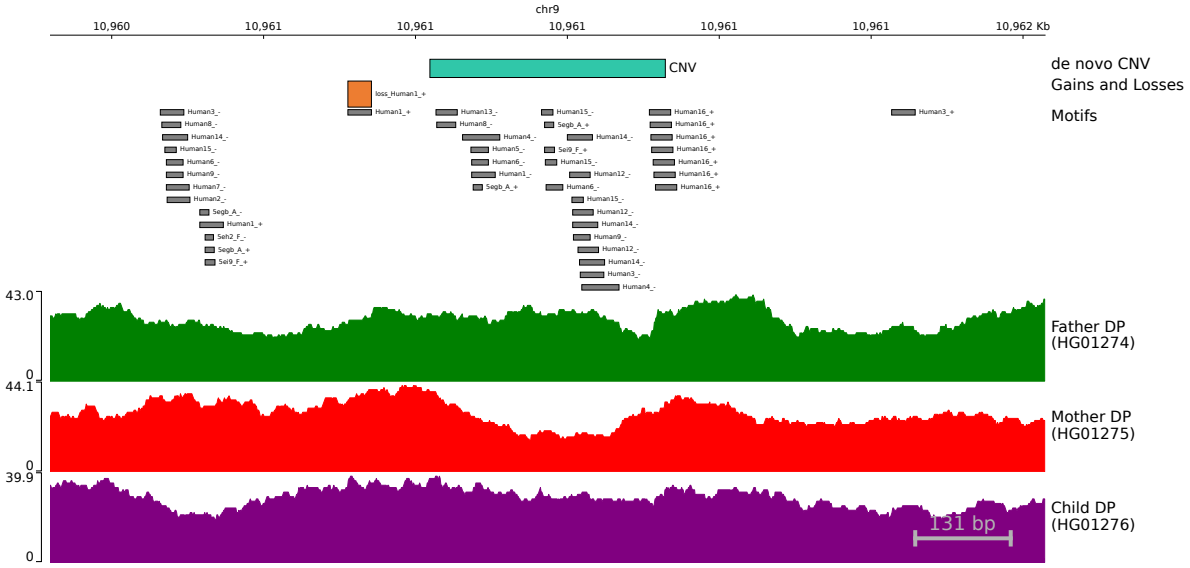


Figure A39, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

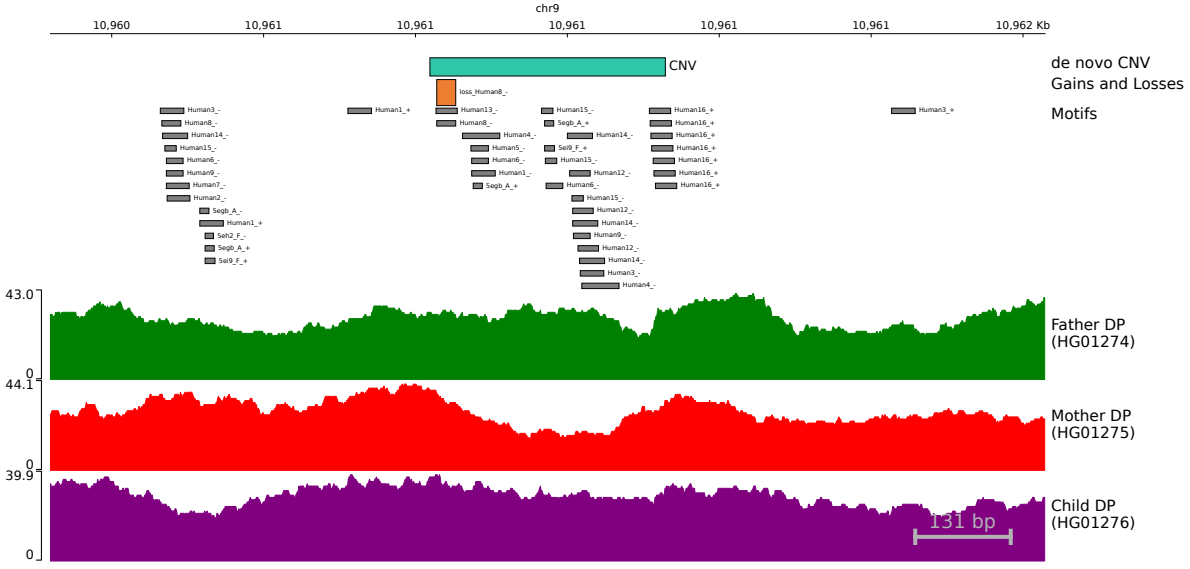


Figure A40, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

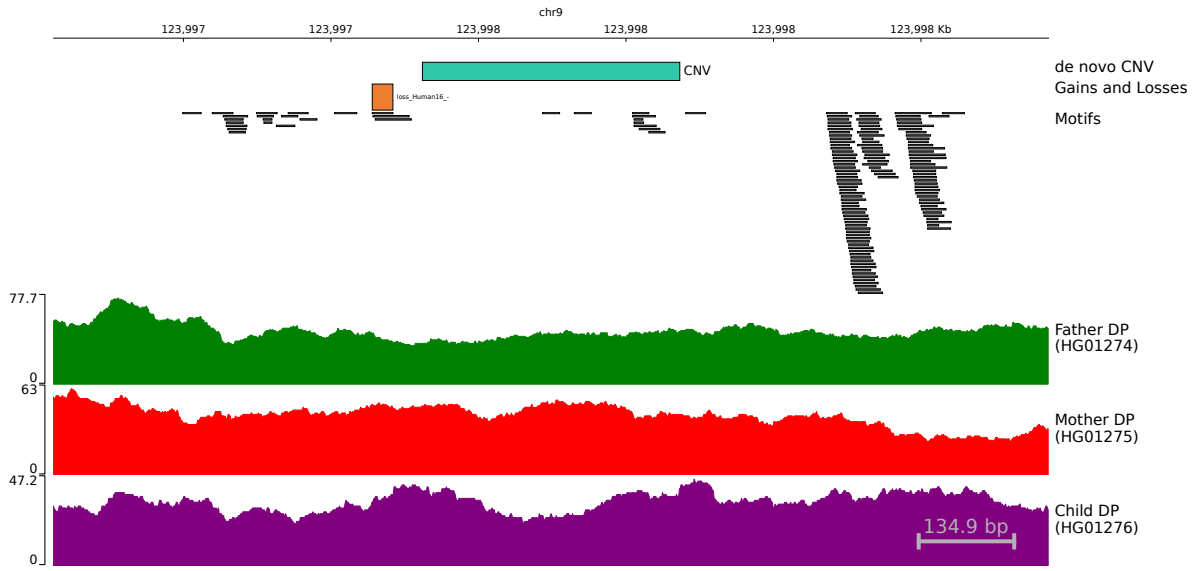


Figure A41, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

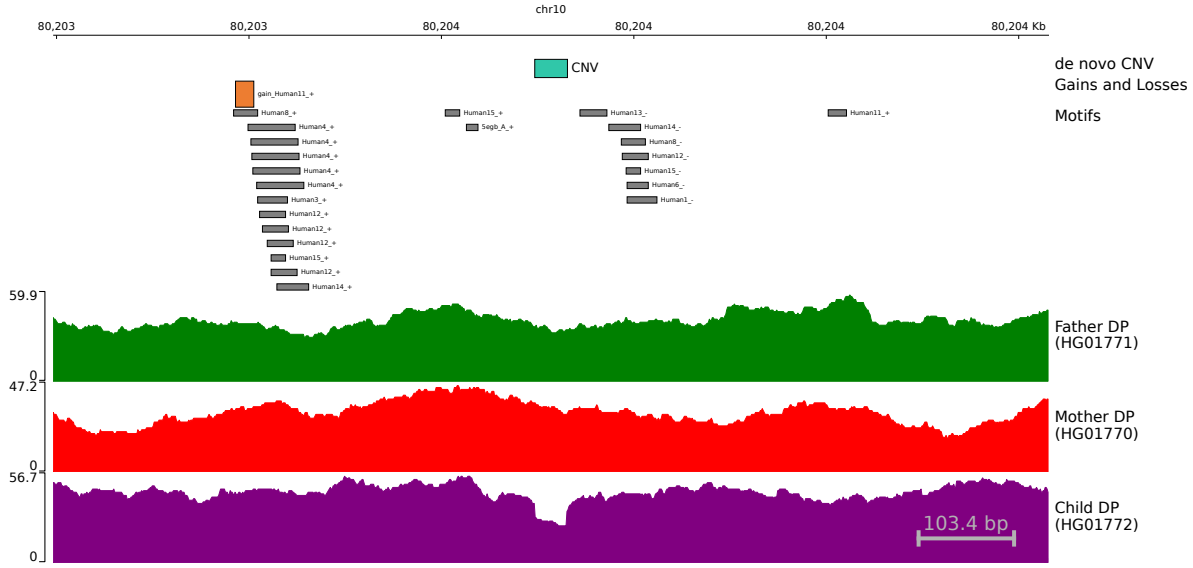


Figure A42, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

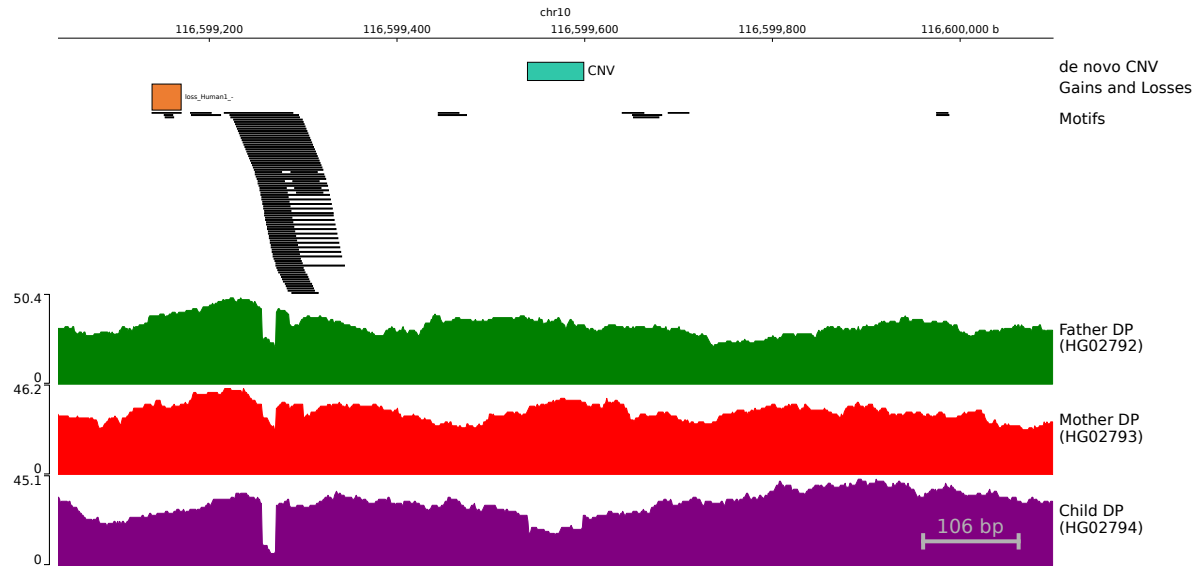


Figure A43, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

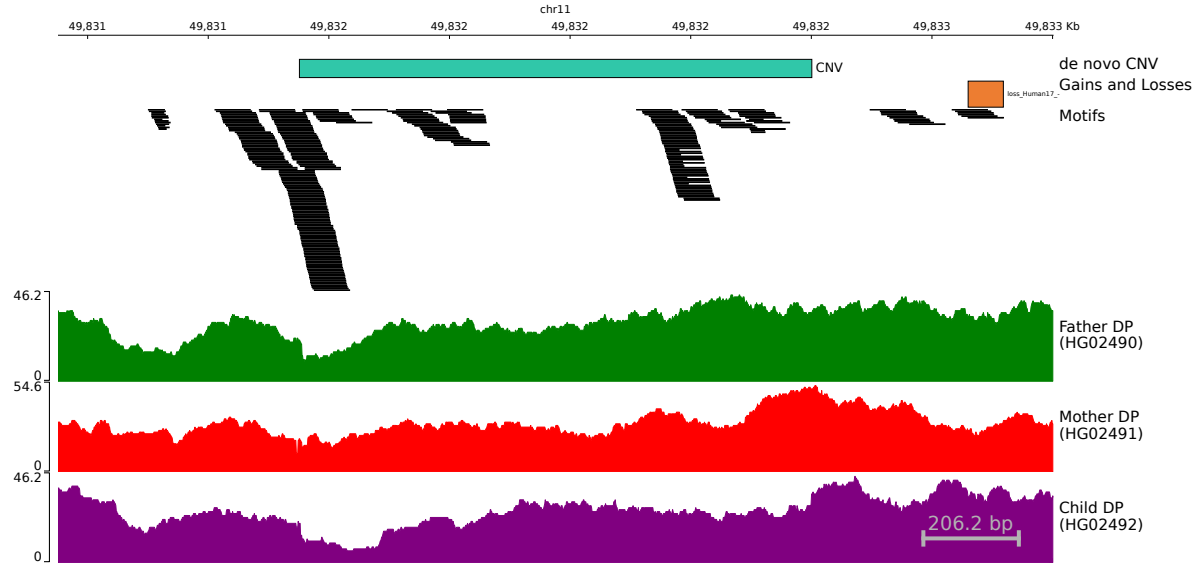


Figure A44, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

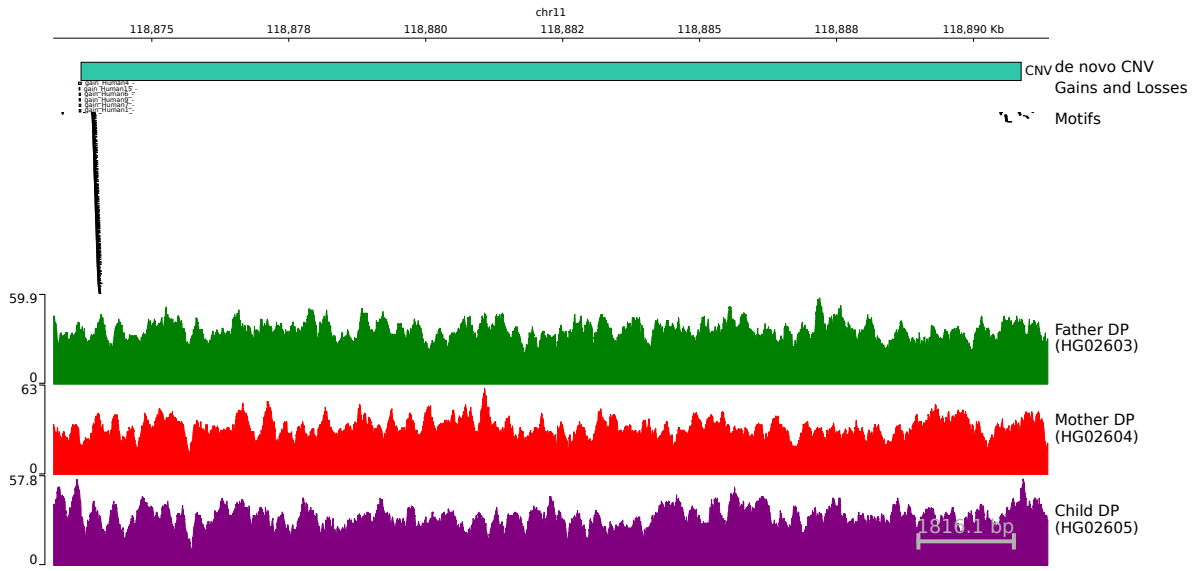


Figure A45, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

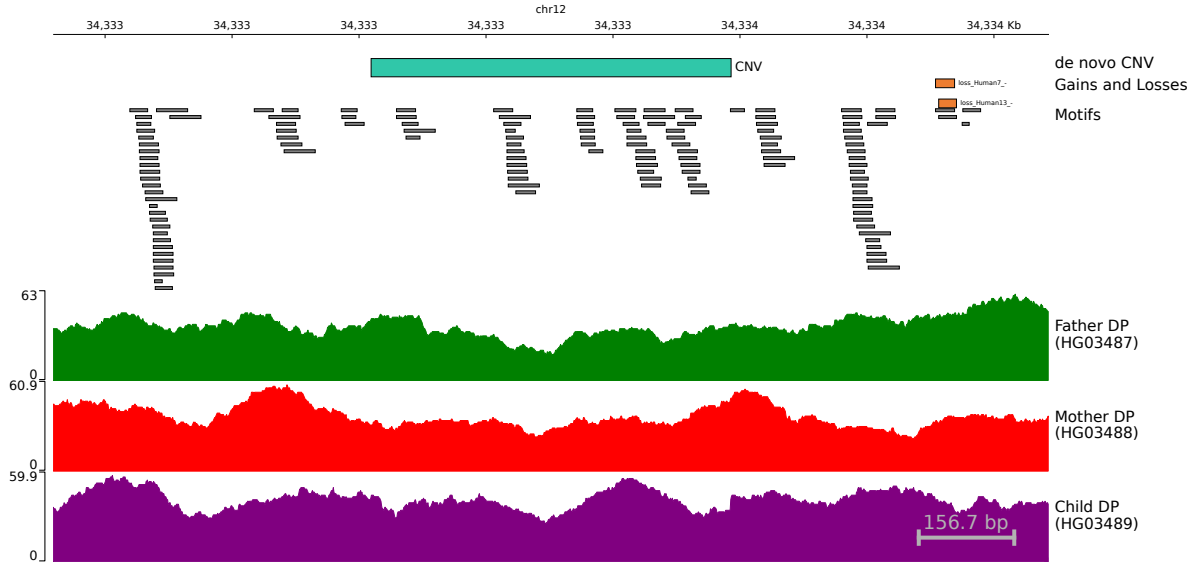


Figure A46, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

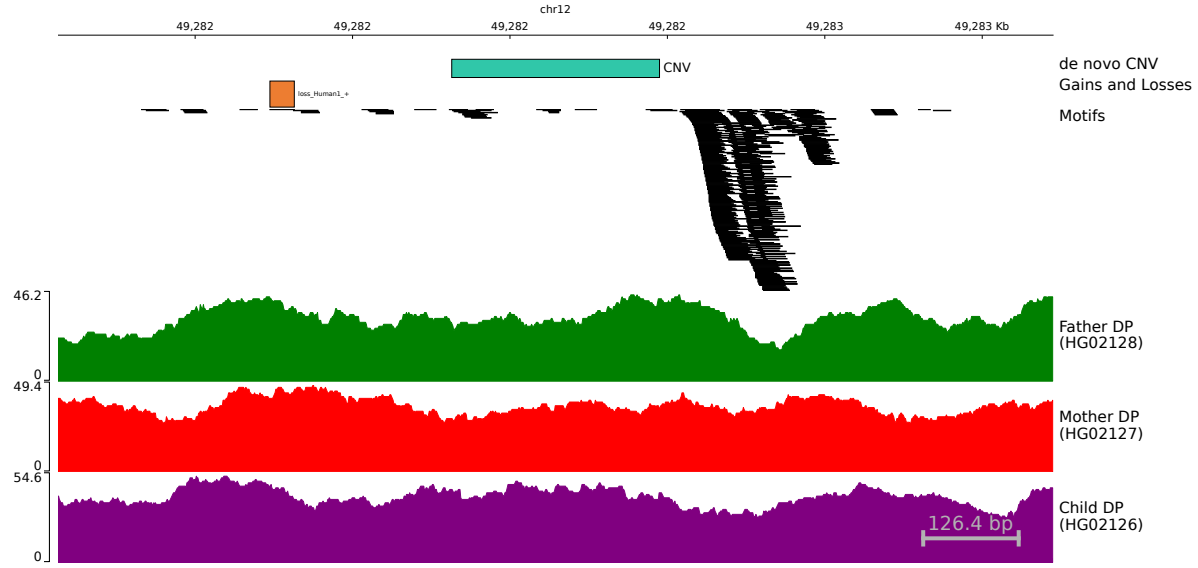


Figure A47, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

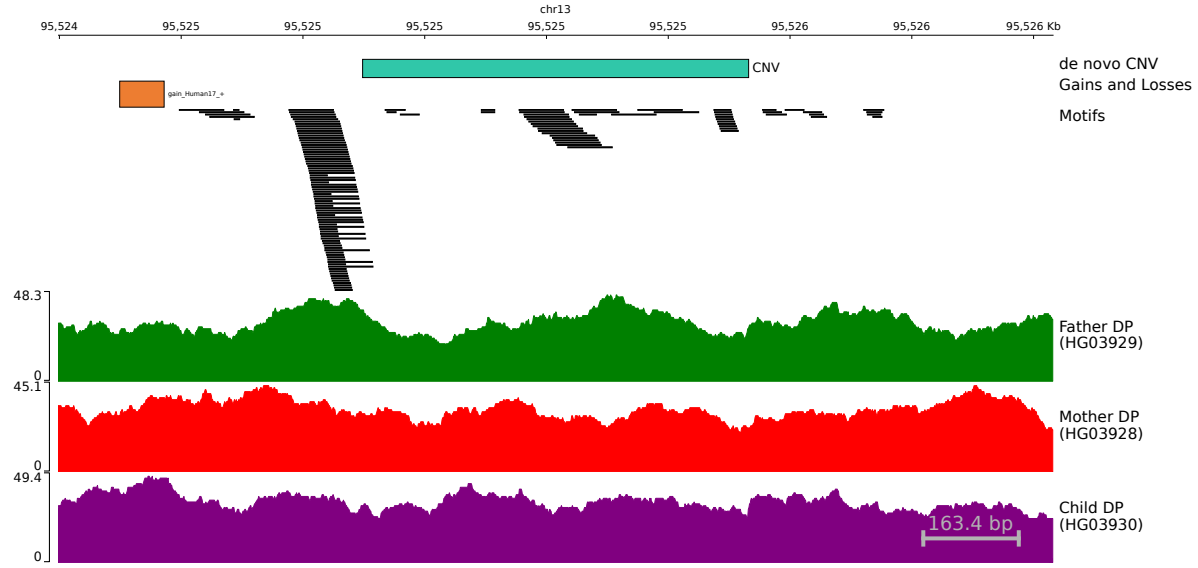


Figure A48, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

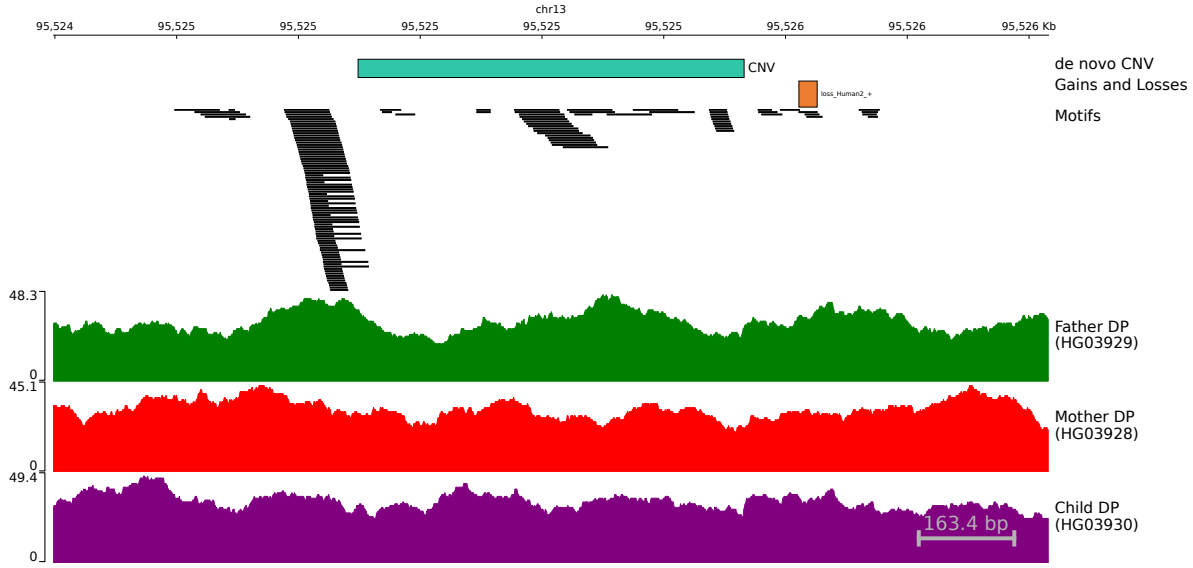


Figure A49, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

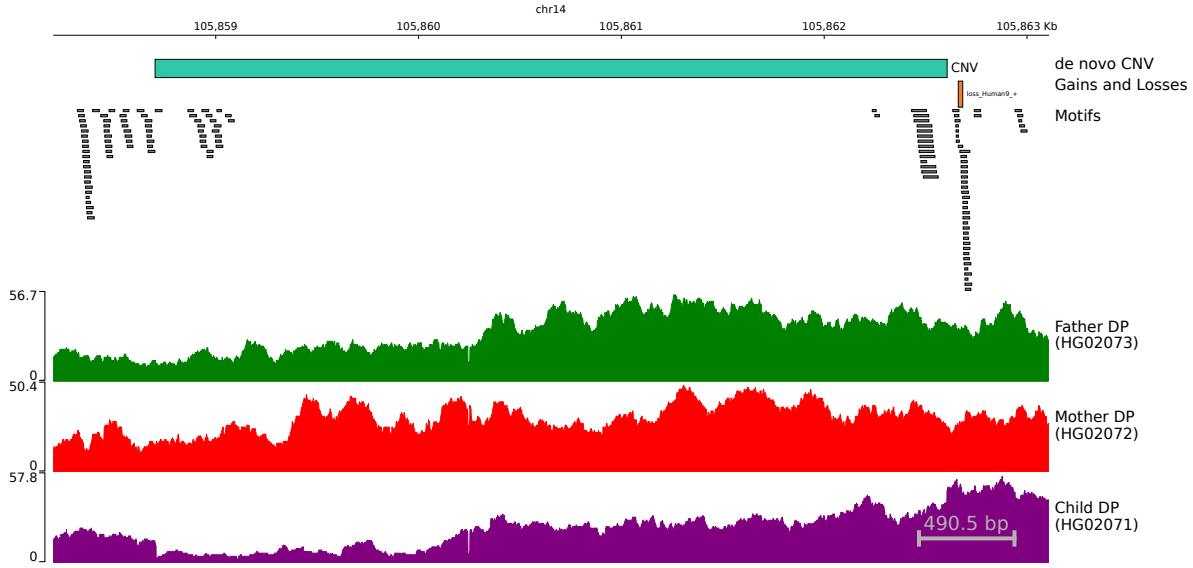


Figure A50, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

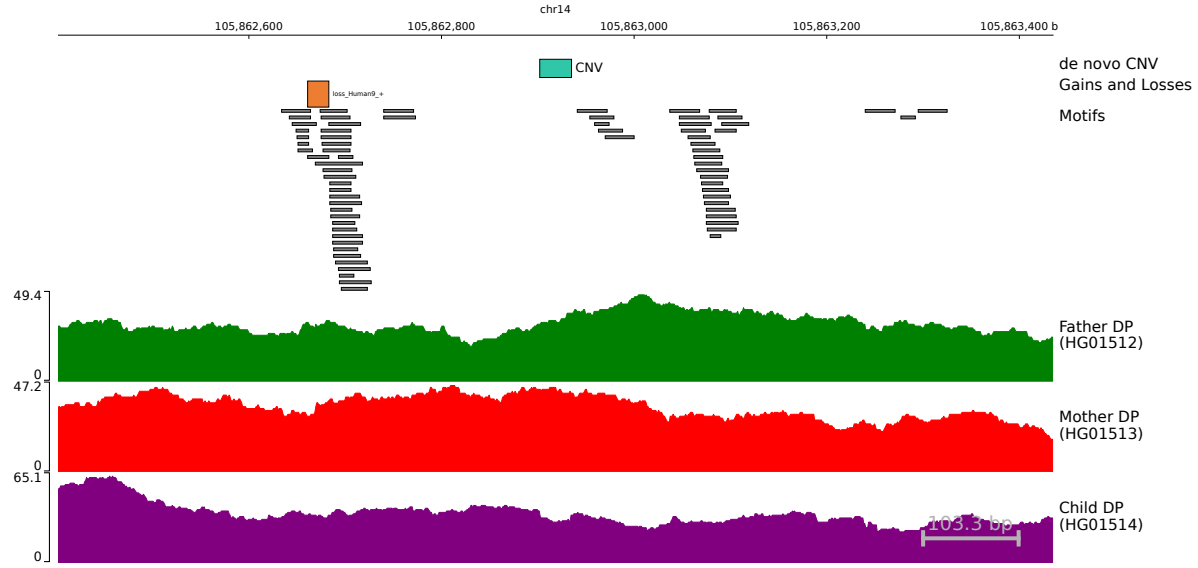


Figure A51, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

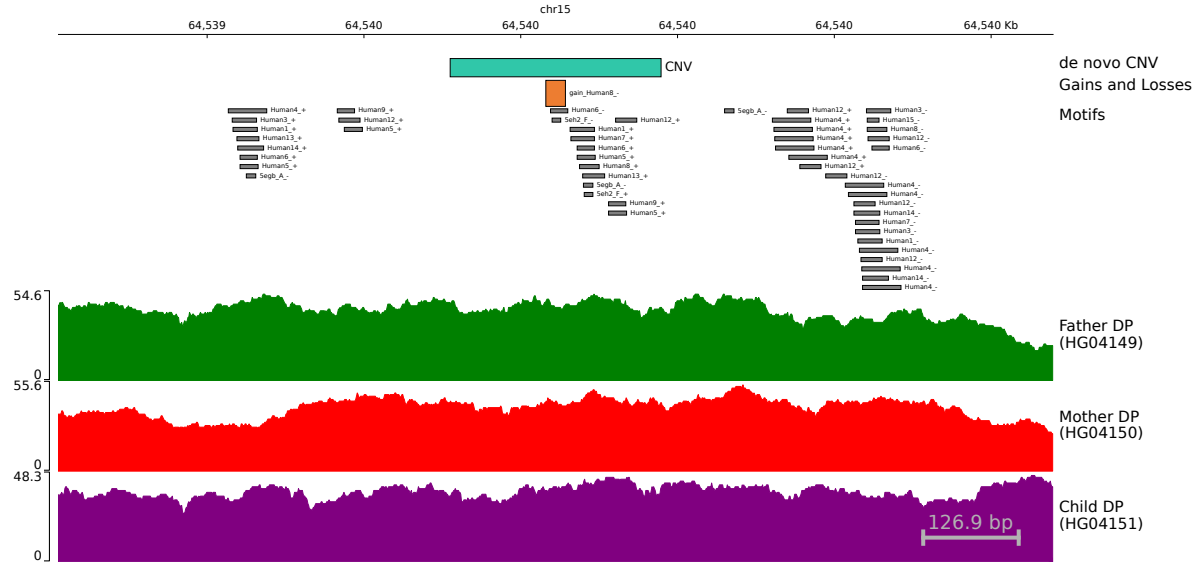


Figure A52, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

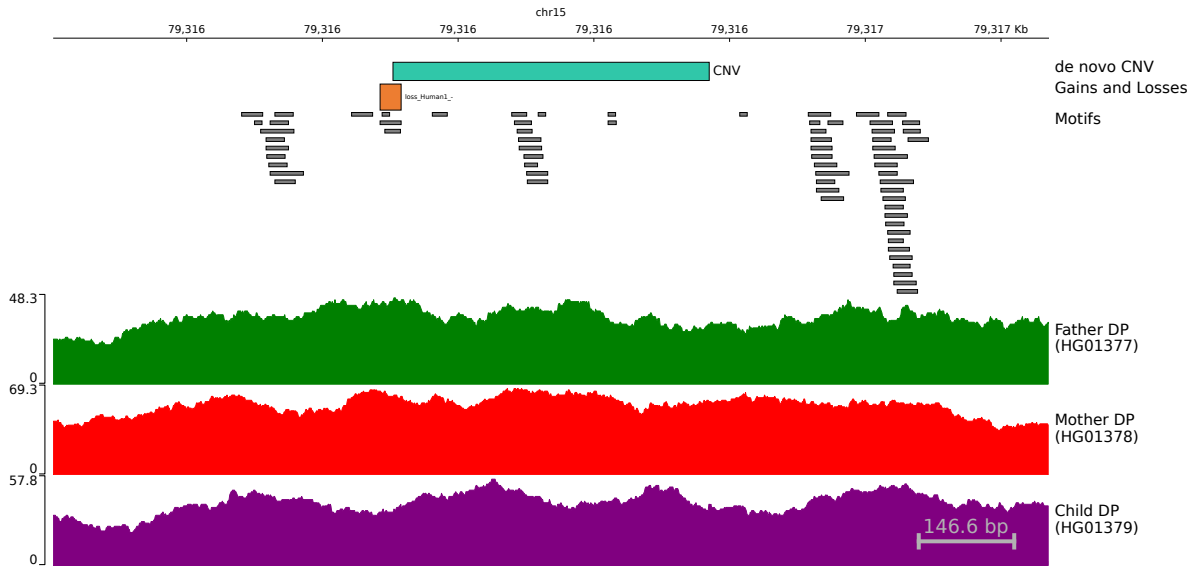


Figure A53, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

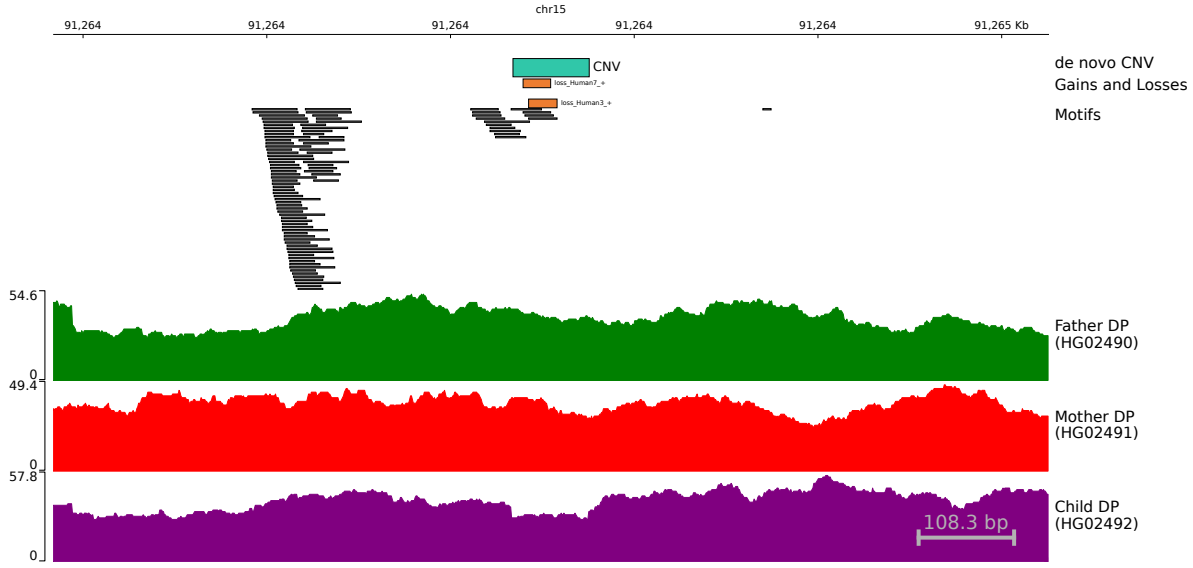


Figure A54, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

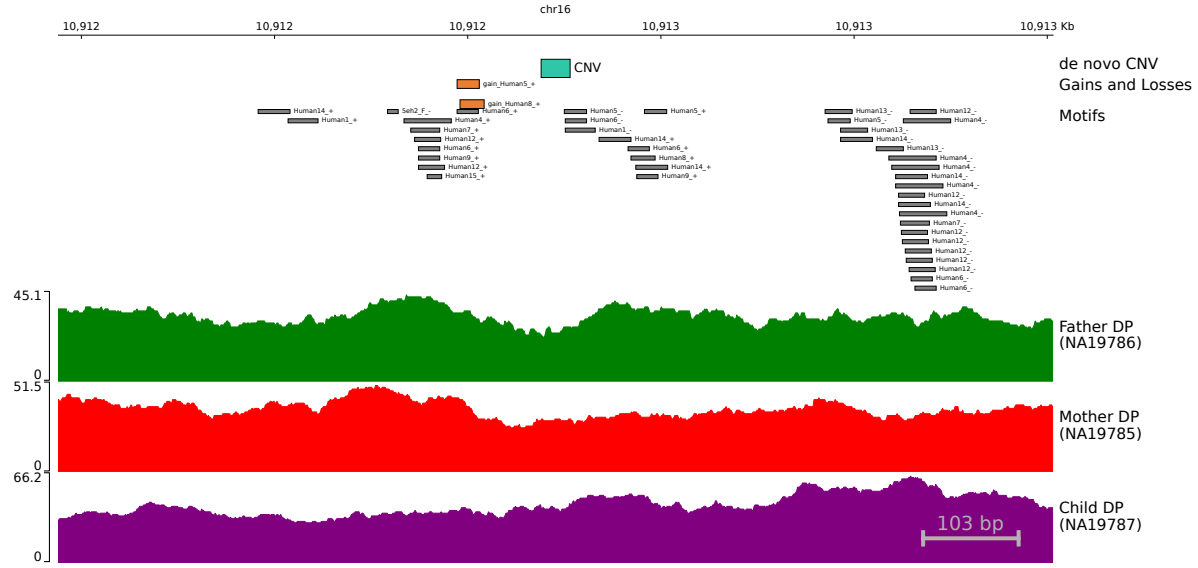


Figure A55, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

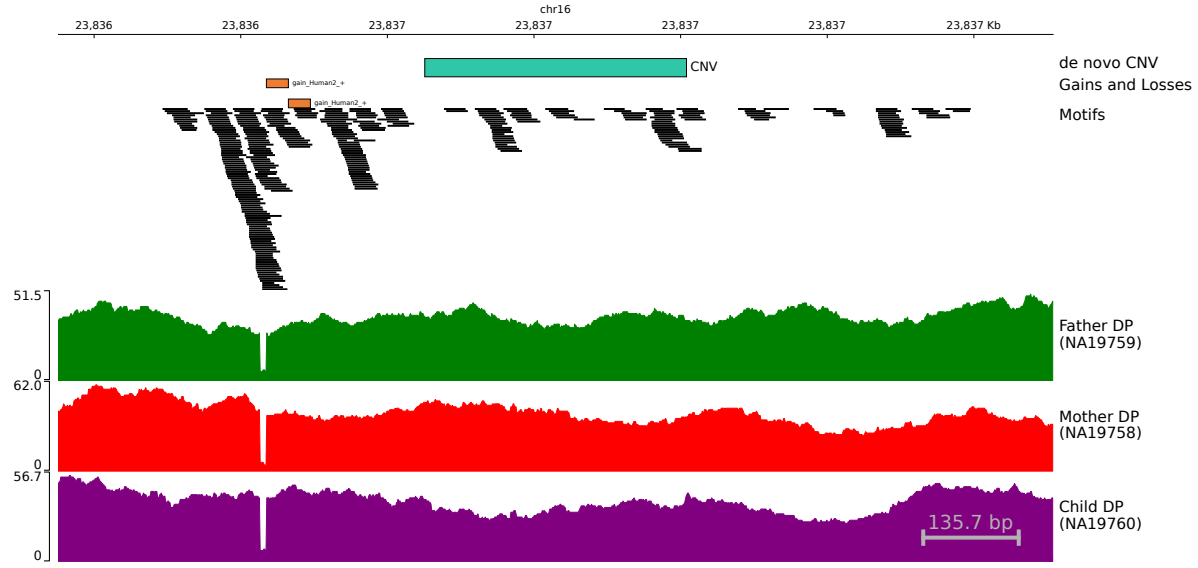


Figure A56, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

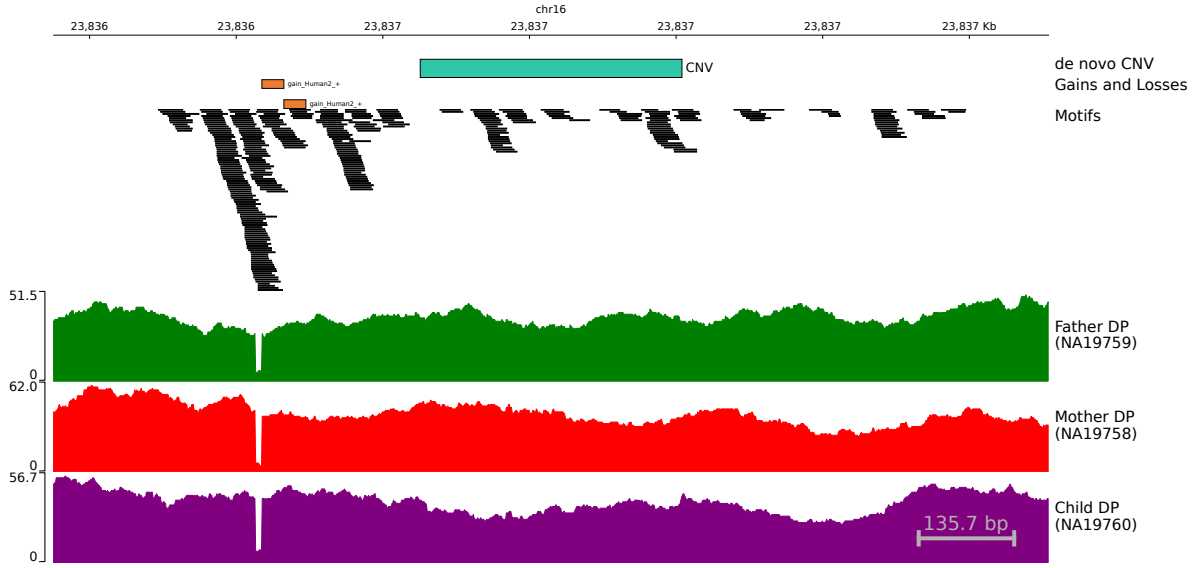


Figure A57, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

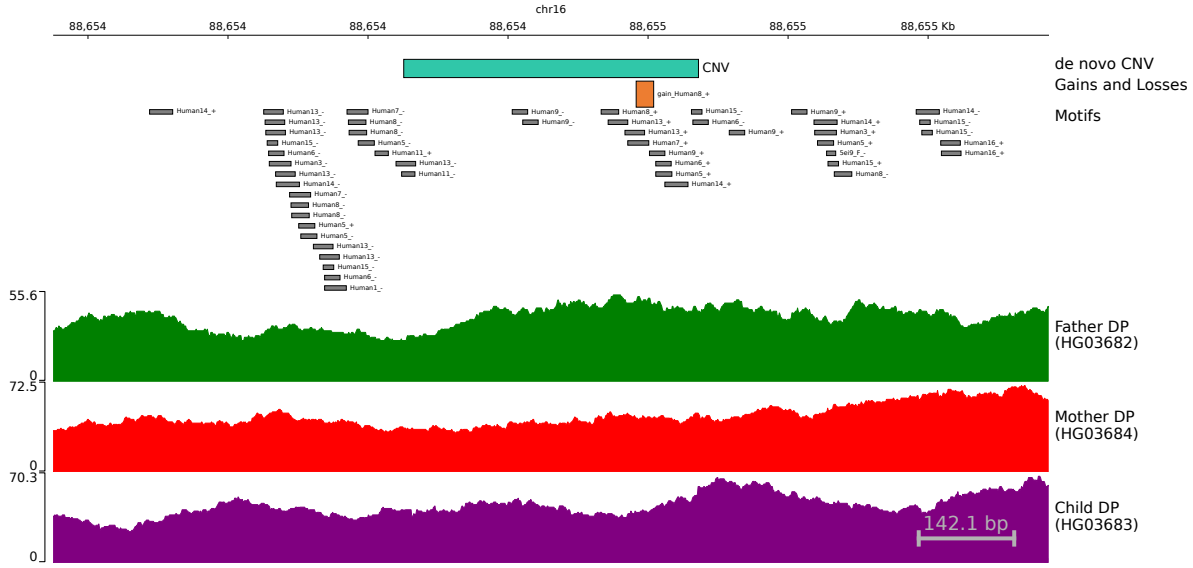


Figure A58, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

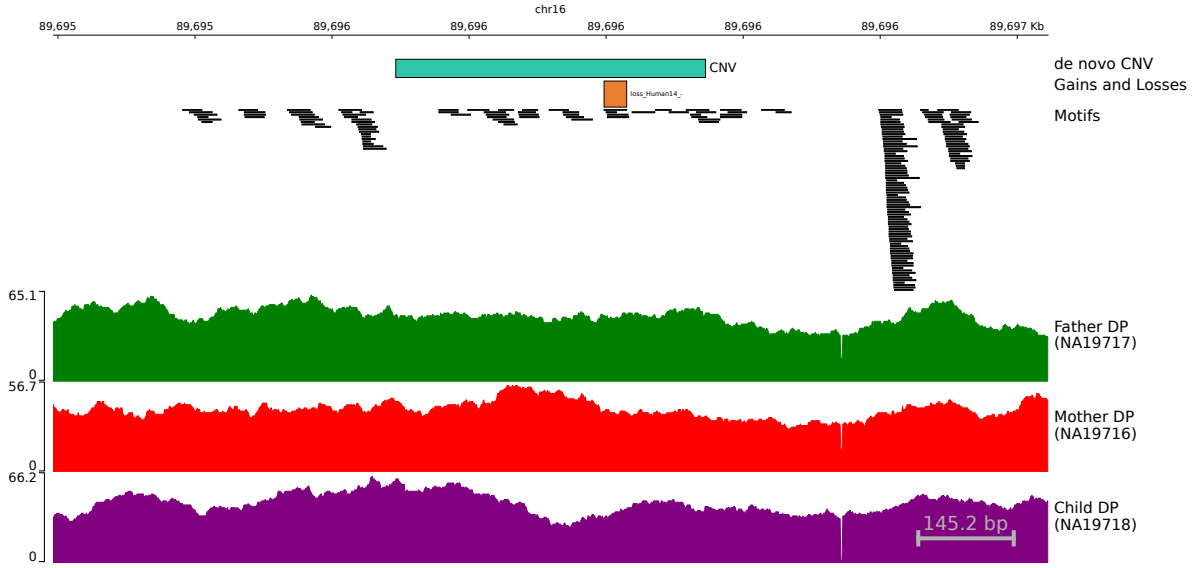


Figure A59, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

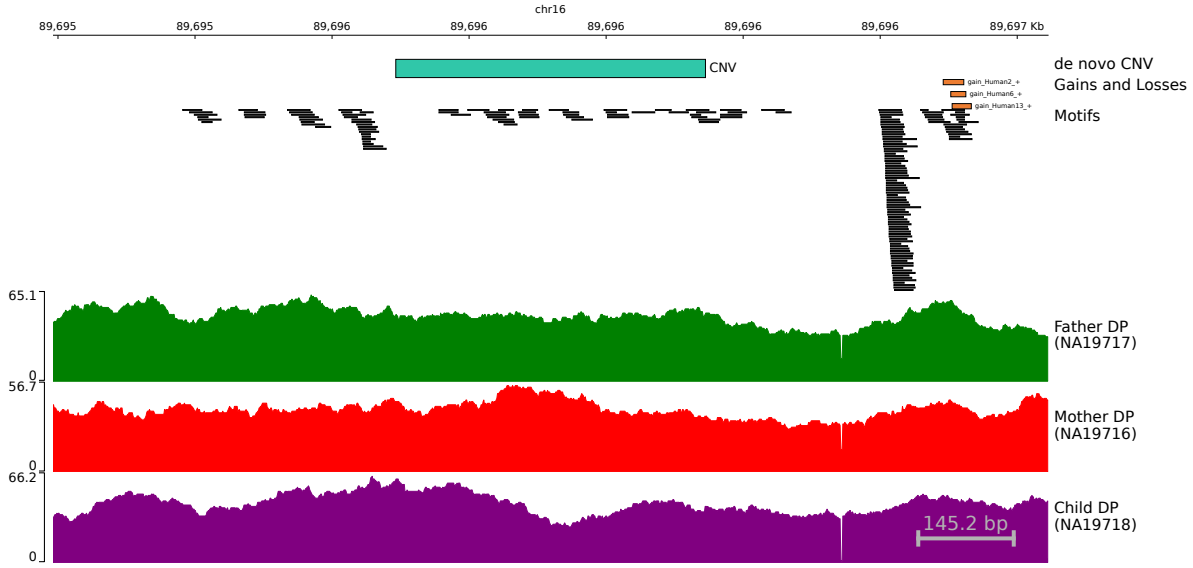


Figure A60, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

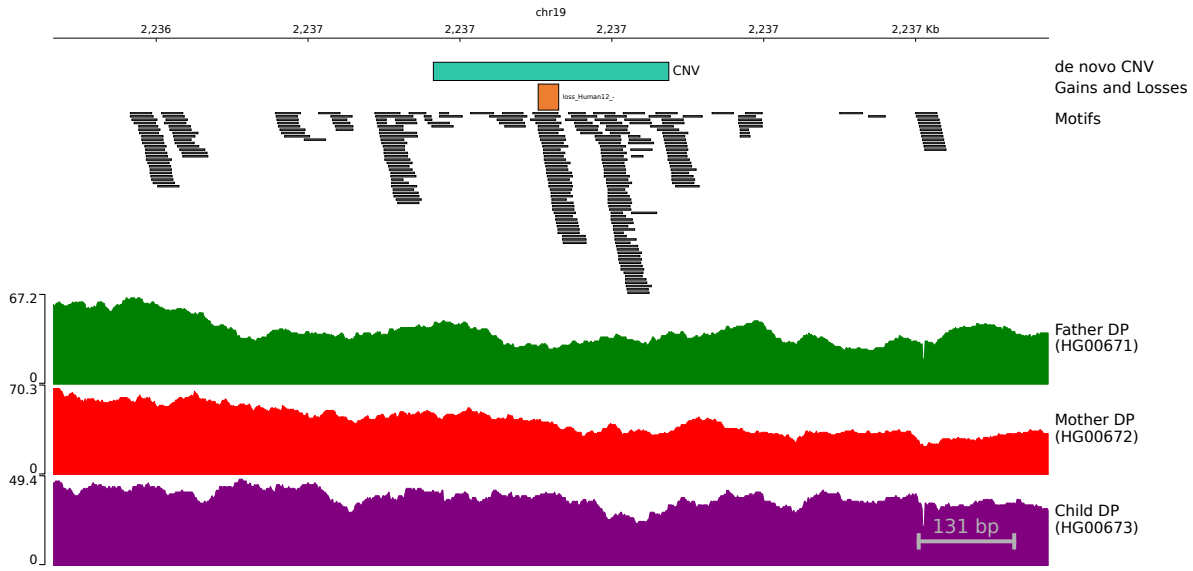


Figure A61, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

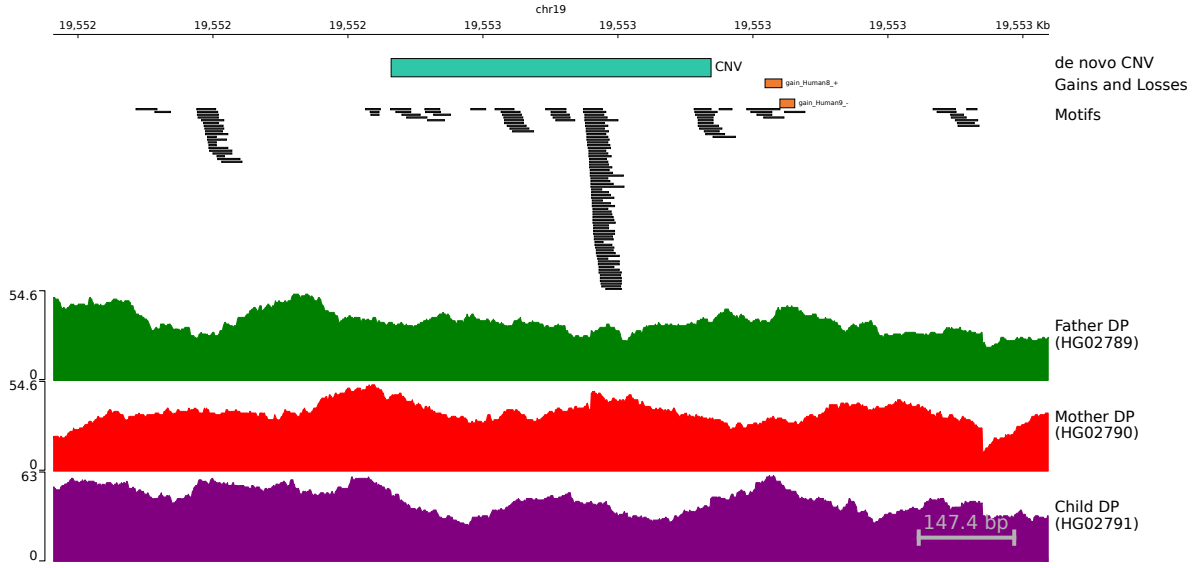


Figure A62, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

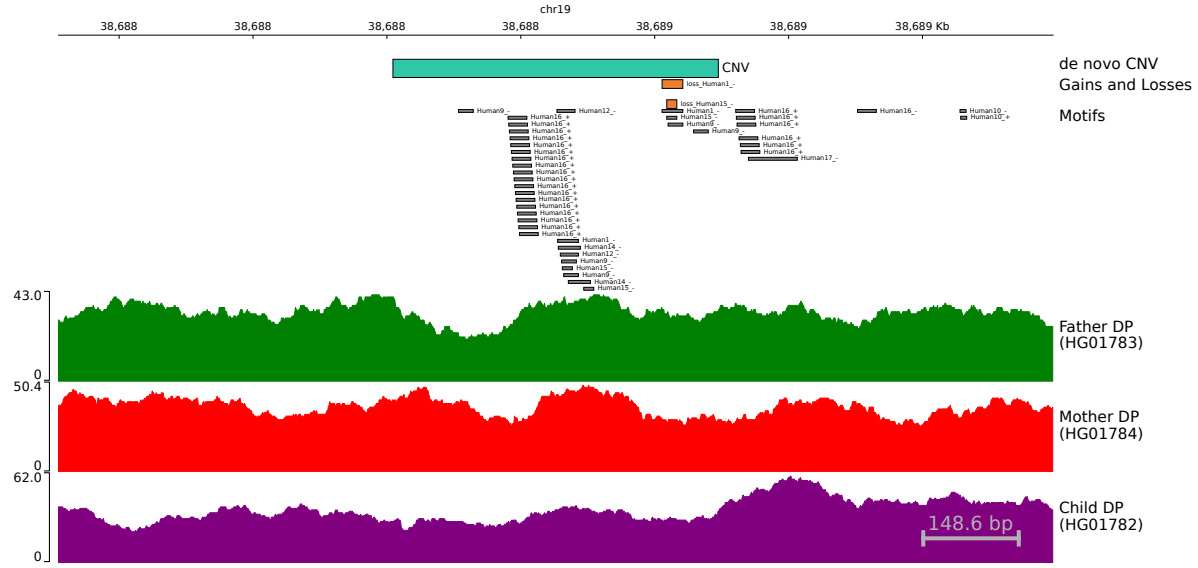


Figure A63, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

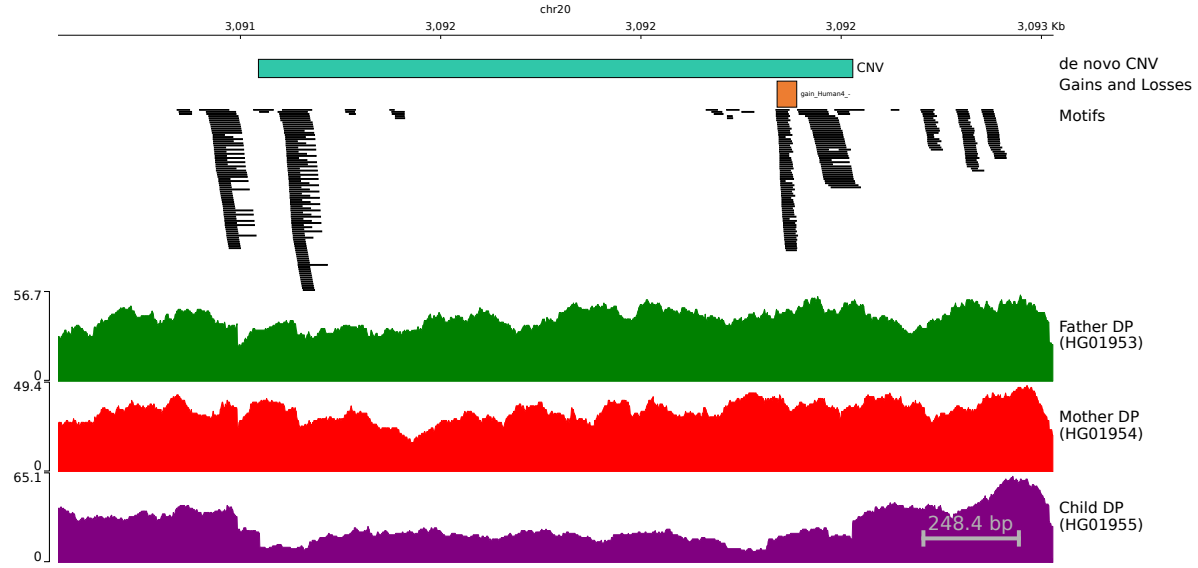


Figure A64, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

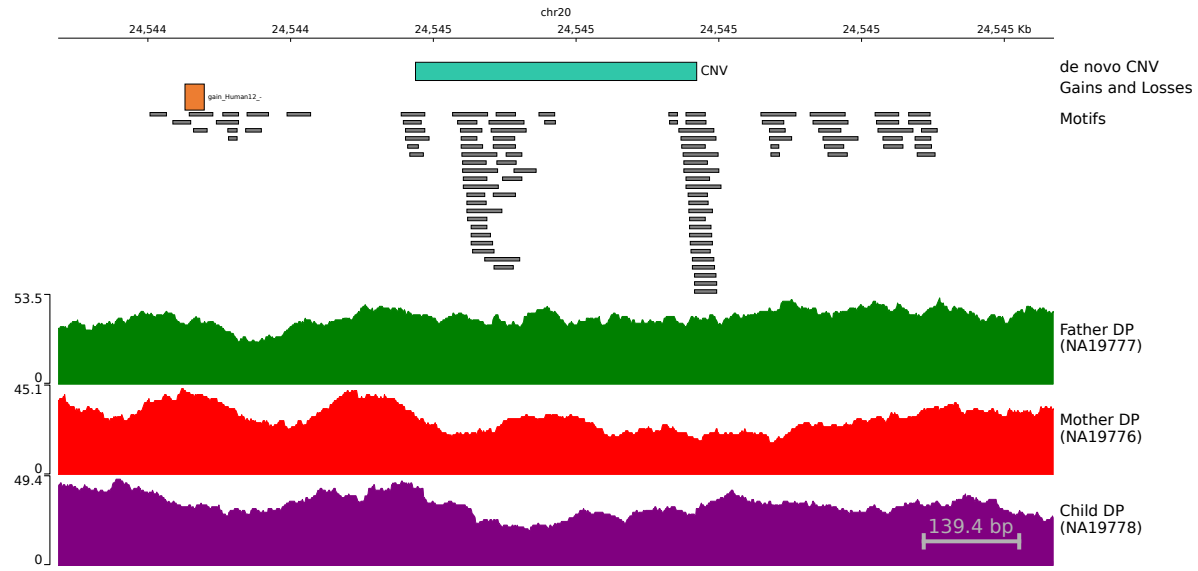


Figure A65, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

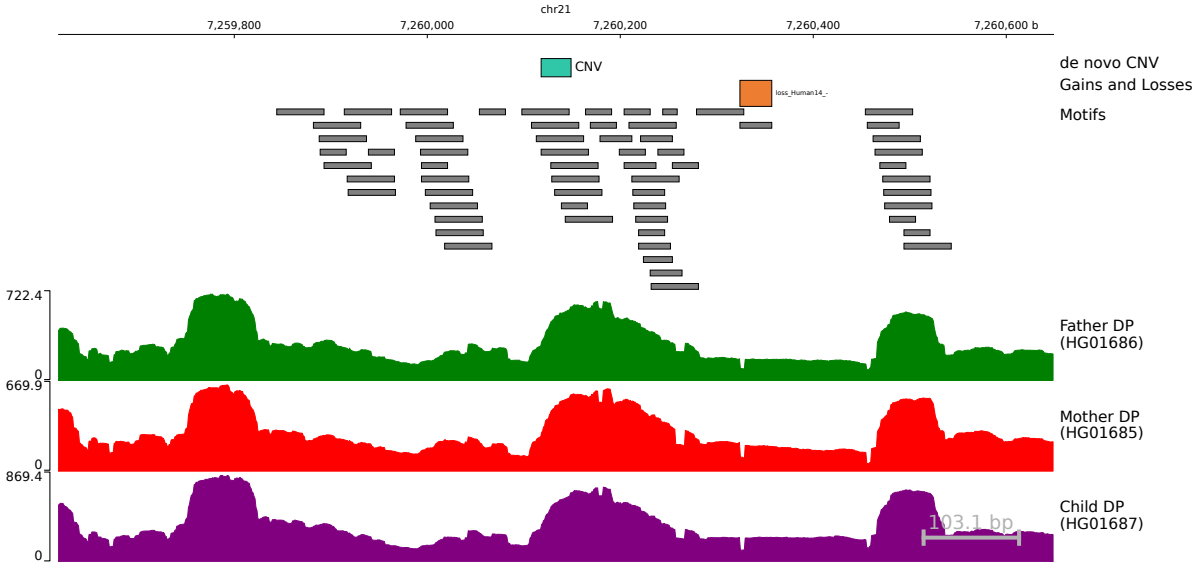


Figure A66, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

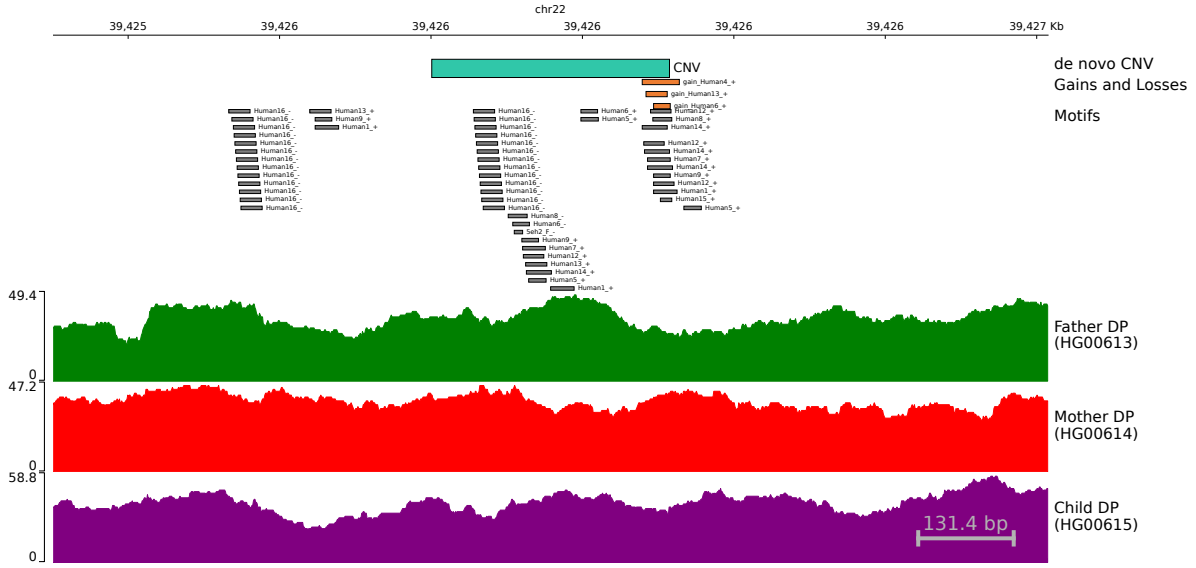


Figure A67, de novo deletion plot. The de novo deletion is displayed, along with their parent's gains/losses (orange) and PRDM9 binding site motifs (grey). BAM read depth for the parent-child trio is also displayed.

Appendix B

ChromoMap plots

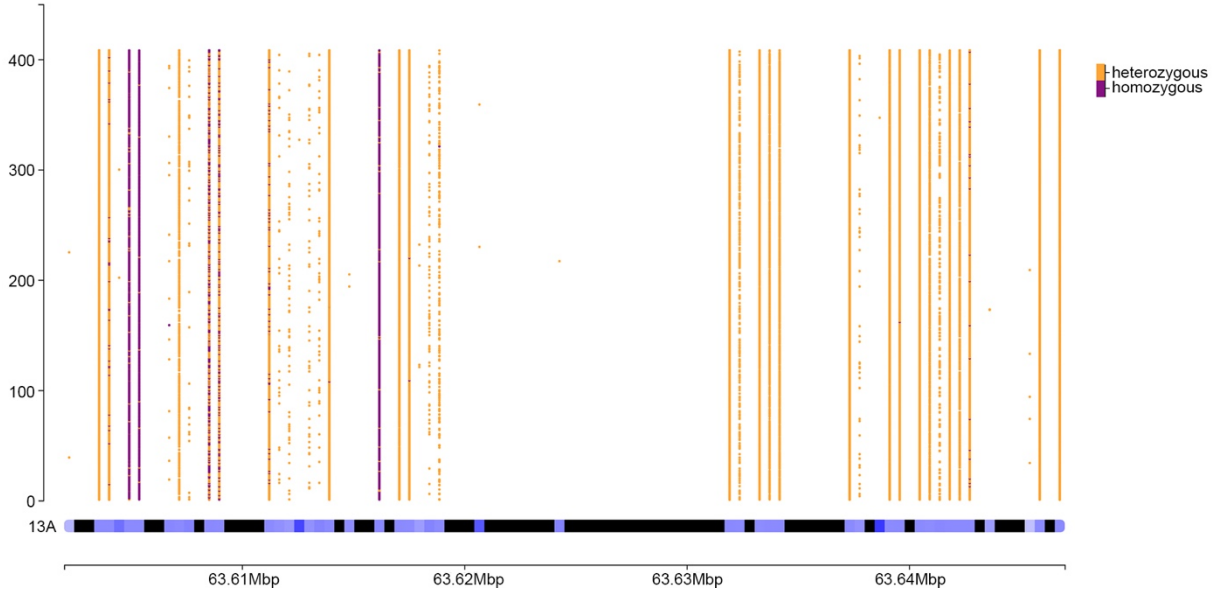


Figure B1, Genotype distribution within the chr13 archaic peak. Each individual has been assigned a number, 0-407, displayed on the y-axis. For each private variant that is present in an individual, a yellow or purple dot is assigned to their corresponding number, depending on whether their genotype is homozygous (purple) or heterozygous (yellow).

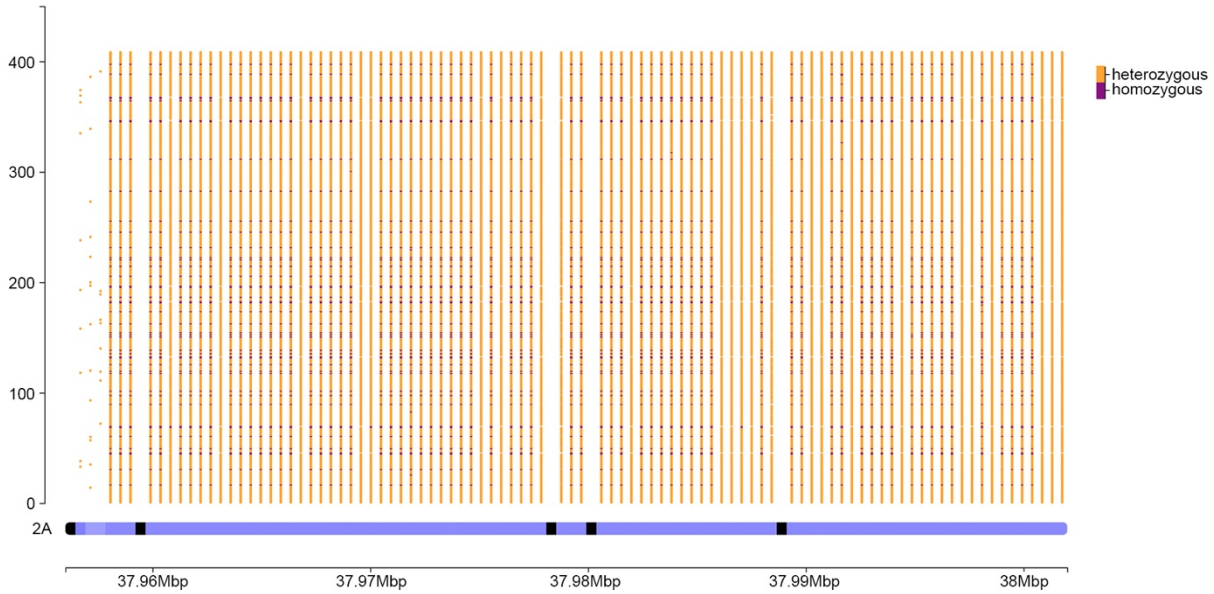


Figure B2, Genotype distribution within the chr2 archaic peak. Each individual has been assigned a number, 0-407, displayed on the y-axis. For each private variant that is present in an individual, a yellow or purple dot is assigned to their corresponding number, depending on whether their genotype is homozygous (purple) or heterozygous (yellow).

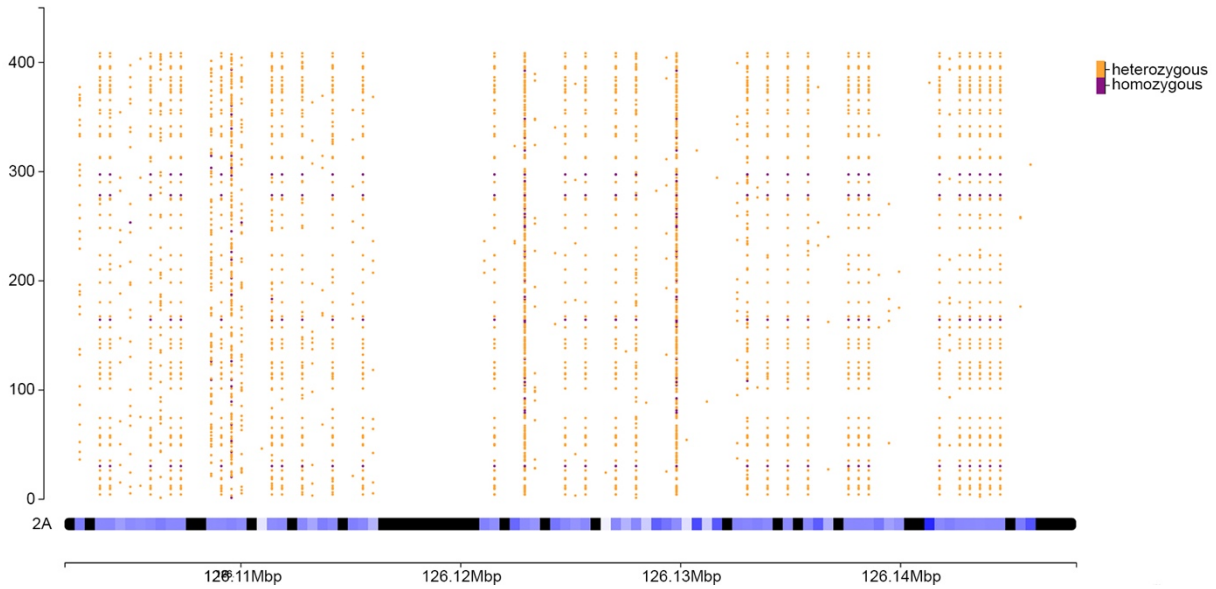


Figure B3, Genotype distribution in a random control region. Each individual has been assigned a number, 0-407, displayed on the y-axis. For each private variant that is present in an individual, a yellow or purple dot is assigned to their corresponding number, depending on whether their genotype is homozygous (purple) or heterozygous (yellow).

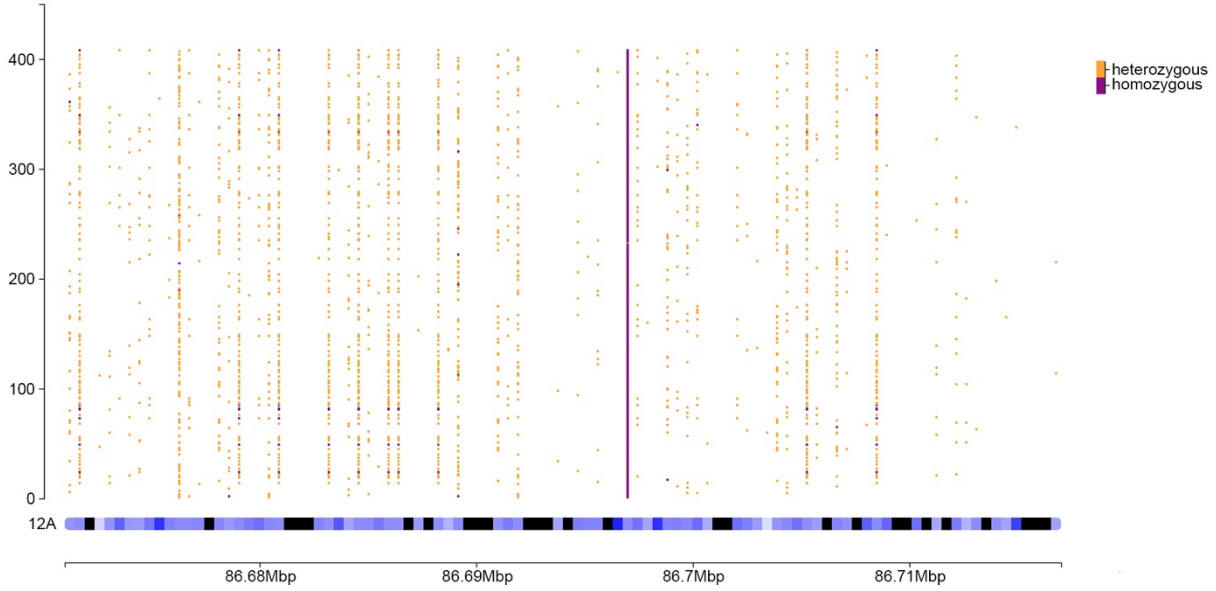


Figure B4, Genotype distribution in a random control region. Each individual has been assigned a number, 0-407, displayed on the y-axis. For each private variant that is present in an individual, a yellow or purple dot is assigned to their corresponding number, depending on whether their genotype is homozygous (purple) or heterozygous (yellow).

Archaic Overlap PCA plots

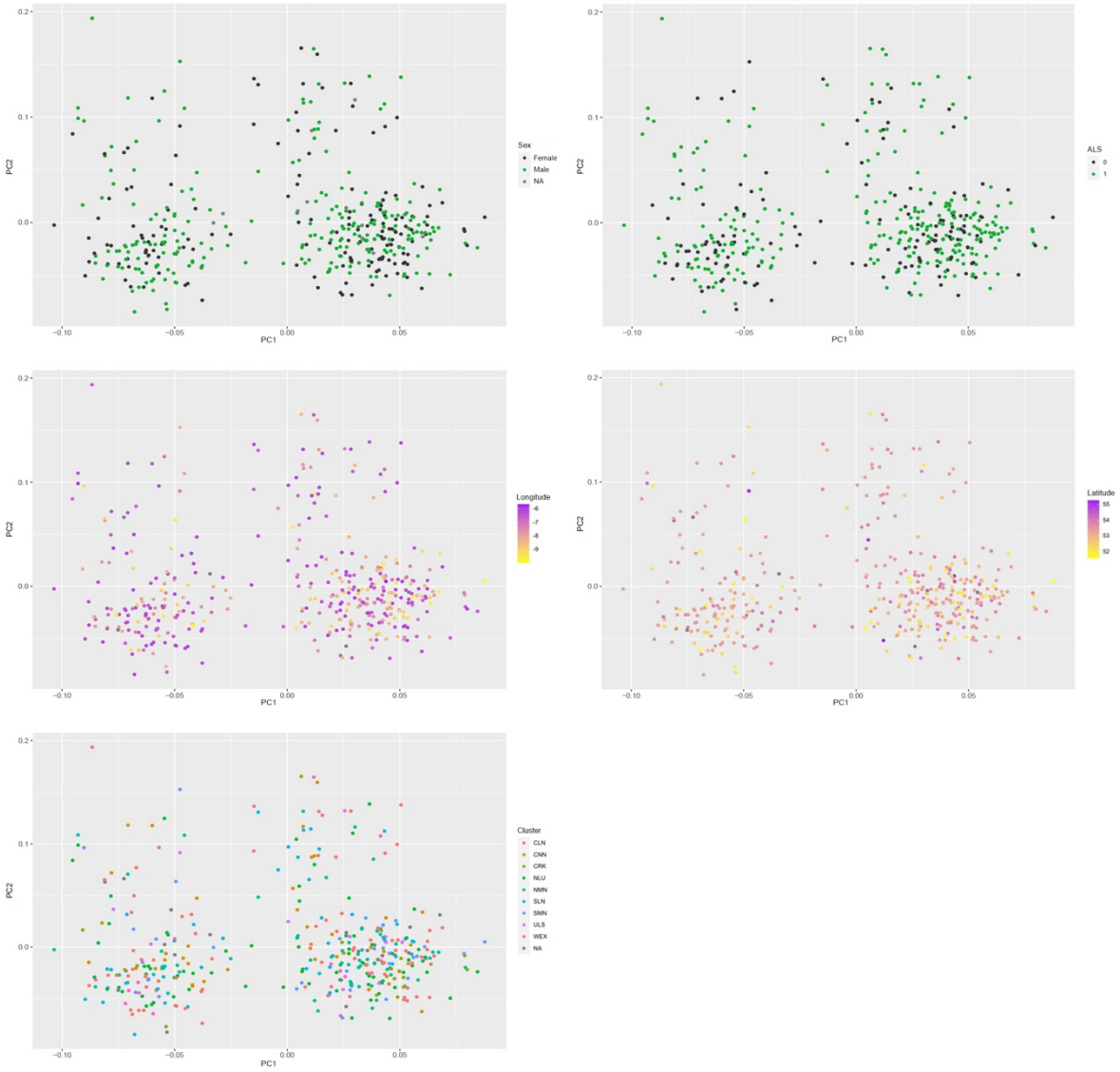


Figure B5, Archaic overlap PCA plots for the Irish dataset. These plots show the distribution of individuals along PC1 and PC2, where the individuals are coloured by their archaic proportion, sex, ALS status, longitude or latitude coordinates, or their genetic cluster.

Private Variant PCA plots

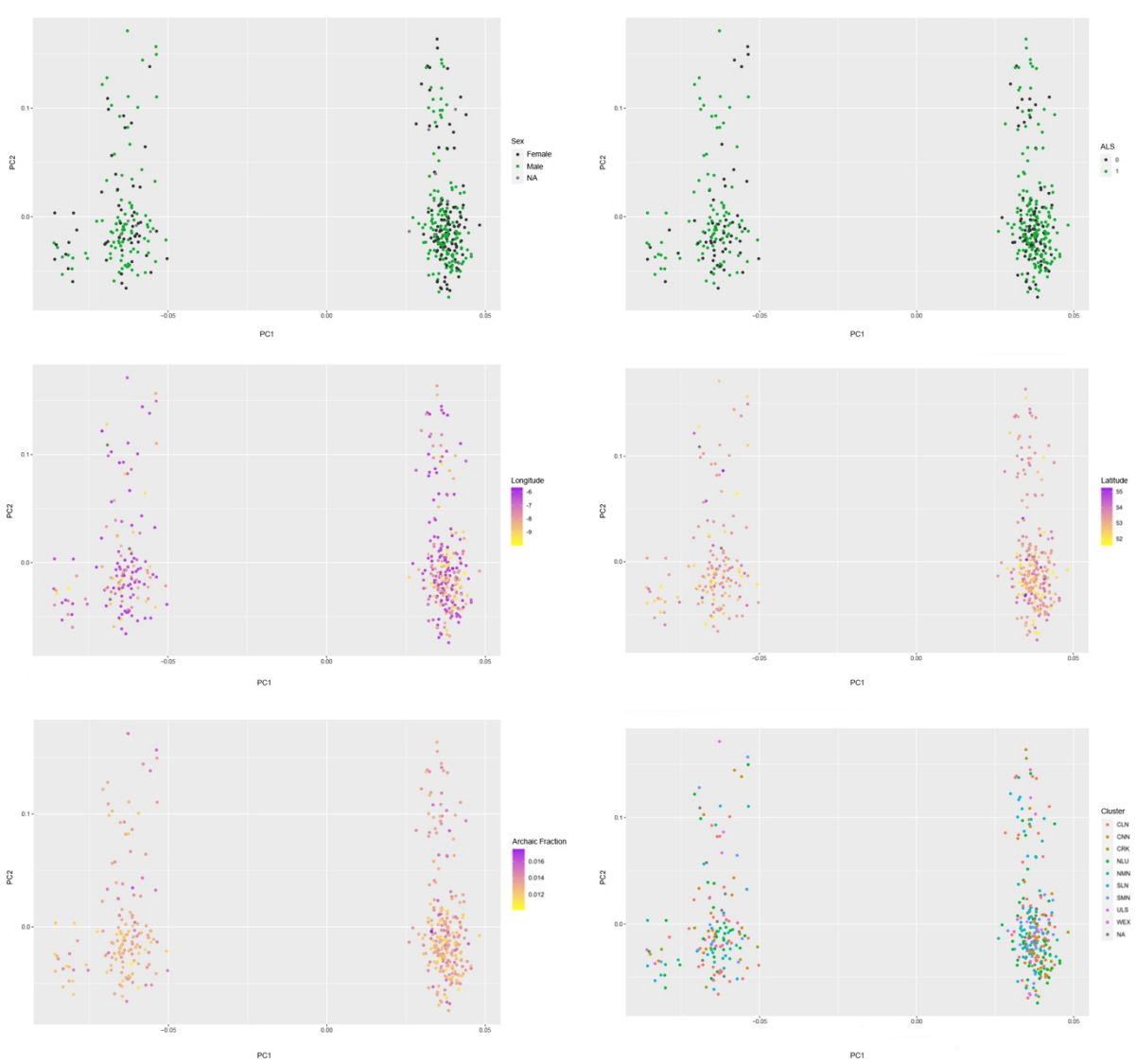


Figure B6, Private Variant PCA plots for the Irish dataset. These plots show the distribution of individuals along PC1 and PC2, where the individuals are coloured by their archaic proportion, sex, ALS status, longitude or latitude coordinates, or their genetic cluster.

Cluster PCA plots

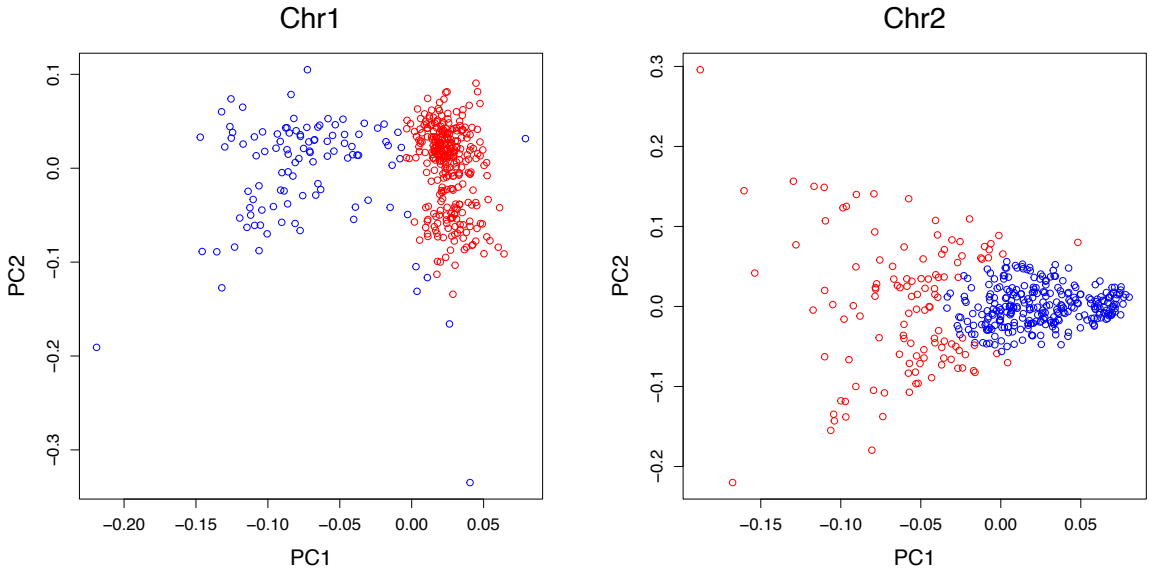


Figure B7, Cluster PCA plots. These plots show the distribution of individuals along PC1 and PC2 when the private variant PCA is conducted for chromosome 1 (left) and 2 (right). Individuals are coloured according to their cluster assigned by mclust.

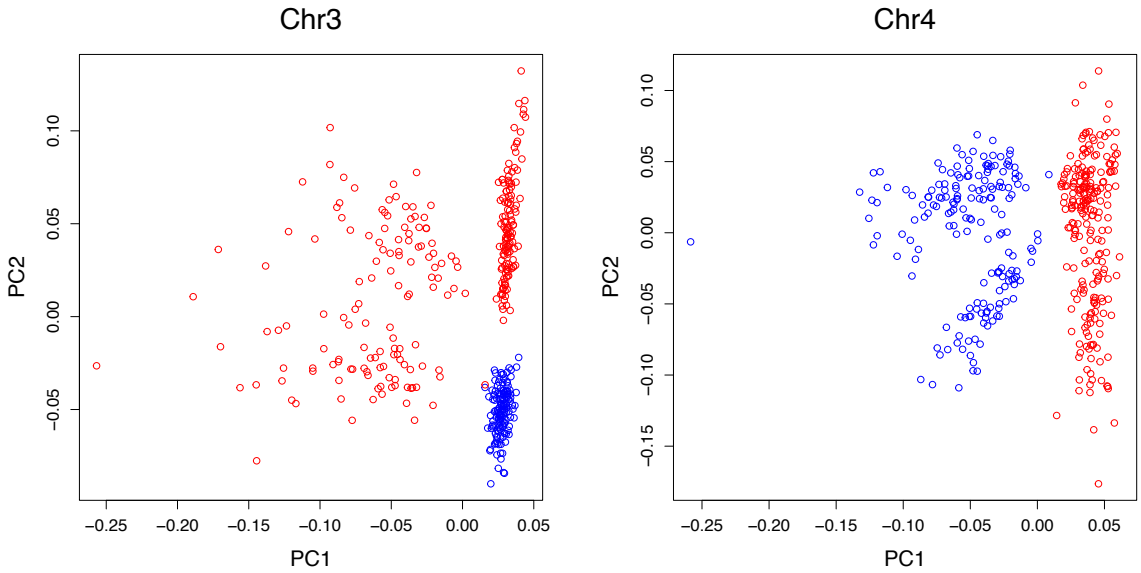


Figure B8, Cluster PCA plots. These plots show the distribution of individuals along PC1 and PC2 when the private variant PCA is conducted for chromosome 3 (left) and 4 (right). Individuals are coloured according to their cluster assigned by mclust.

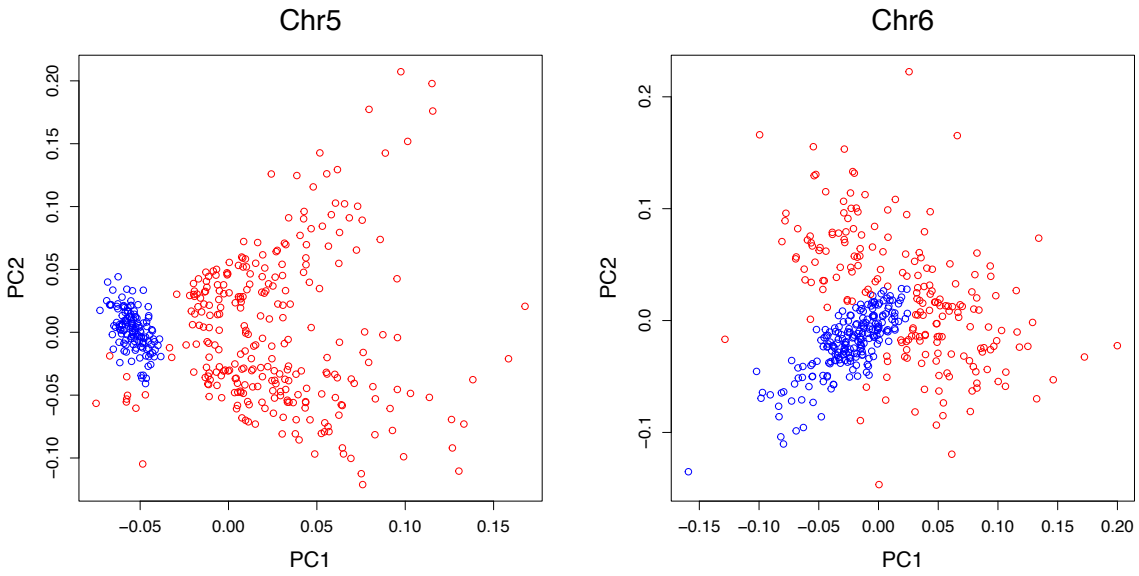


Figure B9, Cluster PCA plots. These plots show the distribution of individuals along PC1 and PC2 when the private variant PCA is conducted for chromosome 5 (left) and 6 (right). Individuals are coloured according to their cluster assigned by mclust.

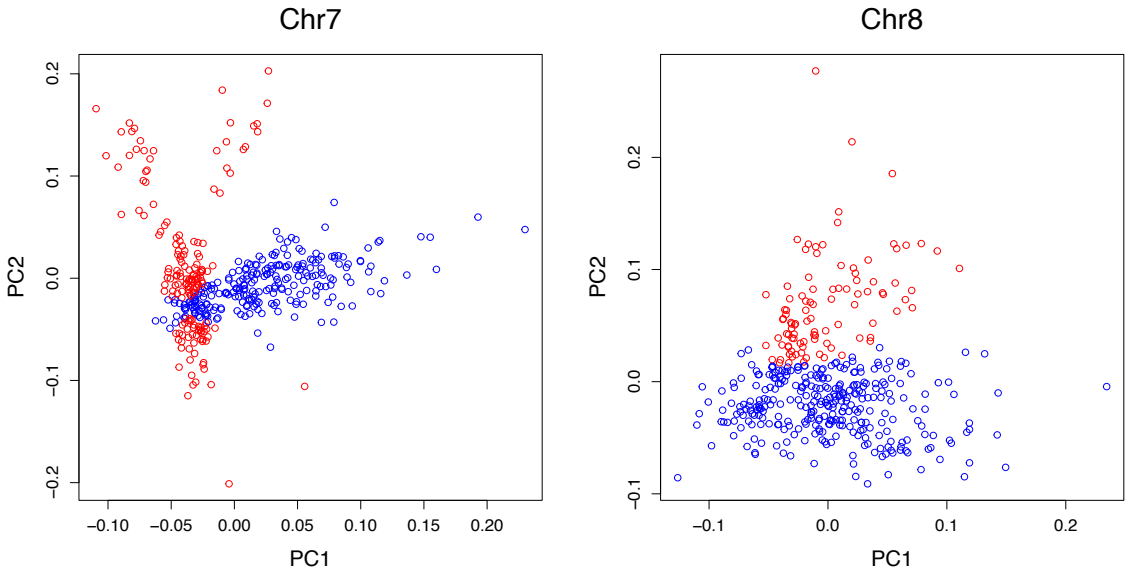


Figure B10, Cluster PCA plots. These plots show the distribution of individuals along PC1 and PC2 when the private variant PCA is conducted for chromosome 7 (left) and 8 (right). Individuals are coloured according to their cluster assigned by mclust.

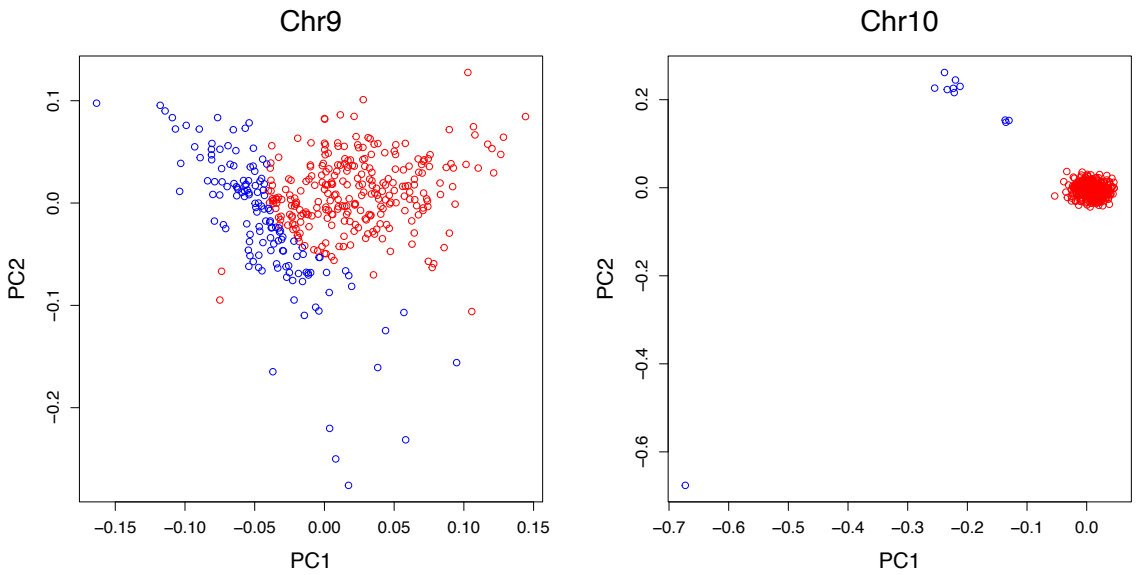


Figure B11, Cluster PCA plots. These plots show the distribution of individuals along PC1 and PC2 when the private variant PCA is conducted for chromosome 9 (left) and 10 (right). Individuals are coloured according to their cluster assigned by mclust.

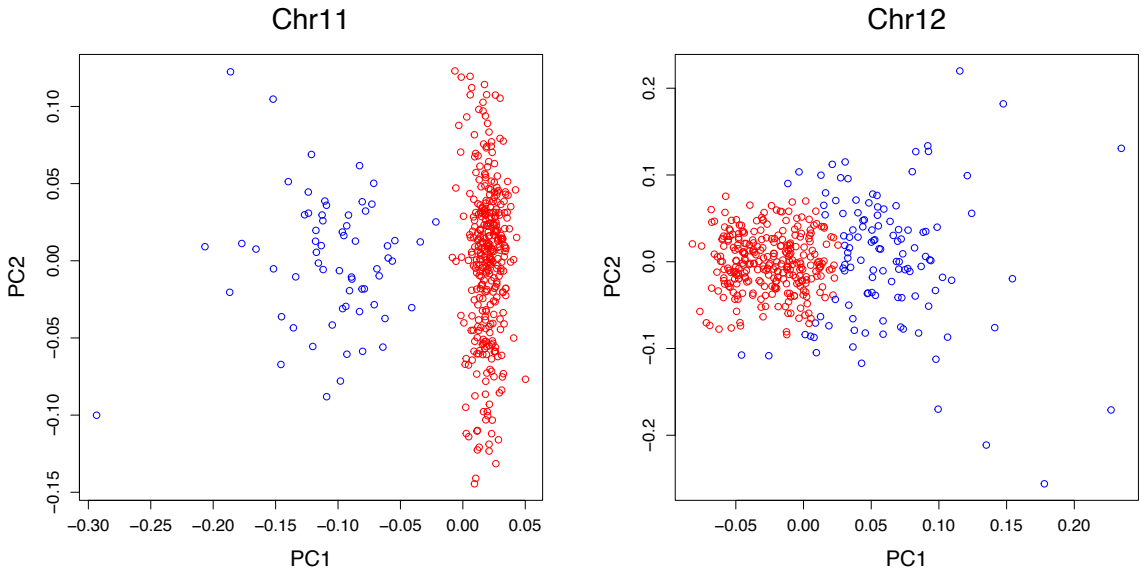


Figure B12, Cluster PCA plots. These plots show the distribution of individuals along PC1 and PC2 when the private variant PCA is conducted for chromosome 11 (left) and 12 (right). Individuals are coloured according to their cluster assigned by mclust.

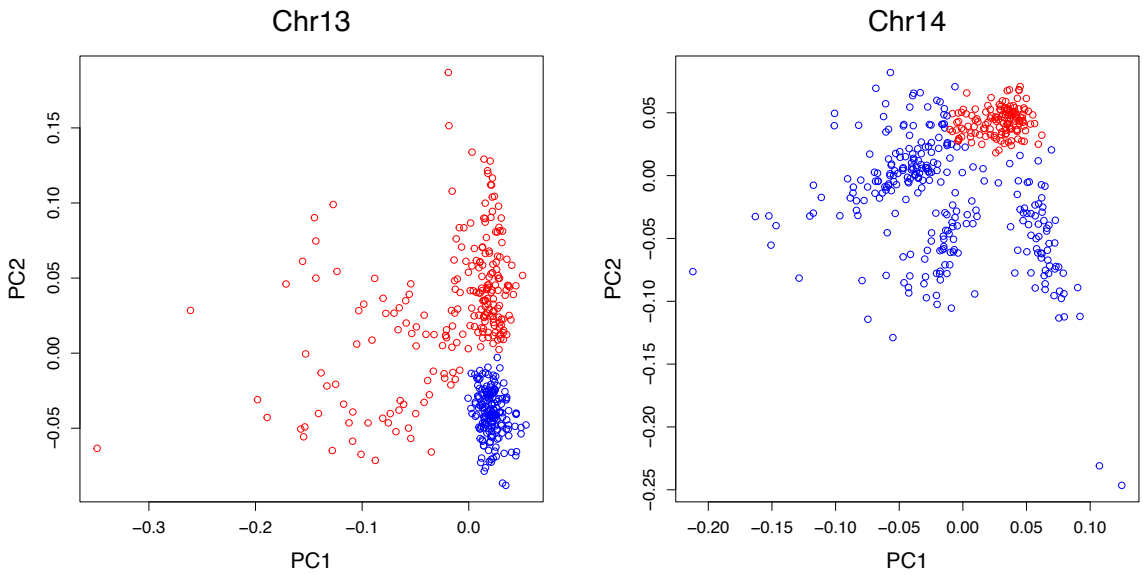


Figure B13, Cluster PCA plots. These plots show the distribution of individuals along PC1 and PC2 when the private variant PCA is conducted for chromosome 13 (left) and 14 (right). Individuals are coloured according to their cluster assigned by mclust.

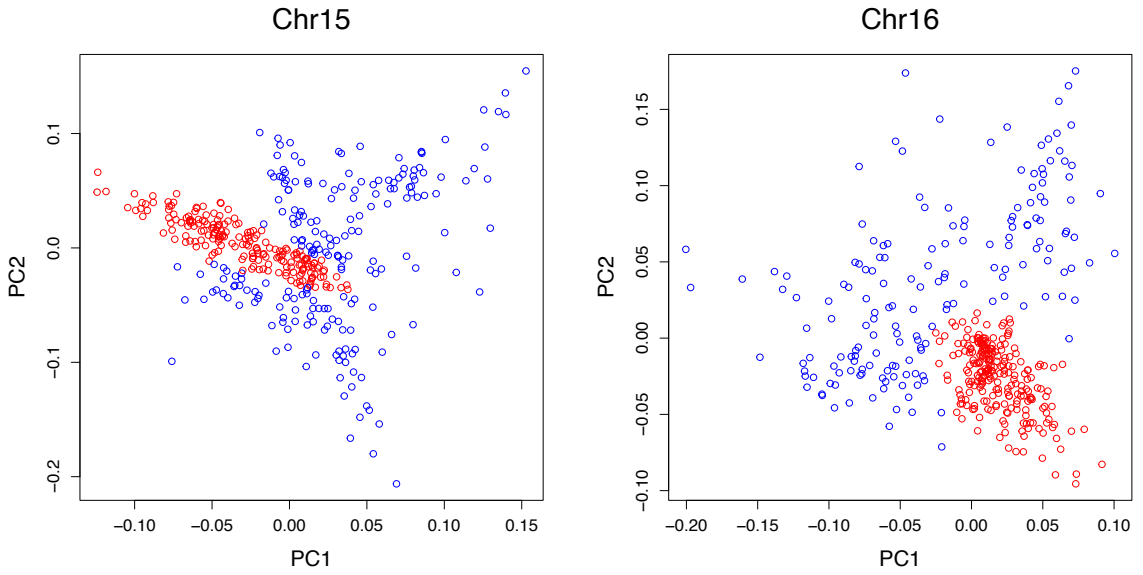


Figure B14, Cluster PCA plots. These plots show the distribution of individuals along PC1 and PC2 when the private variant PCA is conducted for chromosome 15 (left) and 16 (right). Individuals are coloured according to their cluster assigned by mclust.

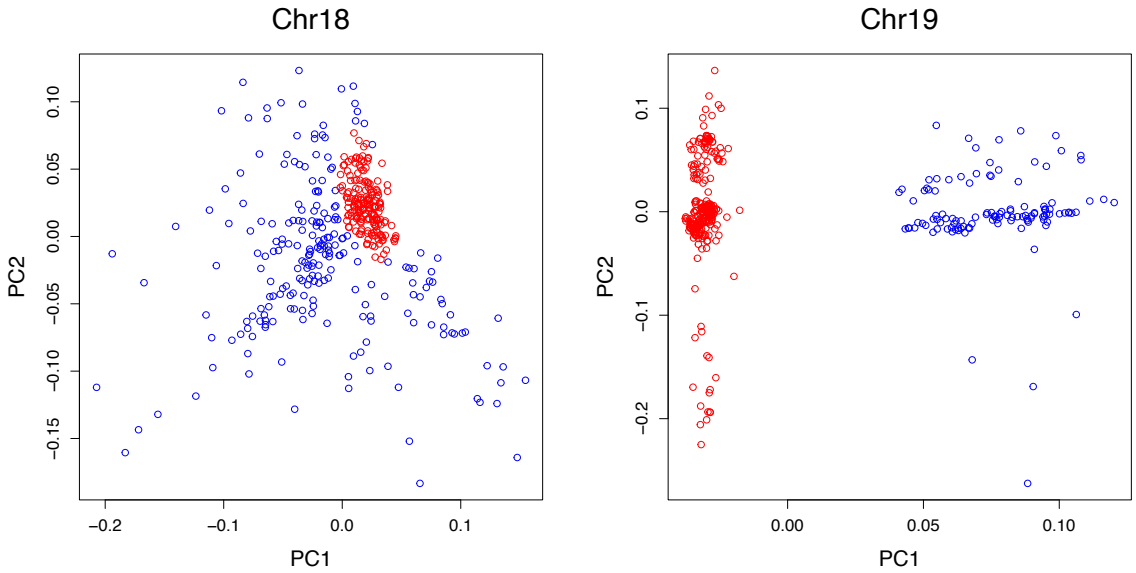


Figure B15, Cluster PCA plots. These plots show the distribution of individuals along PC1 and PC2 when the private variant PCA is conducted for chromosome 18 (left) and 19 (right). Individuals are coloured according to their cluster assigned by mclust.

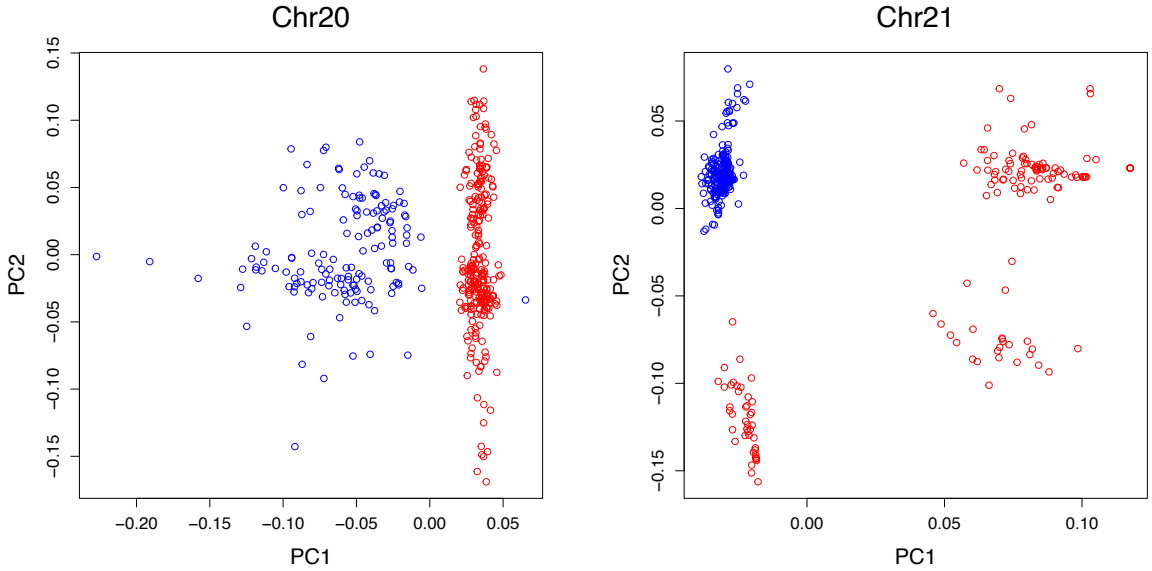


Figure B16, Cluster PCA plots. These plots show the distribution of individuals along PC1 and PC2 when the private variant PCA is conducted for chromosome 20 (left) and 21 (right). Individuals are coloured according to their cluster assigned by mclust.

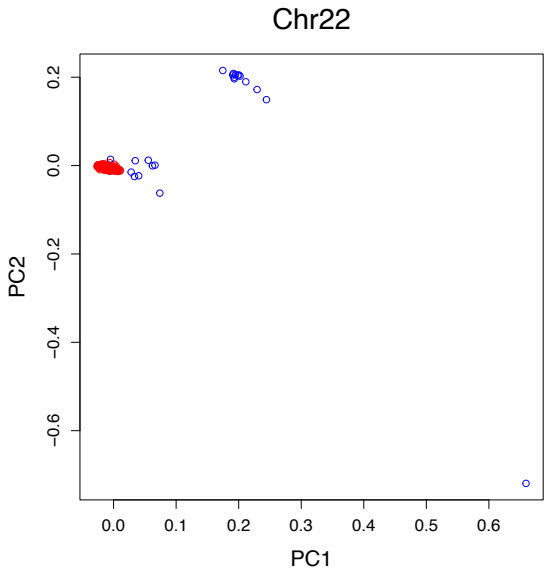


Figure B17, Cluster PCA plot. This plot shows the distribution of individuals along PC1 and PC2 when the private variant PCA is conducted for chromosome 22. Individuals are coloured according to their cluster assigned by mclust.

PC1 divergence chromosomal plots

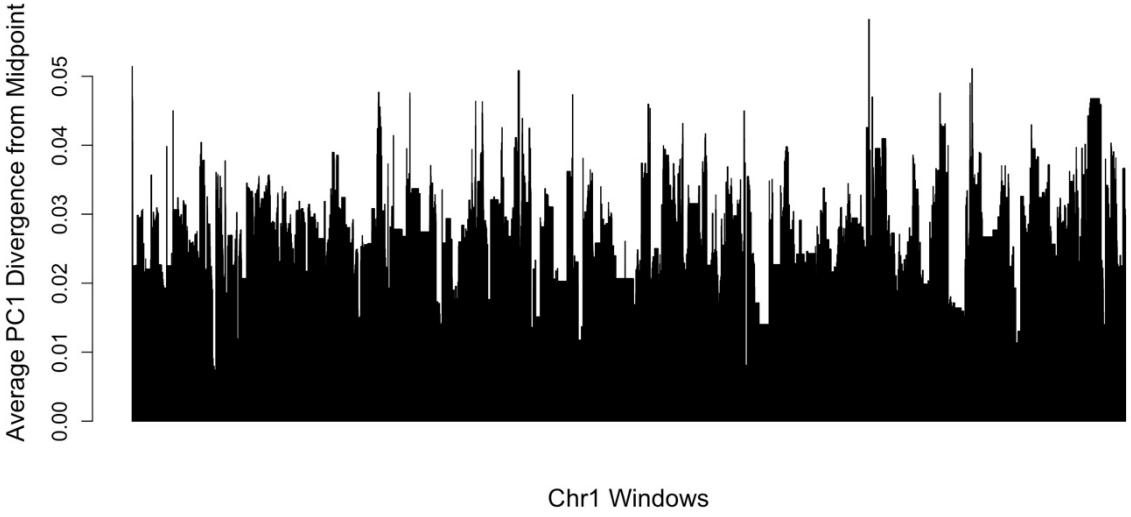


Figure B18, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 1 deviates from the central point between the two genomic clusters.

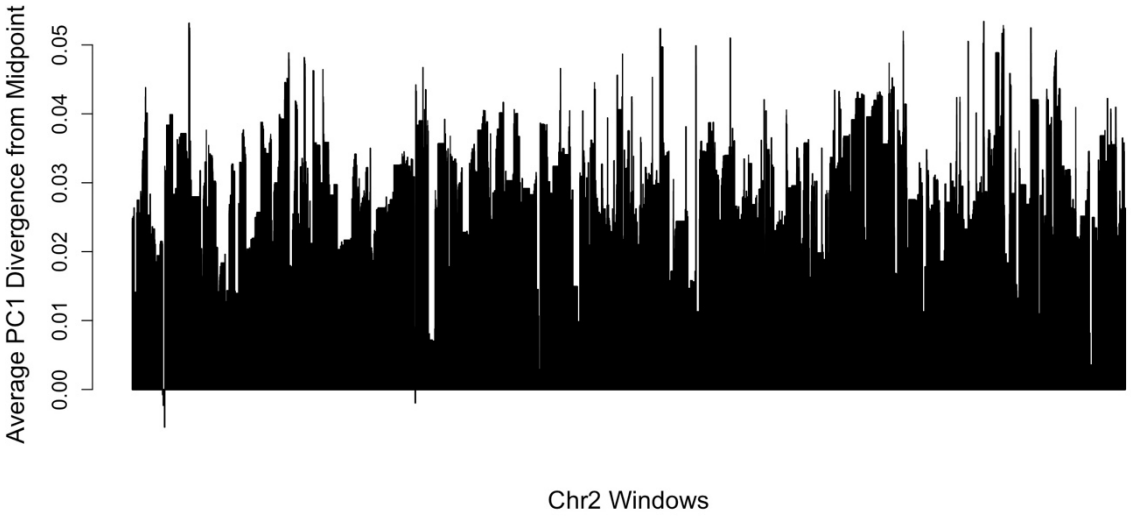


Figure B19, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 2 deviates from the central point between the two genomic clusters.

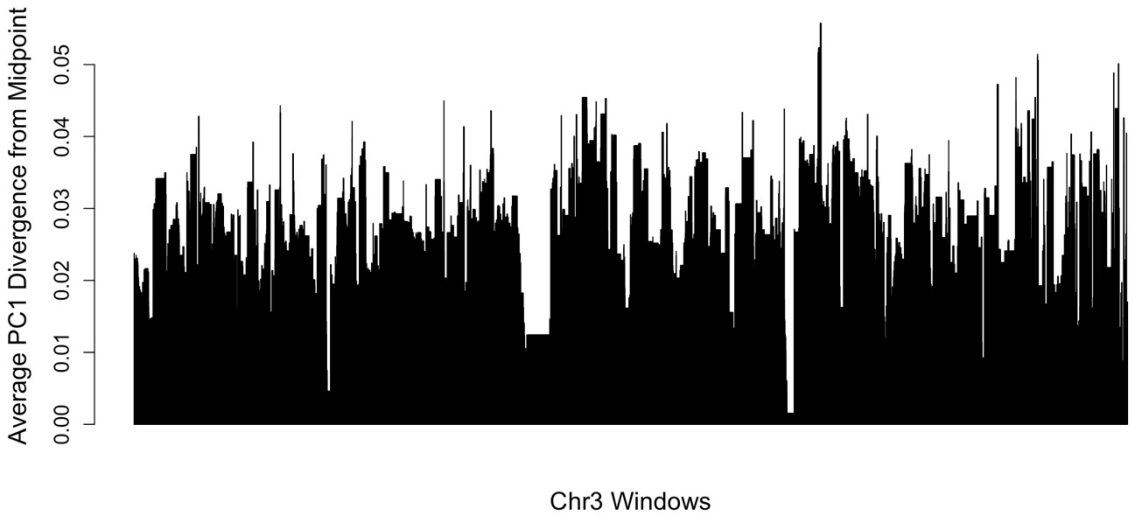


Figure B20, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 3 deviates from the central point between the two genomic clusters.

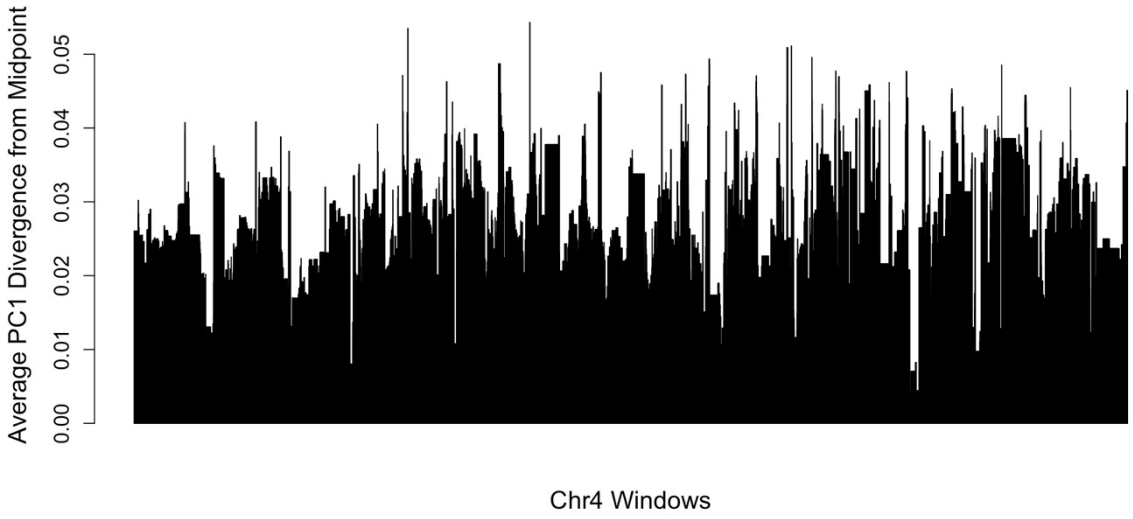


Figure B21, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 4 deviates from the central point between the two genomic clusters.

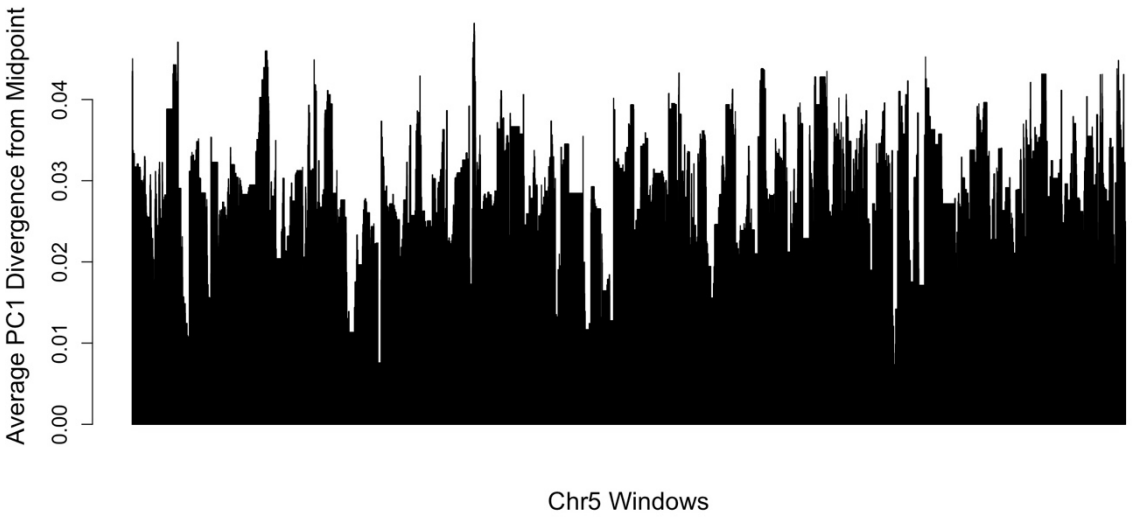


Figure B22, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 5 deviates from the central point between the two genomic clusters.

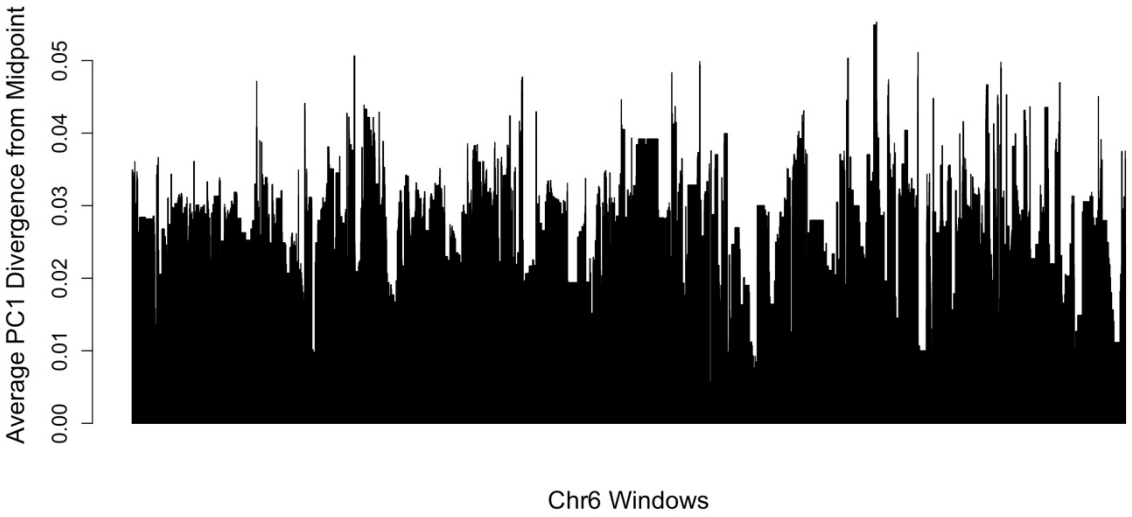


Figure B23, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 6 deviates from the central point between the two genomic clusters.

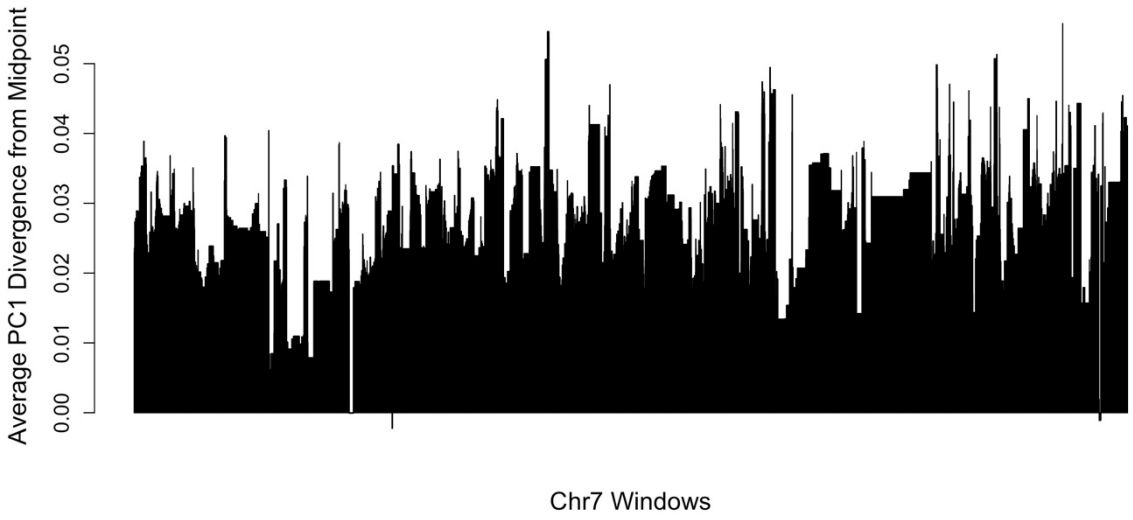


Figure B24, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 7 deviates from the central point between the two genomic clusters.

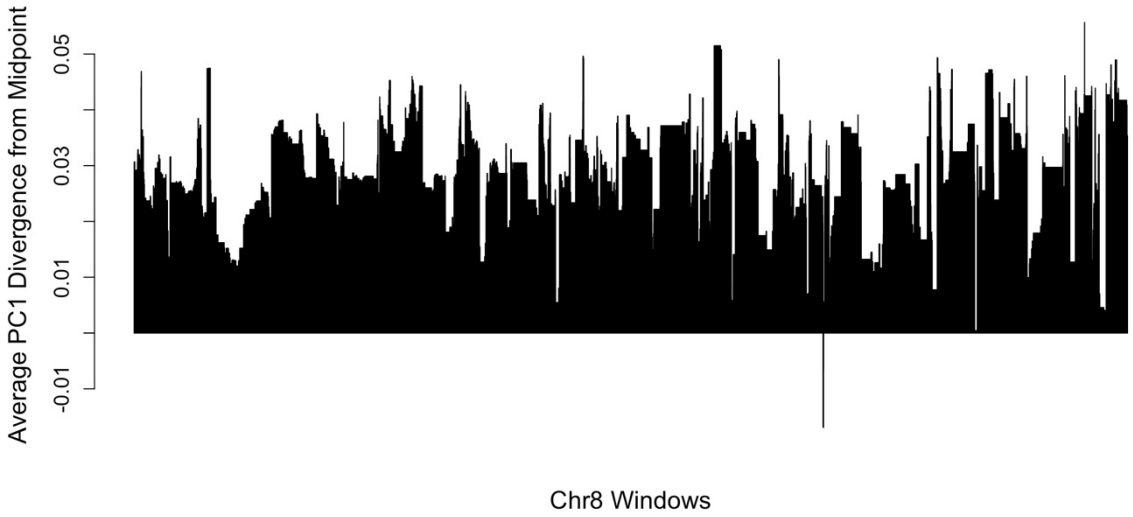


Figure B25, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 8 deviates from the central point between the two genomic clusters.

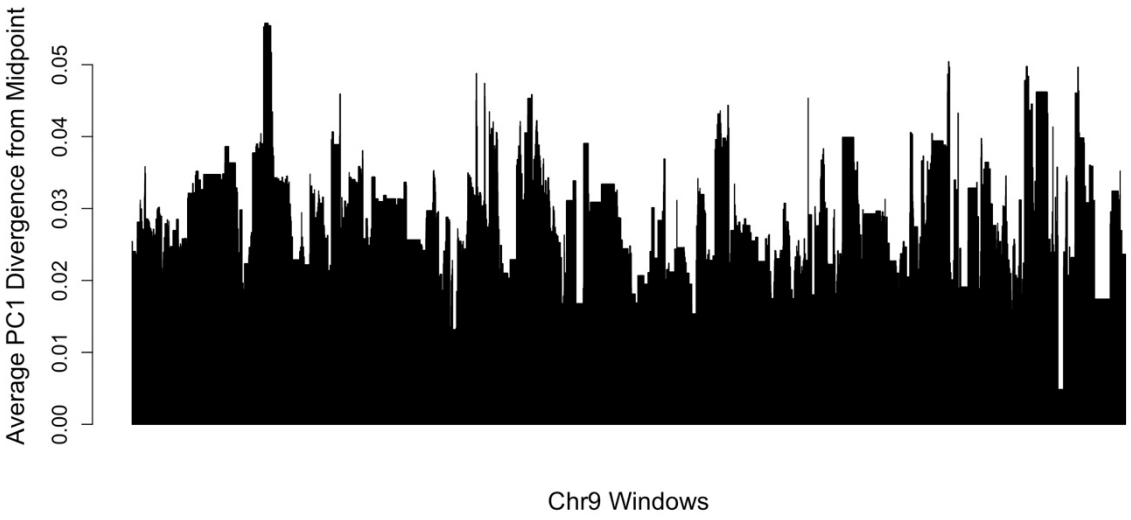


Figure B26, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 9 deviates from the central point between the two genomic clusters.

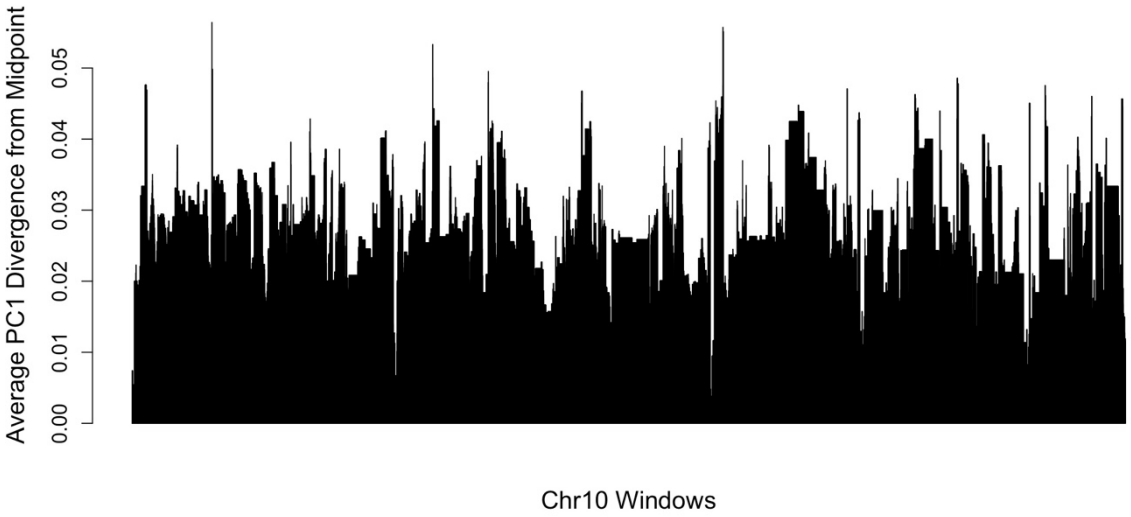


Figure B27, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 10 deviates from the central point between the two genomic clusters.

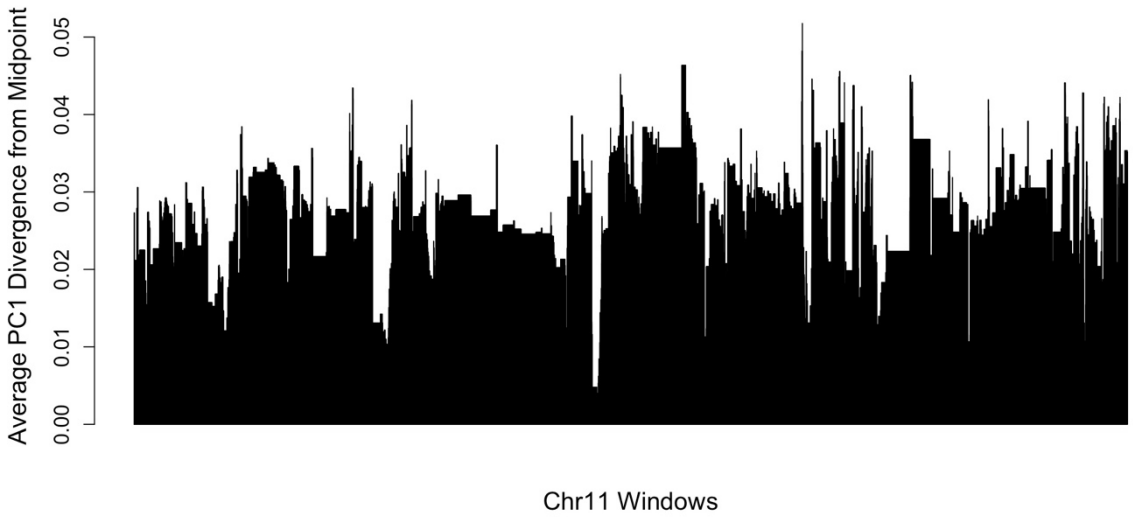


Figure B28, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 11 deviates from the central point between the two genomic clusters.

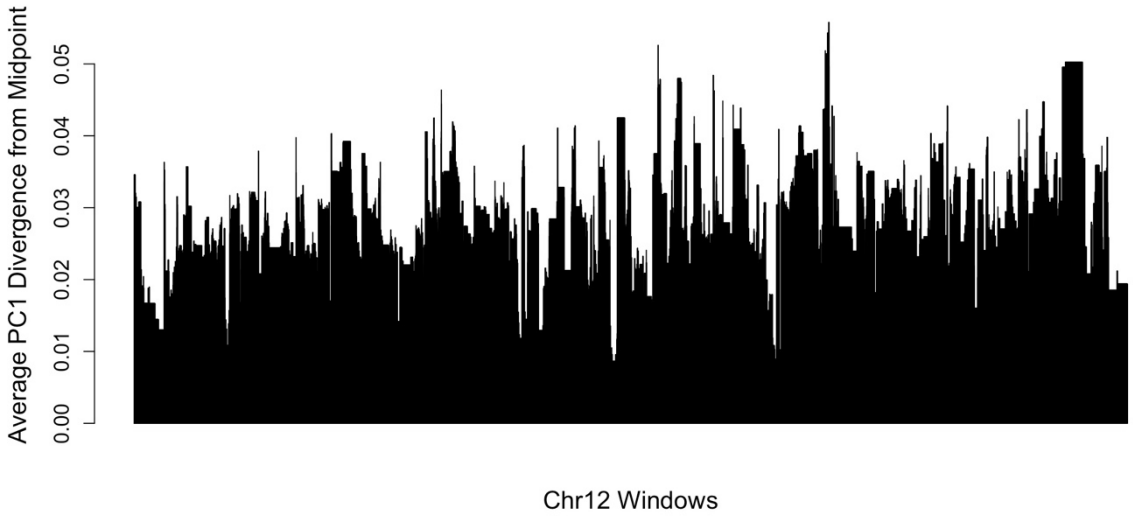


Figure B29, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 12 deviates from the central point between the two genomic clusters.

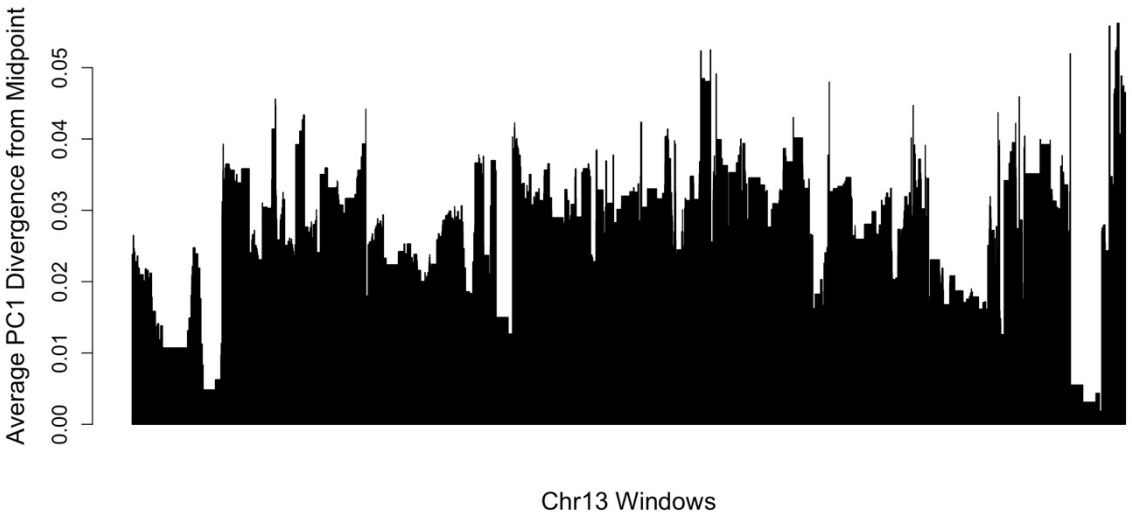


Figure B30, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 13 deviates from the central point between the two genomic clusters.

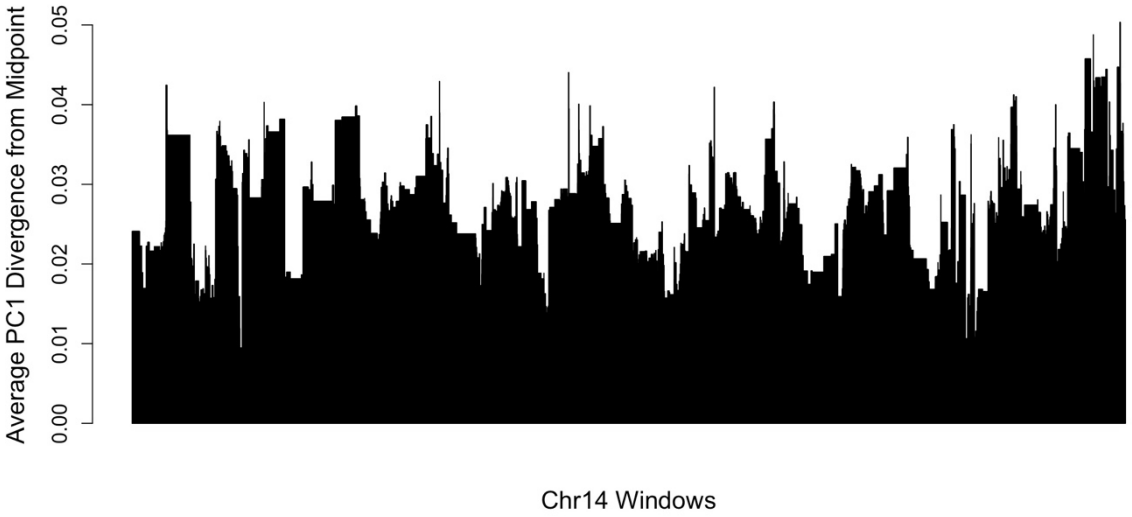


Figure B31, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 14 deviates from the central point between the two genomic clusters.

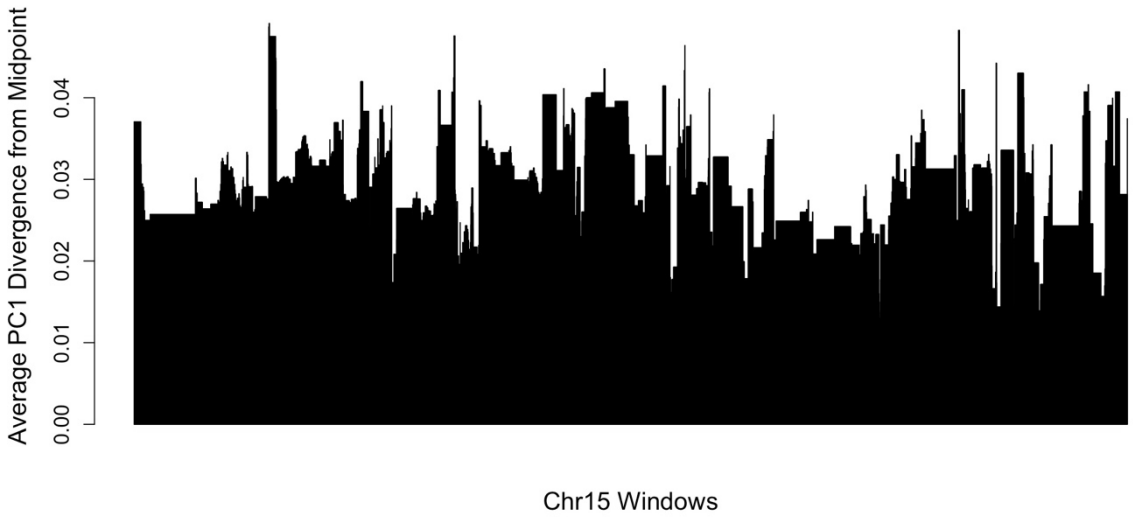


Figure B32, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 15 deviates from the central point between the two genomic clusters.

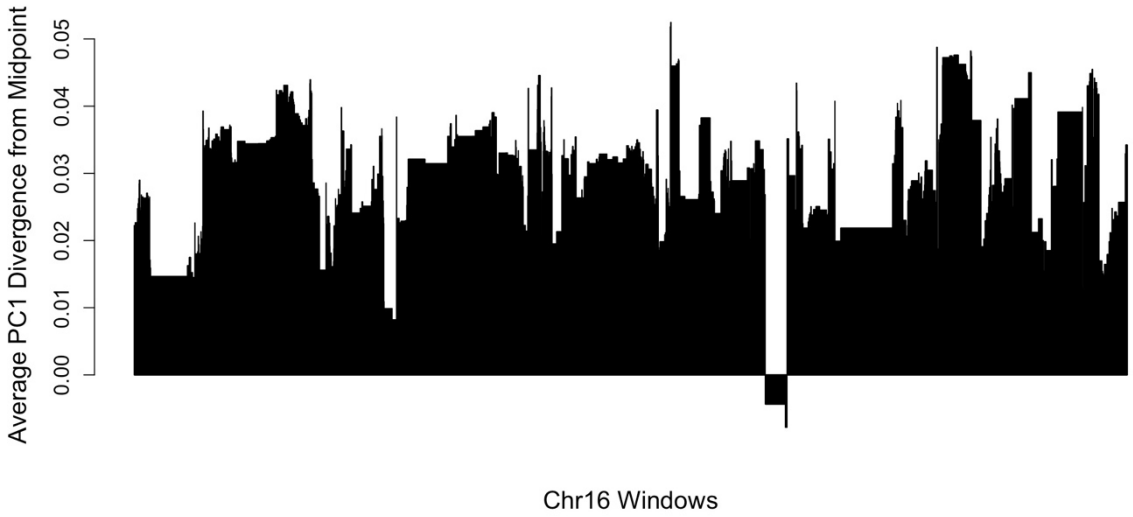


Figure B33, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 16 deviates from the central point between the two genomic clusters.

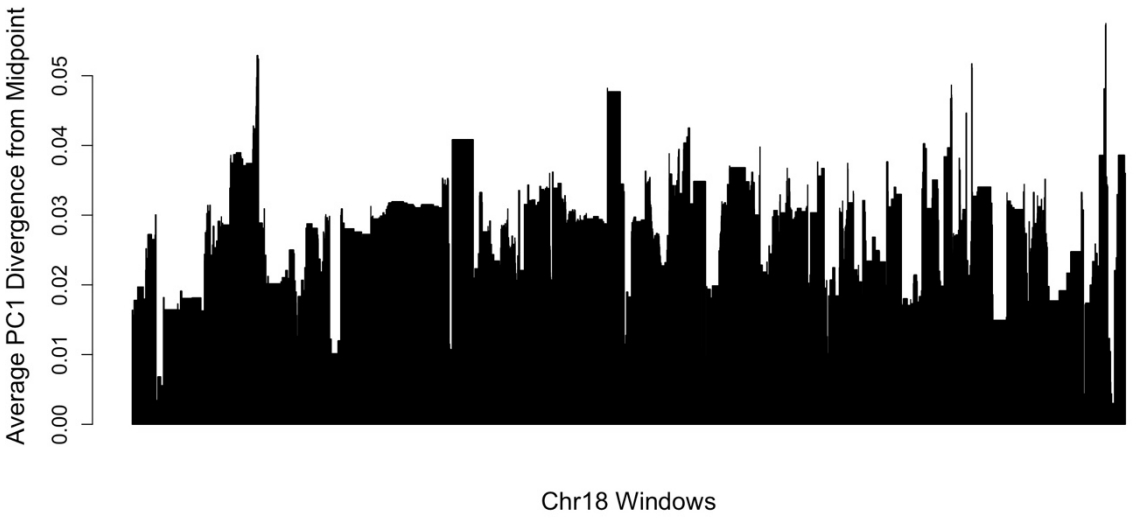


Figure B34, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 18 deviates from the central point between the two genomic clusters.

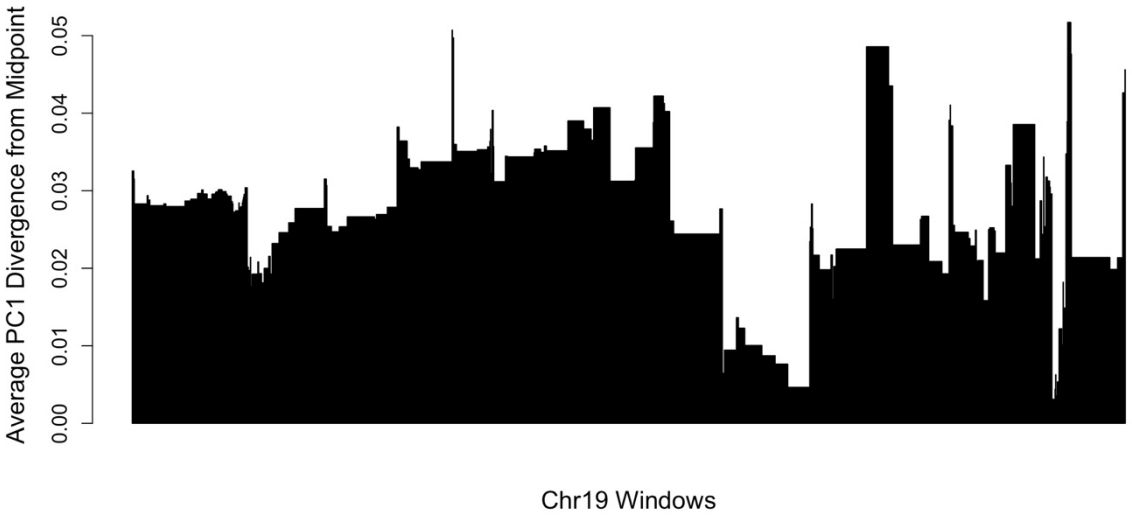


Figure B35, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 19 deviates from the central point between the two genomic clusters.

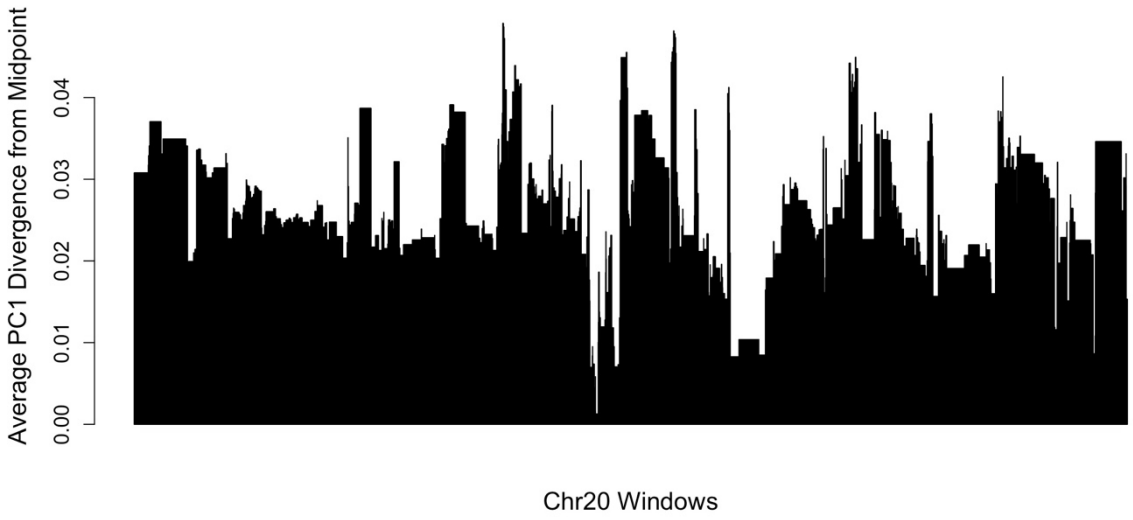


Figure B36, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 20 deviates from the central point between the two genomic clusters.

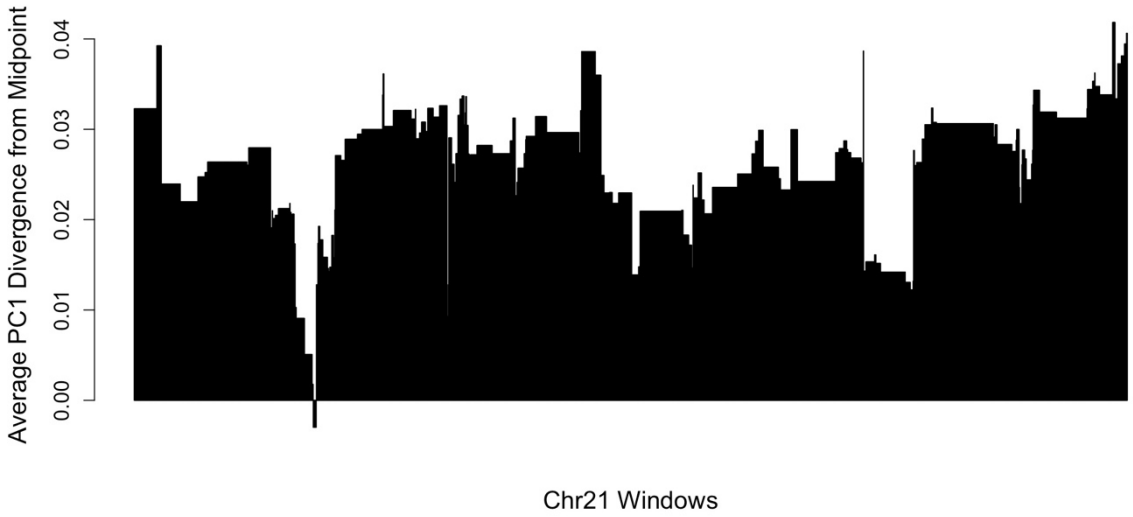


Figure B37, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 21 deviates from the central point between the two genomic clusters.

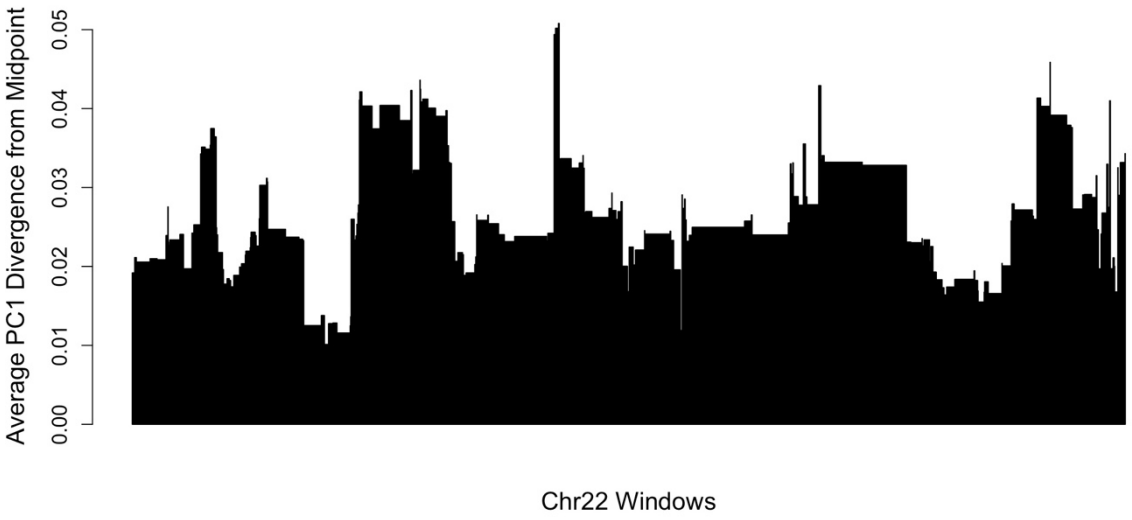


Figure B38, Divergence of average PC1 values from cluster midpoint. This plot tracks how the average PC1 value of all individuals with an archaic call in each window on chromosome 22 deviates from the central point between the two genomic clusters.

Cluster QC

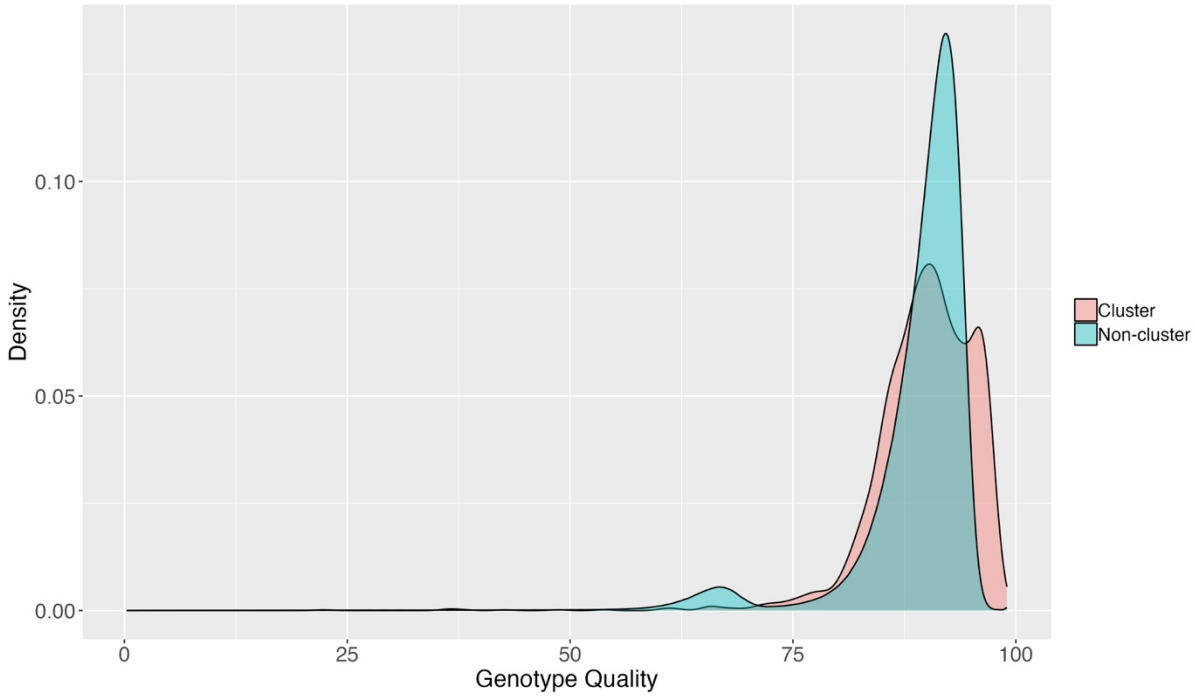


Figure B39, Comparison of GQ for cluster and non-cluster private variants. The average GQ for heterozygous genotypes is compared between private variants within the 683 kb region on chromosome 17 (Cluster) and those elsewhere in the genome (Non-cluster).

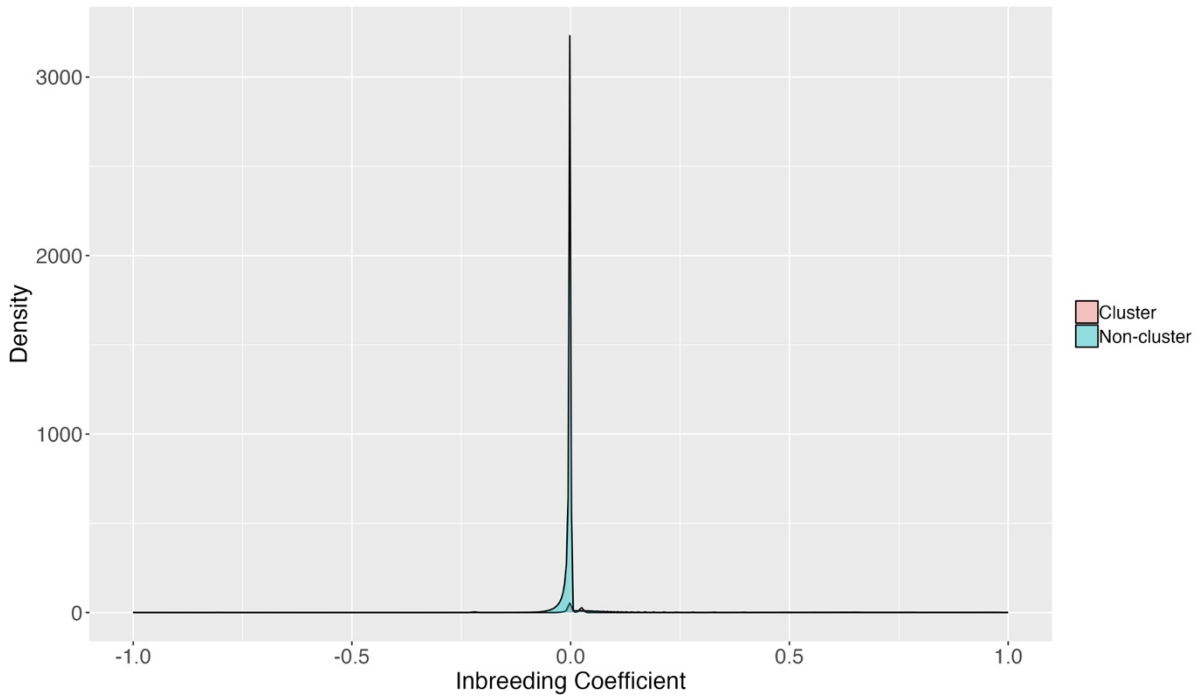


Figure B40, Comparison of inbreeding coefficient for cluster and non-cluster private variants. The inbreeding coefficient is compared between private variants within the 683 kb region on chromosome 17 (Cluster) and those elsewhere in the genome (Non-cluster).

Appendix C

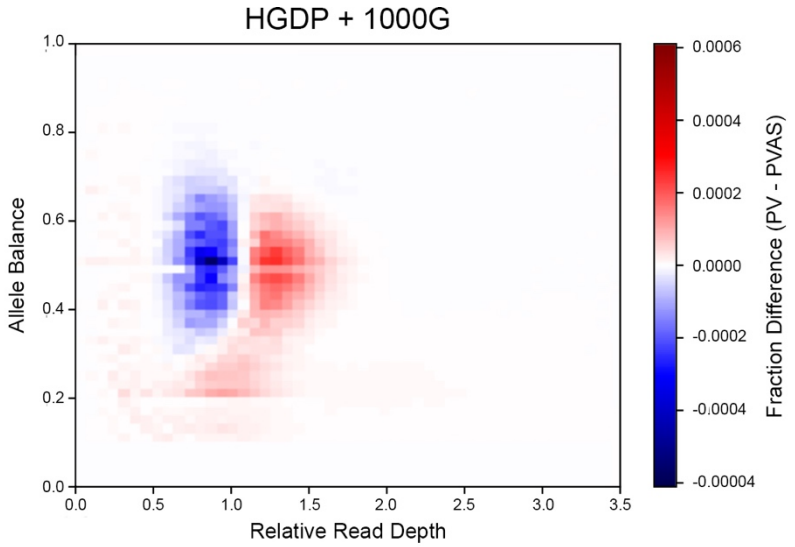


Figure C1, Duplication-screening heatmap of private variants in the HGDP+1000G dataset. Averaged across all individuals in the dataset, this heatmap shows the differences between PV and PVAS, where red represents an excess of PV and blue represents an excess of PVAS. Only bins with an AB between 0.1 and 0.9 were considered to focus on heterozygous genotypes.

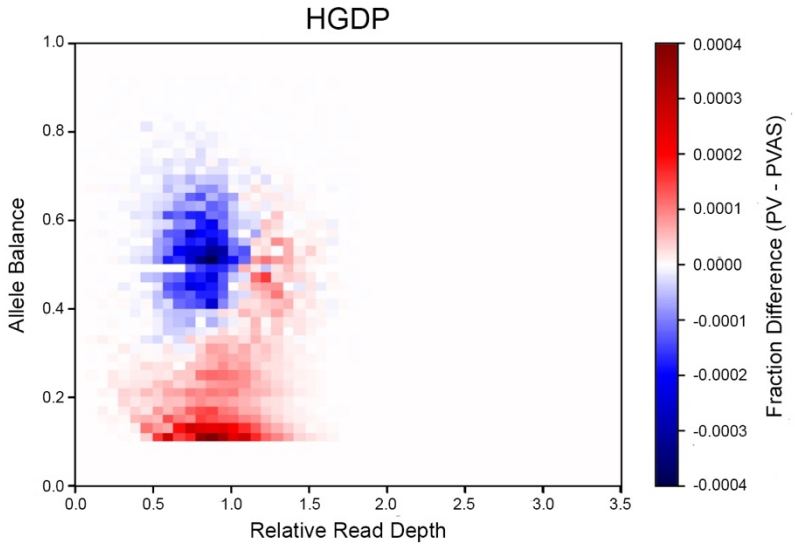


Figure C2, Duplication-screening heatmap of private variants in the HGDP dataset. Averaged across all individuals in the dataset, this heatmap shows the differences between PV and PVAS, where red represents an excess of PV and blue represents an excess of PVAS. Only bins with an AB between 0.1 and 0.9 were considered to focus on heterozygous genotypes.

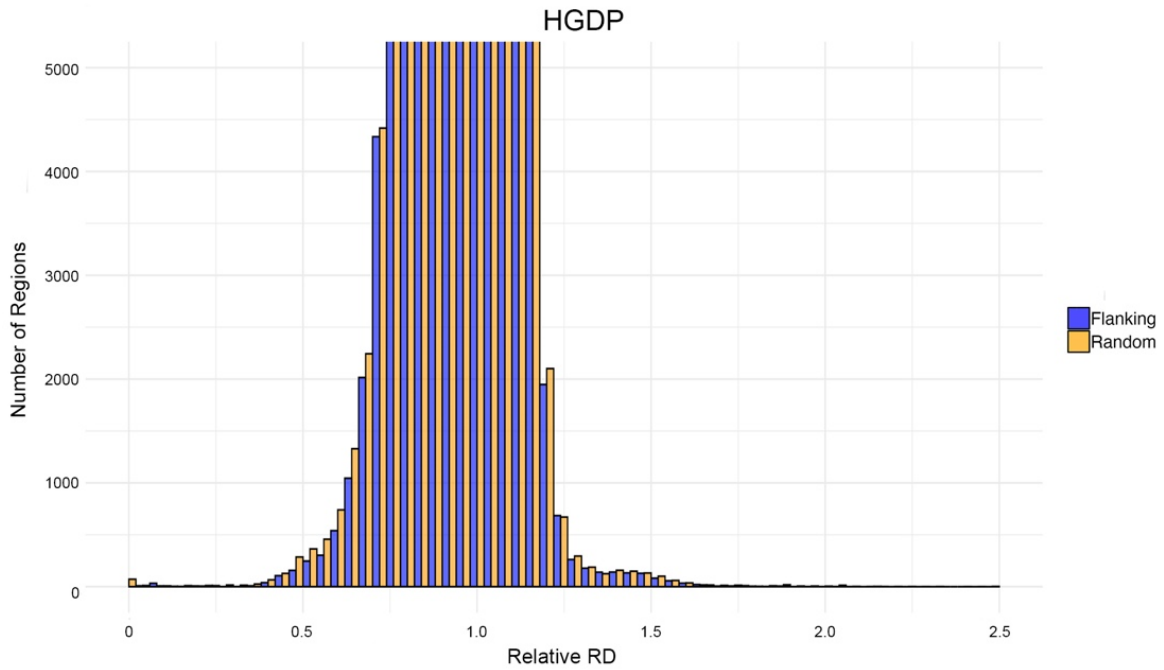


Figure C3, Deletion signatures flanking archaic segments. The distribution of relative RD in 10 kb regions flanking archaic segments is compared to matched random control regions across all HGDP individuals.

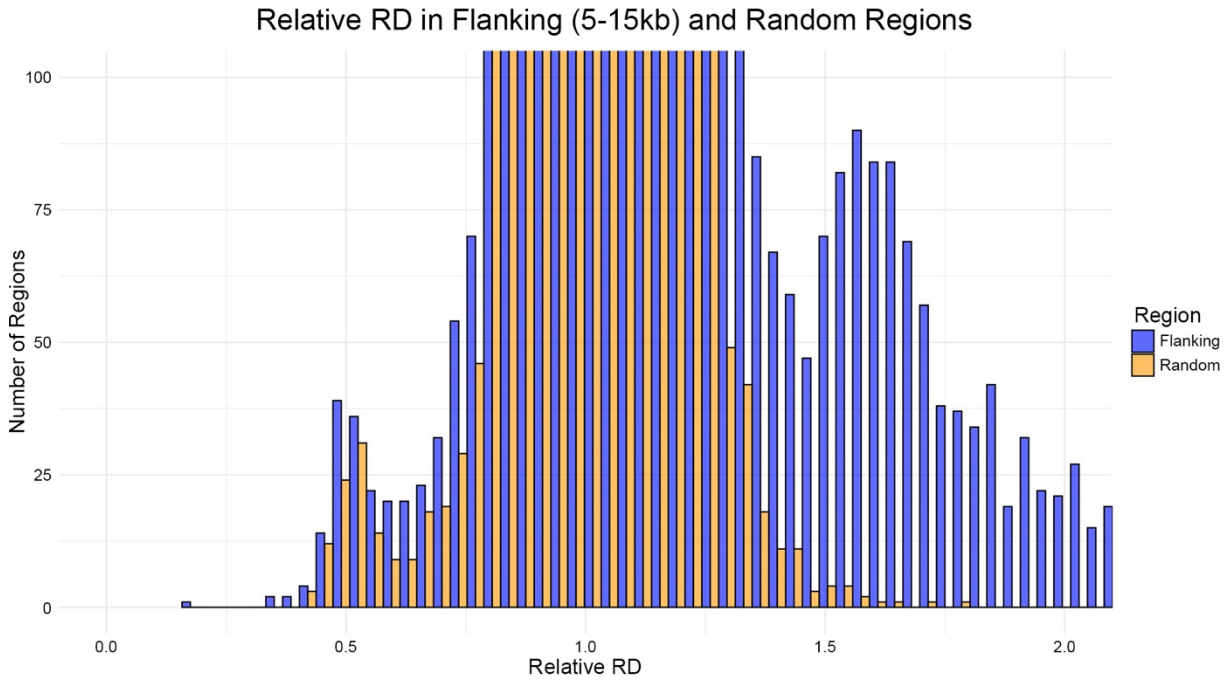


Figure C4, Deletion signatures flanking archaic segments. The distribution of relative RD in 10 kb regions flanking archaic segments, beginning 5 kb further away from the segment edge, is compared to matched random control regions across all parents from the 1000G Trio dataset.

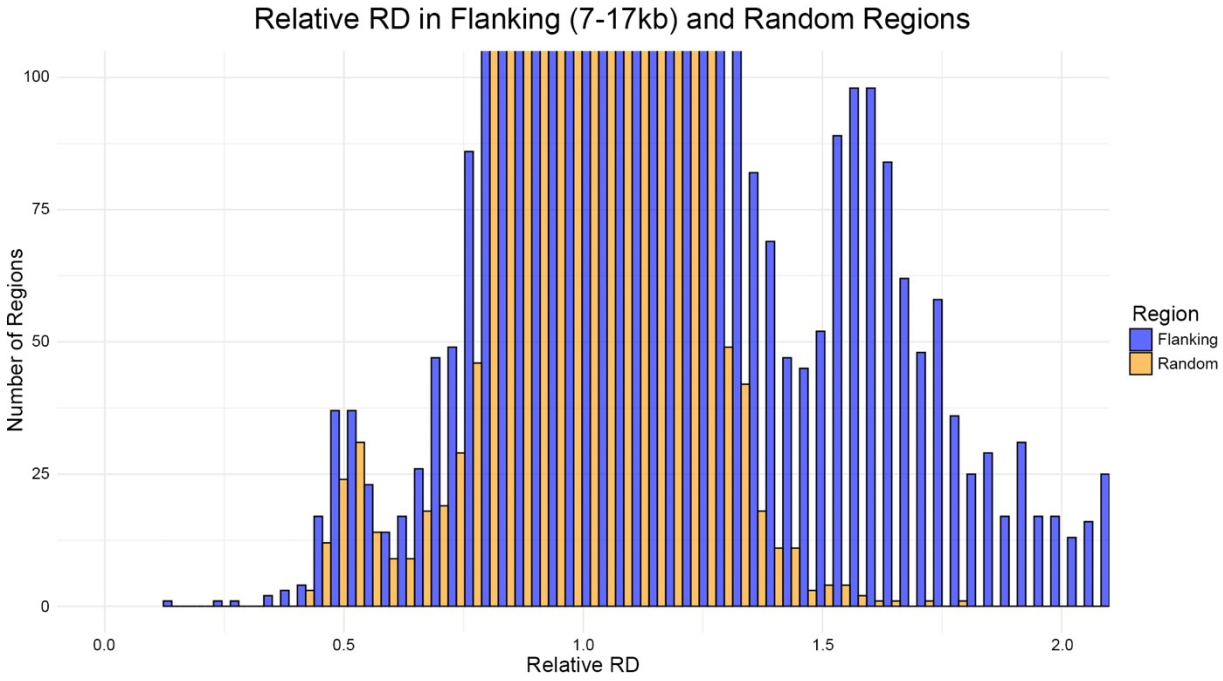


Figure C5, Deletion signatures flanking archaic segments. The distribution of relative RD in 10 kb regions flanking archaic segments, beginning 7 kb further away from the segment edge, is compared to matched random control regions across all parents from the 1000G Trio dataset.

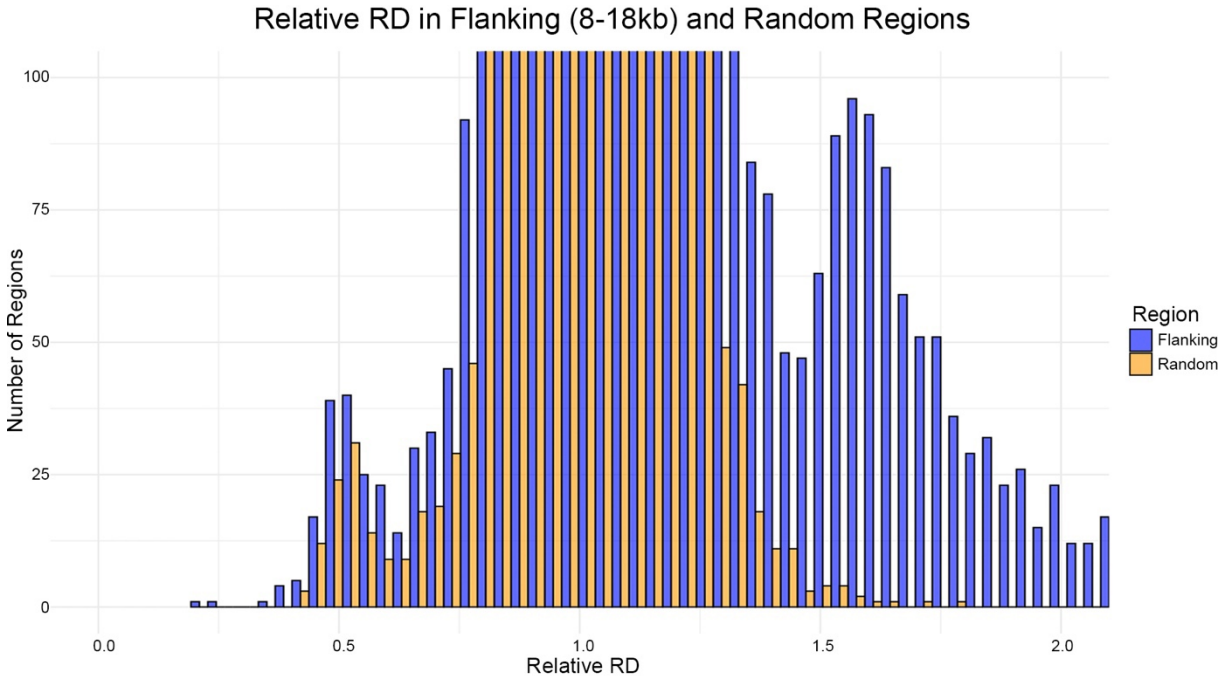


Figure C6, Deletion signatures flanking archaic segments. The distribution of relative RD in 10 kb regions flanking archaic segments, beginning 8 kb further away from the segment edge, is compared to matched random control regions across all parents from the 1000G Trio dataset.

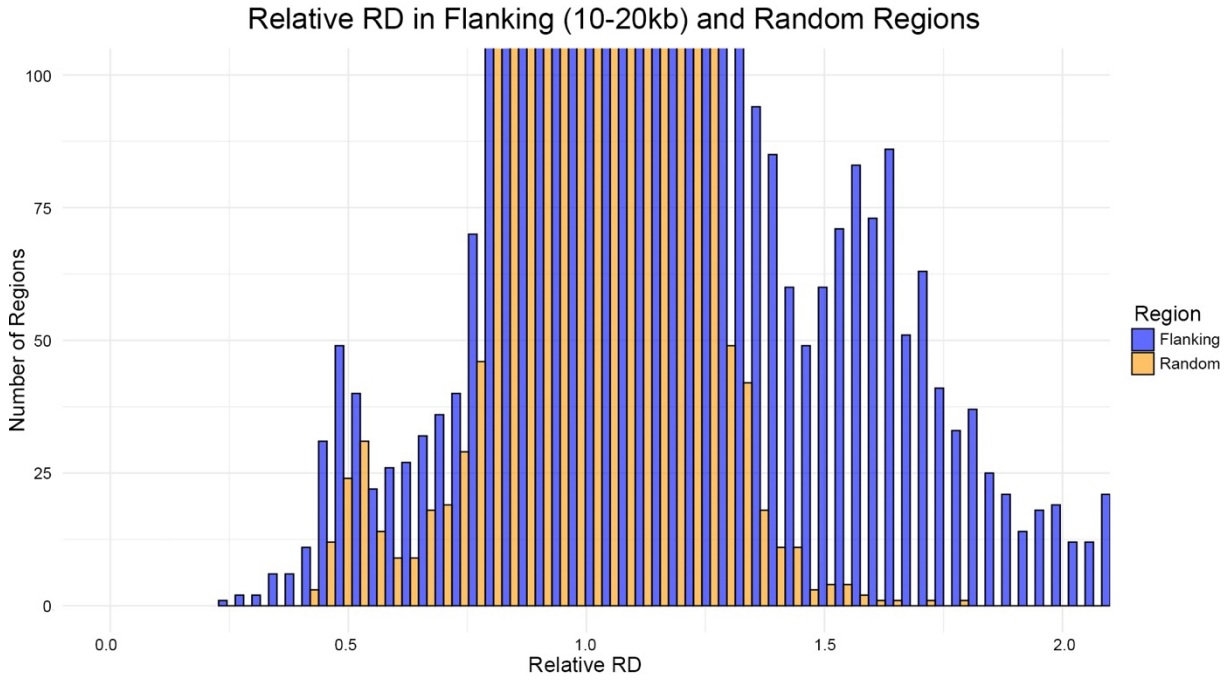


Figure C7, Deletion signatures flanking archaic segments. The distribution of relative RD in 10 kb regions flanking archaic segments, beginning 10 kb further away from the segment edge, is compared to matched random control regions across all parents from the 1000G Trio dataset.

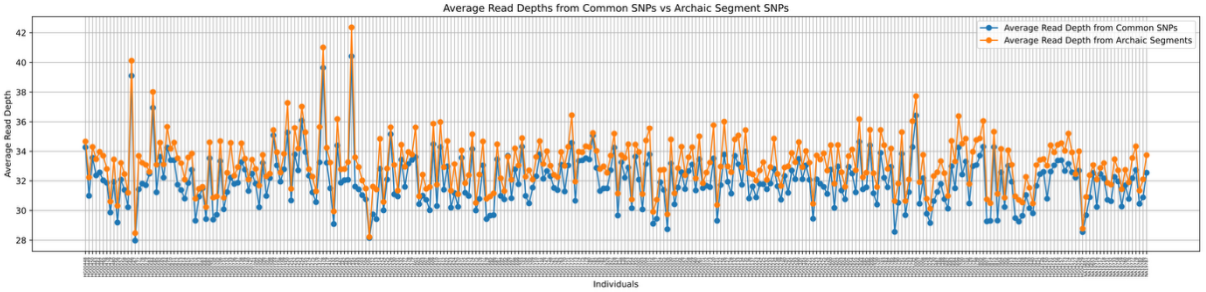


Figure C8, Average RD in archaic segments. This plot shows that the average RD of all variants in archaic segments is lower than the average RD of a list of common SNPs represented on the Infinium Global Screening Array (Illumina, Inc.) per individual.

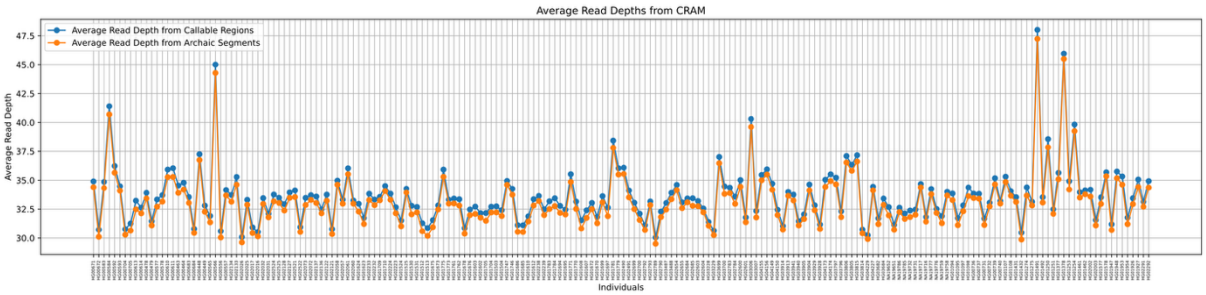


Figure C9, Average CRAM RD in archaic segments. This plot shows that the average CRAM RD of all archaic segments is slightly lower than that of the overall average CRAM depth per individual.

References

AbuAlia, K. F. N., Damm, E., Ullrich, K. K., Mukaj, A., Parvanov, E., Forejt, J., & Odenthal-Hesse, L. (2024). Natural variation in the zinc-finger-encoding exon of Prdm9 affects hybrid sterility phenotypes in mice. *Genetics*, *226*(3), iyae004. https://doi.org/10.1093/genetics/iyae004

Altemose, N., Noor, N., Bitoun, E., Tumian, A., Imbeault, M., Chapman, J. R., Aricescu, A. R., & Myers, S. R. (2017). A map of human PRDM9 binding provides evidence for novel behaviors of PRDM9 and other zinc-finger proteins in meiosis. *eLife*, *6*, e28383. https://doi.org/10.7554/eLife.28383

Anand, L., & Rodriguez Lopez, C. M. (2022). ChromoMap: An R package for interactive visualization of multi-omics data and annotation of chromosomes. *BMC Bioinformatics*, *23*(1). https://doi.org/10.1186/s12859-021-04556-z

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., … Schloss, J. A. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Auton, A., Fledel-Alon, A., Pfeifer, S., Venn, O., Ségurel, L., Street, T., Leffler, E. M., Bowden, R., Aneas, I., Broxholme, J., Humburg, P., Iqbal, Z., Lunter, G., Maller, J., Hernandez, R. D., Melton, C., Venkat, A., Nobrega, M. A., Bontrop, R., … McVean, G. (2012). A fine-scale chimpanzee genetic map from population sequencing. *Science (New York, N.Y.)*, *336*(6078), 193–198. https://doi.org/10.1126/science.1216872

Baker, C. L., Walker, M., Kajita, S., Petkov, P. M., & Paigen, K. (2014). PRDM9 binding organizes hotspot nucleosomes and limits Holliday junction migration. *Genome Research*, *24*(5), 724–732. https://doi.org/10.1101/gr.170167.113

Bateson, W. (1909). Heredity and variation in modern lights. In *Seward, A.C. Ed., Darwin and Modern Science* (pp. 85–101). Cambridge University Press.

Berg, I. L., Neumann, R., Lam, K.-W. G., Sarbajna, S., Odenthal-Hesse, L., May, C. A., & Jeffreys, A. J. (2010). PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nature Genetics*, *42*(10), 859–863. https://doi.org/10.1038/ng.658

Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., Blanché, H., Deleuze, J. F., Cann, H., Mallick, S., Reich, D., Sandhu, M. S., Skoglund, P., Scally, A., Xue, Y., … Tyler-Smith, C. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science*, *367*(6484). https://doi.org/10.1126/science.aay5012

Bhattacharyya, T., Gregorova, S., Mihola, O., Anger, M., Sebestova, J., Denny, P., Simecek, P., & Forejt, J. (2013). Mechanistic basis of infertility of mouse intersubspecific hybrids. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(6), E468-477. https://doi.org/10.1073/pnas.1219126110

Borde, V., & de Massy, B. (2013). Programmed induction of DNA double strand breaks during meiosis: Setting up communication between DNA and the chromosome structure. *Current Opinion in Genetics & Development, Genome Architecture and Expression*, *23*(2), 147–155. https://doi.org/10.1016/j.gde.2012.12.002

Boulton, A., Myers, R. S., & Redfield, R. J. (1997). The hotspot conversion paradox and the evolution of meiotic recombination. *Proceedings of the National Academy of Sciences*, *94*(15), 8058–8063. https://doi.org/10.1073/pnas.94.15.8058

Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R. D., & Petukhova, G. V. (2012). Genetic recombination is directed away from functional genomic elements in mice. *Nature*, *485*(7400), 642–645. https://doi.org/10.1038/nature11089

Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S., & Akey, J. M. (2018). Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell*, *173*(1). https://doi.org/10.1016/j.cell.2018.02.031

Bulik-Sullivan, B., Loh, P. R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K. H., Holmans, P. A., Lee, P., Collier, D. A., Huang, H., Pers, T. H., Agartz, I., Agerbo, E., … O'Donovan, M. C. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, *47*(3). https://doi.org/10.1038/ng.3211

Burssed, B., Zamariolli, M., Bellucco, F. T., & Melaragno, M. I. (2022). Mechanisms of structural chromosomal rearrangement formation. *Molecular Cytogenetics*, *15*, 23. https://doi.org/10.1186/s13039-022-00600-6

Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., Paul Flicek, Germer, S., Brand, H., Hall, I. M., Talkowski, M. E., … Zody, M. C. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, *185*(18), 3426-3440.e19. https://doi.org/10.1016/j.cell.2022.08.004

Carvalho, C. M. B., & Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews. Genetics*, *17*(4), 224–238. https://doi.org/10.1038/nrg.2015.25

Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E. J., Rodriguez, O. L., Guo, L., Collins, R. L., Fan, X., Wen, J., Handsaker, R. E., Fairley, S., Kronenberg, Z. N., Kong, X., Hormozdiari, F., Lee, D., Wenger, A. M., … Lee, C. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, *10*(1), 1784. https://doi.org/10.1038/s41467-018-08148-z

Chen, L., Wolf, A. B., Fu, W., Li, L., & Akey, J. M. (2020). Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals. *Cell*, *180*(4). https://doi.org/10.1016/j.cell.2020.01.012

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A. J., Kruglyak, S., & Saunders, C. T. (2016). Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics (Oxford, England)*, *32*(8), 1220–1222. https://doi.org/10.1093/bioinformatics/btv710

Coe, B. P., Stessman, H. A. F., Sulovari, A., Geisheker, M. R., Bakken, T. E., Lake, A. M., Dougherty, J. D., Lein, E. S., Hormozdiari, F., Bernier, R. A., & Eichler, E. E. (2019). Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nature Genetics*, *51*(1), 106–116. https://doi.org/10.1038/s41588-018-0288-4

Coll Macià, M., Skov, L., Peter, B. M., & Schierup, M. H. (2021). Different historical generation intervals in human populations inferred from Neanderthal fragment lengths and mutation signatures. *Nature Communications*, *12*(5317).

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C. H., Kristiansson, K., MacArthur, D. G., MacDonald, J. R., Onyiah, I., Pang, A. W. C., Robson, S., … Hurles, M. E. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, *464*(7289), 704–712. https://doi.org/10.1038/nature08516

Cooper, G. M., Coe, B. P., Girirajan, S., Rosenfeld, J. A., Vu, T. H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., Abdel-Hamid, H., Bader, P., McCracken, E., Niyazov, D., Leppig, K., Thiese, H., Hummel, M., Alexander, N., Gorski, J., … Eichler, E. E. (2011). A copy number variation morbidity map of developmental delay. *Nature Genetics*, *43*(9), 838–846. https://doi.org/10.1038/ng.909

Coyne, J. A. (1985). The genetic basis of Haldane's rule. *Nature*, *314*(6013), 736–738. https://doi.org/10.1038/314736a0

Dabney, J., Meyer, M., & Pääbo, S. (2013). Ancient DNA Damage. *Cold Spring Harbor Perspectives in Biology*, *5*(7), a012567. https://doi.org/10.1101/cshperspect.a012567

Dalal, V., Pasupuleti, N., Chaubey, G., Rai, N., & Shinde, V. (2023). Advancements and Challenges in Ancient DNA Research: Bridging the Global North–South Divide. *Genes*, *14*(2), 479. https://doi.org/10.3390/genes14020479

Dannemann, M., & Kelso, J. (2017). The Contribution of Neanderthals to Phenotypic Variation in Modern Humans. *American Journal of Human Genetics*, *101*(4). https://doi.org/10.1016/j.ajhg.2017.09.010

Dasmahapatra, K. K., Silva-Vásquez, A., Chung, J.-W., & Mallet, J. (2007). Genetic analysis of a wild-caught hybrid between non-sister Heliconius butterfly species. *Biology Letters*, *3*(6), 660–663. https://doi.org/10.1098/rsbl.2007.0401

Davies, B., Hatton, E., Altemose, N., Hussin, J. G., Pratto, F., Zhang, G., Hinch, A. G., Moralli, D., Biggs, D., Diaz, R., Preece, C., Li, R., Bitoun, E., Brick, K., Green, C. M., Camerini-Otero, R. D., Myers, S. R., & Donnelly, P. (2016). Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. *Nature*, *530*(7589), 171–176. https://doi.org/10.1038/nature16931

DeGiorgio, M., Jakobsson, M., & Rosenberg, N. A. (2009). Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proceedings of the National Academy of Sciences*, *106*(38), 16057–16062. https://doi.org/10.1073/pnas.0903341106

Dittwald, P., Gambin, T., Szafranski, P., Li, J., Amato, S., Divon, M. Y., Rodríguez Rojas, L. X., Elton, L. E., Scott, D. A., Schaaf, C. P., Torres-Martinez, W., Stevens, A. K., Rosenfeld, J. A., Agadi, S., Francis, D., Kang, S.-H. L., Breman, A., Lalani, S. R., Bacino, C. A., … Stankiewicz, P. (2013). NAHR-mediated copy-number variants in a clinical population: Mechanistic insights into both genomic disorders and Mendelizing traits. *Genome Research*, *23*(9), 1395–1409. https://doi.org/10.1101/gr.152454.112

Dobzhansky, Th. (1934). Studies on hybrid sterility. *Zeitschrift Für Zellforschung Und Mikroskopische Anatomie*, *21*(2), 169–223. https://doi.org/10.1007/BF00374056

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for Ancient Admixture between Closely Related Populations. *Molecular Biology and Evolution*, *28*(8), 2239–2252. https://doi.org/10.1093/molbev/msr048

Durvasula, A., & Sankararaman, S. (2019). A statistical model for reference-free inference of archaic local ancestry. *PLoS Genetics*, *15*(5). https://doi.org/10.1371/journal.pgen.1008175

Enard, D., & Petrov, D. A. (2018). Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. *Cell*, *175*(2), 360-371.e13. https://doi.org/10.1016/j.cell.2018.08.034

Eram, M. S., Bustos, S. P., Lima-Fernandes, E., Siarheyeva, A., Senisterra, G., Hajian, T., Chau, I., Duan, S., Wu, H., Dombrovski, L., Schapira, M., Arrowsmith, C. H., & Vedadi, M. (2014). Trimethylation of histone H3 lysine 36 by human methyltransferase PRDM9 protein. *The Journal of Biological Chemistry*, *289*(17), 12177–12188. https://doi.org/10.1074/jbc.M113.523183

Fang, L. L., Peede, D., Vecchyo, D. O.-D., McTavish, E. J., & Huerta-Sánchez, E. (2024). Leveraging shared ancestral variation to detect local introgression. *PLOS Genetics*, *20*(1), e1010155. https://doi.org/10.1371/journal.pgen.1010155

Fu, Q., Hajdinjak, M., Moldovan, O. T., Constantin, S., Mallick, S., Skoglund, P., Patterson, N., Rohland, N., Lazaridis, I., Nickel, B., Viola, B., Prüfer, K., Meyer, M., Kelso, J., Reich, D., & Pääbo, S. (2015). An early modern human from Romania with a recent Neanderthal ancestor. *Nature*, *524*(7564), 216–219. https://doi.org/10.1038/nature14558

Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S. M., Bondarev, A. A., Johnson, P. L. F., Petri, A. A., Prüfer, K., de Filippo, C., Meyer, M., Zwyns, N., Salazar-Garcia, D. C., Kuzmin, Y. V., Keates, S. G., Kosintsev, P. A., Razhev, D. I., Richards, M. P., Peristov, N. V., … Pääbo, S. (2014). The genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*, *514*(7523), 445–449. https://doi.org/10.1038/nature13810

Fu, Q., Mittnik, A., Johnson, P. L. F., Bos, K., Lari, M., Bollongino, R., Sun, C., Giemsch, L., Schmitz, R., Burger, J., Ronchitelli, A. M., Martini, F., Cremonesi, R. G., Svoboda, J., Bauer, P., Caramelli, D., Castellano, S., Reich, D., Pääbo, S., &

Krause, J. (2013). A revised timescale for human evolution based on ancient mitochondrial genomes. *Current Biology*, *23*(7). https://doi.org/10.1016/j.cub.2013.02.044

Gómez-Robles, A. (2019). Dental evolutionary rates and its implications for the Neanderthal–modern human divergence. *Science Advances*, *5*(5). https://doi.org/10.1126/sciadv.aaw1268

Gong, T., Hayes, V. M., & Chan, E. K. F. (2020). Detection of somatic structural variants from short-read next-generation sequencing data. *Briefings in Bioinformatics*, *22*(3), bbaa056. https://doi.org/10.1093/bib/bbaa056

Gower, G., Picazo, P. I., Fumagalli, M., & Racimo, F. (2021). Detecting adaptive introgression in human evolution using convolutional neural networks. *eLife*, *10*, e64669. https://doi.org/10.7554/eLife.64669

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: Scanning for occurrences of a given motif. *Bioinformatics (Oxford, England)*, *27*(7), 1017–1018. https://doi.org/10.1093/bioinformatics/btr064

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H. Y., Hansen, N. F., Durand, E. Y., Malaspinas, A. S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., … Pääbo, S. (2010). A draft sequence of the neandertal genome. *Science*, *328*(5979). https://doi.org/10.1126/science.1188021

Gregorova, S., Gergelits, V., Chvatalova, I., Bhattacharyya, T., Valiskova, B., Fotopulosova, V., Jansa, P., Wiatrowska, D., & Forejt, J. (2018). Modulation of Prdm9-controlled meiotic chromosome asynapsis overrides hybrid sterility in mice. *eLife*, *7*, e34282. https://doi.org/10.7554/eLife.34282

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., & Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics*, *43*(10), 1031–1034. https://doi.org/10.1038/ng.937

Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B. V., Hjartarson, E., Sigurdsson, G. T., Stacey, S. N., Frigge, M. L., Holm, H., Saemundsdottir, J., Helgadottir, H. T., Johannsdottir, H., Sigfusson, G., Thorgeirsson, G., … Stefansson, K. (2015). Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics*, *47*(5), 435–444. https://doi.org/10.1038/ng.3247

Halldorsson, B. V., Palsson, G., Stefansson, O. A., Jonsson, H., Hardarson, M. T., Eggertsson, H. P., Gunnarsson, B., Oddsson, A., Halldorsson, G. H., Zink, F., Gudjonsson, S. A., Frigge, M. L., Thorleifsson, G., Sigurdsson, A., Stacey, S. N., Sulem, P., Masson, G., Helgason, A., Gudbjartsson, D. F., … Stefansson, K. (2019). Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science*, *363*(6425), eaau1043. https://doi.org/10.1126/science.aau1043

Harris, K., & Nielsen, R. (2016). The genetic cost of neanderthal introgression. *Genetics*, *203*(2), 881–891. https://doi.org/10.1534/genetics.116.186890

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., Hillman-Jackson, J., Kuhn, R. M., Pedersen, J. S., Pohl, A., Raney, B. J., Rosenbloom, K. R., Siepel, A., Smith, K. E., Sugnet, C. W., … Kent, W. J. (2006). The UCSC Genome Browser Database: Update 2006. *Nucleic Acids Research*, *34*(Database issue), D590-598. https://doi.org/10.1093/nar/gkj144

Ho, S. S., Urban, A. E., & Mills, R. E. (2020). Structural variation in the sequencing era. *Nature Reviews Genetics*, *21*(3), 171–189. https://doi.org/10.1038/s41576-019-0180-9

Huang, X., Hackl, J., & Kuhlwilm, M. (2025). Decoding genomic landscapes of introgression. *Trends in Genetics*. https://doi.org/10.1016/j.tig.2025.07.001

Hublin, J.-J., Ben-Ncer, A., Bailey, S. E., Freidline, S. E., Neubauer, S., Skinner, M. M., Bergmann, I., Le Cabec, A., Benazzi, S., Harvati, K., & Gunz, P. (2017). New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature*, *546*(7657), 289–292. https://doi.org/10.1038/nature22336

Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., Ni, P., Wang, B., Ou, X., Huasang, Luosang, J., Cuo, Z. X. P., Li, K., Gao, G., Yin, Y., … Nielsen, R. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, *512*(7513), 194–197. https://doi.org/10.1038/nature13408

Hvala, J. A., Frayer, M. E., & Payseur, B. A. (2018). Signatures of hybridization and speciation in genomic patterns of ancestry. *Evolution*, *72*(8), 1540–1552. https://doi.org/10.1111/evo.13509

Iasi, L. N. M., Chintalapati, M., Skov, L., Mesa, A. B., Hajdinjak, M., Peter, B. M., & Moorjani, P. (2024). Neanderthal ancestry through time: Insights from genomes of ancient and present-day humans. *Science*, *386*(6727), eadq3010. https://doi.org/10.1126/science.adq3010

International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320. https://doi.org/10.1038/nature04226

Itsara, A., Wu, H., Smith, J. D., Nickerson, D. A., Romieu, I., London, S. J., & Eichler, E. E. (2010). De novo rates and selection of large copy number variation. *Genome Research*, *20*(11), 1469–1481. https://doi.org/10.1101/gr.107680.110

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., … Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, *36*(4), 338–345. https://doi.org/10.1038/nbt.4060

Jégou, B., Sankararaman, S., Rolland, A. D., Reich, D., & Chalmel, F. (2017). Meiotic Genes Are Enriched in Regions of Reduced Archaic Ancestry. *Molecular Biology and Evolution*, *34*(8), 1974–1980. https://doi.org/10.1093/molbev/msx141

Jeste, S. S., & Geschwind, D. H. (2014). Disentangling the heterogeneity of autism spectrum disorder through genetic findings. *Nature Reviews. Neurology*, *10*(2), 74–81. https://doi.org/10.1038/nrneurol.2013.278

Juric, I., Aeschbacher, S., & Coop, G. (2016a). The Strength of Selection against Neanderthal Introgression. *PLOS Genetics*, *12*(11), e1006340. https://doi.org/10.1371/journal.pgen.1006340

Juric, I., Aeschbacher, S., & Coop, G. (2016b). The Strength of Selection against Neanderthal Introgression. *PLoS Genetics*, *12*(11). https://doi.org/10.1371/journal.pgen.1006340

Koenig, Z., Yohannes, M. T., Nkambule, L. L., Zhao, X., Goodrich, J. K., Kim, H. A., Wilson, M. W., Tiao, G., Hao, S. P., Sahakian, N., Chao, K. R., Walker, M. A., Lyu, Y., Rehm, H. L., Neale, B. M., Talkowski, M. E., Daly, M. J., Brand, H., Karczewski, K. J., … Martin, A. R. (2024). A harmonized public resource of deeply

sequenced diverse human genomes. *bioRxiv*, 2023.01.23.525248. https://doi.org/10.1101/2023.01.23.525248

Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G. B., Jonasdottir, A., Gylfason, A., Kristinsson, K. T., Gudjonsson, S. A., Frigge, M. L., Helgason, A., Thorsteinsdottir, U., & Stefansson, K. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, *467*(7319), 1099–1103. https://doi.org/10.1038/nature09525

Kurki, M. I., Karjalainen, J., Palta, P., Sipilä, T. P., Kristiansson, K., Donner, K. M., Reeve, M. P., Laivuori, H., Aavikko, M., Kaunisto, M. A., Loukola, A., Lahtela, E., Mattsson, H., Laiho, P., Della Briotta Parolo, P., Lehisto, A. A., Kanai, M., Mars, N., Rämö, J., … Palotie, A. (2023). FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*, *613*(7944), 508–518. https://doi.org/10.1038/s41586-022-05473-8

Lam, I., & Keeney, S. (2015a). Mechanism and Regulation of Meiotic Recombination Initiation. *Cold Spring Harbor Perspectives in Biology*, *7*(1), a016634. https://doi.org/10.1101/cshperspect.a016634

Lam, I., & Keeney, S. (2015b). Non-paradoxical evolutionary stability of the recombination initiation landscape in yeast. *Science (New York, N.Y.)*, *350*(6263), 932–937. https://doi.org/10.1126/science.aad0814

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., … The Wellcome Trust: (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Lange, J., Yamada, S., Tischfield, S. E., Pan, J., Kim, S., Zhu, X., Socci, N. D., Jasin, M., & Keeney, S. (2016). The landscape of mouse meiotic double-strand break formation, processing and repair. *Cell*, *167*(3), 695-708.e16. https://doi.org/10.1016/j.cell.2016.09.035

Larson, E. L., Vanderpool, D., Sarver, B. A. J., Callahan, C., Keeble, S., Provencio, L. L., Kessler, M. D., Stewart, V., Nordquist, E., Dean, M. D., & Good, J. M. (2018). The Evolution of Polymorphic Hybrid Incompatibilities in House Mice. *Genetics*, *209*(3), 845–859. https://doi.org/10.1534/genetics.118.300840

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, *15*(6), R84. https://doi.org/10.1186/gb-2014-15-6-r84

Lee, J. A., Carvalho, C. M. B., & Lupski, J. R. (2007). A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. *Cell*, *131*(7), 1235–1247. https://doi.org/10.1016/j.cell.2007.11.037

Lesecque, Y., Glémin, S., Lartillot, N., Mouchiroud, D., & Duret, L. (2014). The Red Queen Model of Recombination Hotspots Evolution in the Light of Archaic and Modern Human Genomes. *PLOS Genetics*, *10*(11), e1004790. https://doi.org/10.1371/journal.pgen.1004790

Li, H., & Durbin, R. (2011). Inference of Human Population History From Whole Genome Sequence of A Single Individual. *Nature*, *475*(7357), 493–496. https://doi.org/10.1038/nature10231

Lieber, M. R. (2010). The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End Joining Pathway. *Annual Review of Biochemistry*, *79*, 181–211. https://doi.org/10.1146/annurev.biochem.052308.093131

Lifschytz, E., & Lindsley, D. L. (1972). The role of X-chromosome inactivation during spermatogenesis (Drosophila-allocycly-chromosome evolution-male sterility-dosage compensation). *Proceedings of the National Academy of Sciences of the United States of America*, *69*(1), 182–186. https://doi.org/10.1073/pnas.69.1.182

Liu, P., Carvalho, C. M. B., Hastings, P. J., & Lupski, J. R. (2012). Mechanisms for recurrent and complex human genomic rearrangements. *Current Opinion in Genetics & Development*, *22*(3), 211–220. https://doi.org/10.1016/j.gde.2012.02.012

Lupski, J. R., Reid, J. G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D. C. Y., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D. A., McGuire, A. L., Zhang, F., Stankiewicz, P., Halperin, J. J., Yang, C., Gehman, C., Guo, D., Irikat, R. K., Tom, W., … Gibbs, R. A. (2010). Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *The New England Journal of Medicine*, *362*(13), 1181–1191. https://doi.org/10.1056/NEJMoa0908094

Maheshwari, S., & Barbash, D. A. (2011). The genetics of hybrid incompatibilities. *Annual Review of Genetics*, *45*, 331–355. https://doi.org/10.1146/annurev-genet-110410-132514

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends in Ecology & Evolution, Special Issue: Invasions, Guest Edited by Michael E. Hochberg and Nicholas J. Gotelli*, *20*(5), 229–237. https://doi.org/10.1016/j.tree.2005.02.010

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., … Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from

142 diverse populations. *Nature*, *538*(7624), 201–206. https://doi.org/10.1038/nature18964

Martin, S. H., & Jiggins, C. D. (2017). Interpreting the genomic landscape of introgression. *Current Opinion in Genetics & Development, Evolutionary Genetics*, *47*, 69–74. https://doi.org/10.1016/j.gde.2017.08.007

McDonald-McGinn, D. M., Sullivan, K. E., Marino, B., Philip, N., Swillen, A., Vorstman, J. A. S., Zackai, E. H., Emanuel, B. S., Vermeesch, J. R., Morrow, B. E., Scambler, P. J., & Bassett, A. S. (2015). 22q11.2 deletion syndrome. *Nature Reviews Disease Primers*, *1*(1), 15071. https://doi.org/10.1038/nrdp.2015.71

McDougall, I., Brown, F. H., & Fleagle, J. G. (2005). Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature*, *433*(7027), 733–736. https://doi.org/10.1038/nature03258

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

McVey, M., & Lee, S. E. (2008). MMEJ repair of double-strand breaks (director's cut): Deleted sequences and alternative endings. *Trends in Genetics : TIG*, *24*(11), 529–538. https://doi.org/10.1016/j.tig.2008.08.007

Meyer, M., Arsuaga, J.-L., de Filippo, C., Nagel, S., Aximu-Petri, A., Nickel, B., Martínez, I., Gracia, A., de Castro, J. M. B., Carbonell, E., Viola, B., Kelso, J., Prüfer, K., & Pääbo, S. (2016). Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature*, *531*(7595), 504–507. https://doi.org/10.1038/nature17405

Meyer, M., Kircher, M., Gansauge, M. T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer, K., De Filippo, C., Sudmant, P. H., Alkan, C., Fu, Q., Do, R.,

Rohland, N., Tandon, A., Siebauer, M., Green, R. E., Bryc, K., … Pääbo, S. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science*, *338*(6104), 222–226. https://doi.org/10.1126/science.1224344

Mihola, O., Trachtulec, Z., Vlcek, C., Schimenti, J. C., & Forejt, J. (2009). A mouse speciation gene encodes a meiotic histone H3 methyltransferase. *Science (New York, N.Y.)*, *323*(5912), 373–375. https://doi.org/10.1126/science.1163601

Muller, H. J. (1942). Isolating mechanisms, evolution, and temperature. *Biological Symposia*, *6*, 71–125.

Myers, S., Bowden, R., Tumian, A., Bontrop, R. E., Freeman, C., MacFie, T. S., McVean, G., & Donnelly, P. (2010). Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science (New York, N.Y.)*, *327*(5967), 876–879. https://doi.org/10.1126/science.1182363

Oliver, P. L., Goodstadt, L., Bayes, J. J., Birtle, Z., Roach, K. C., Phadnis, N., Beatson, S. A., Lunter, G., Malik, H. S., & Ponting, C. P. (2009). Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. *PLoS Genetics*, *5*(12), e1000753. https://doi.org/10.1371/journal.pgen.1000753

Osborne, L. R., Li, M., Pober, B., Chitayat, D., Bodurtha, J., Mandel, A., Costa, T., Grebe, T., Cox, S., Tsui, L. C., & Scherer, S. W. (2001). A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nature Genetics*, *29*(3), 321–325. https://doi.org/10.1038/ng753

Padhi, A., Shen, B., Jiang, J., Zhou, Y., Liu, G. E., & Ma, L. (2017). Ruminant-specific multiple duplication events of PRDM9 before speciation. *BMC Evolutionary Biology*, *17*(1), 79. https://doi.org/10.1186/s12862-017-0892-4

Paigen, K., & Petkov, P. M. (2018). PRDM9 and its role in genetic recombination. *Trends in Genetics : TIG*, *34*(4), 291–300. https://doi.org/10.1016/j.tig.2017.12.017

Parks, M. M., Lawrence, C. E., & Raphael, B. J. (2015). Detecting non-allelic homologous recombination from high-throughput sequencing data. *Genome Biology*, *16*(1), 72. https://doi.org/10.1186/s13059-015-0633-1

Parvanov, E. D., Tian, H., Billings, T., Saxl, R. L., Spruce, C., Aithal, R., Krejci, L., Paigen, K., & Petkov, P. M. (2017). PRDM9 interactions with other proteins provide a link between recombination hotspots and the chromosomal axis in meiosis. *Molecular Biology of the Cell*, *28*(3), 488–499. https://doi.org/10.1091/mbc.E16-09-0686

Patel, A., Horton, J. R., Wilson, G. G., Zhang, X., & Cheng, X. (2016). Structural basis for human PRDM9 action at recombination hot spots. *Genes & Development*, *30*(3), 257–265. https://doi.org/10.1101/gad.274928.115

Petr, M., Pääbo, S., Kelso, J., & Vernot, B. (2019). Limits of long-term selection against Neandertal introgression. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(5). https://doi.org/10.1073/pnas.1814338116

Pfennig, A., & Lachance, J. (2024). The evolutionary fate of Neanderthal DNA in 30,780 admixed genomes with recent African-like ancestry. *bioRxiv*, 2024.07.25.605203. https://doi.org/10.1101/2024.07.25.605203

Plagnol, V., & Wall, J. D. (2006). Possible ancestral structure in human populations. *PLoS Genetics*, *2*(7). https://doi.org/10.1371/journal.pgen.0020105

Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G. V., & Camerini-Otero, R. D. (2014). Recombination initiation maps of individual human genomes. *Science (New York, N.Y.)*, *346*(6211), 1256442. https://doi.org/10.1126/science.1256442

Presgraves, D. C. (2005). Recombination Enhances Protein Adaptation in *Drosophila melanogaster*. *Current Biology*, *15*(18), 1651–1656. https://doi.org/10.1016/j.cub.2005.07.065

Presgraves, D. C. (2008). Sex chromosomes and speciation in Drosophila. *Trends in Genetics: TIG*, *24*(7), 336–343. https://doi.org/10.1016/j.tig.2008.04.007

Presgraves, D. C. (2010). The molecular evolutionary basis of species formation. *Nature Reviews. Genetics*, *11*(3), 175–180. https://doi.org/10.1038/nrg2718

Preska Steinberg, A., Lin, M., & Kussell, E. (2022). Core genes can have higher recombination rates than accessory genes within global microbial populations. *eLife*, *11*, e78533. https://doi.org/10.7554/eLife.78533

Project MinE ALS Sequencing Consortium. (2018). Project MinE: Study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. *European Journal of Human Genetics: EJHG*, *26*(10), 1537–1546. https://doi.org/10.1038/s41431-018-0177-4

Prüfer, K., De Filippo, C., Grote, S., Mafessoni, F., Korlević, P., Hajdinjak, M., Vernot, B., Skov, L., Hsieh, P., Peyrégne, S., Reher, D., Hopfe, C., Nagel, S., Maricic, T., Fu, Q., Theunert, C., Rogers, R., Skoglund, P., Chintalapati, M., … Pääbo, S. (2017). A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science*, *358*(6363). https://doi.org/10.1126/science.aao1887

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., De Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., … Pääbo, S. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, *505*(7481), 43–49. https://doi.org/10.1038/nature12886

Prüfer, K., Stenzel, U., Hofreiter, M., Pääbo, S., Kelso, J., & Green, R. E. (2010). Computational challenges in the analysis of ancient DNA. *Genome Biology*, *11*(5), R47. https://doi.org/10.1186/gb-2010-11-5-r47

Qin, P., & Stoneking, M. (2015). Denisovan Ancestry in East Eurasian and Native American Populations. *Molecular Biology and Evolution*, *32*(10), 2665–2674. https://doi.org/10.1093/molbev/msv141

Quach, H., Rotival, M., Pothlichet, J., Loh, Y.-H. E., Dannemann, M., Zidane, N., Laval, G., Patin, E., Harmant, C., Lopez, M., Deschamps, M., Naffakh, N., Duffy, D., Coen, A., Leroux-Roels, G., Clément, F., Boland, A., Deleuze, J.-F., Kelso, J., … Quintana-Murci, L. (2016). Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. *Cell*, *167*(3), 643-656.e17. https://doi.org/10.1016/j.cell.2016.09.024

Quilodrán, C. S., Rio, J., Tsoupas, A., & Currat, M. (2023). Past human expansions shaped the spatial pattern of Neanderthal ancestry. *Science Advances*, *9*(42), eadg9817. https://doi.org/10.1126/sciadv.adg9817

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18), i333–i339. https://doi.org/10.1093/bioinformatics/bts378

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., Gonzalez, J. R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., … Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–454. https://doi.org/10.1038/nature05329

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L. F., Maricic, T., Good, J. M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E. E., …

Pääbo, S. (2010). Genetic history of an archaic hominin group from Denisova cave in Siberia. *Nature*, *468*(7327). https://doi.org/10.1038/nature09710

Reilly, P. F., Tjahjadi, A., Miller, S. L., Akey, J. M., & Tucci, S. (2022). The contribution of Neanderthal introgression to modern human traits. *Current Biology*, *32*(18). https://doi.org/10.1016/j.cub.2022.08.027

Rinker, D. C., Simonti, C. N., McArthur, E., Shaw, D., Hodges, E., & Capra, J. A. (2020). Neanderthal introgression reintroduced functional ancestral alleles lost in Eurasian populations. *Nature Ecology and Evolution*, *4*(10), 1332–1341. https://doi.org/10.1038/s41559-020-1261-z

Rosenfeld, J. A., & Patel, A. (2017). Chromosomal Microarrays: Understanding Genetics of Neurodevelopmental Disorders and Congenital Anomalies. *Journal of Pediatric Genetics*, *6*(1), 42–50. https://doi.org/10.1055/s-0036-1584306

Sams, A. J., Dumaine, A., Nédélec, Y., Yotova, V., Alfieri, C., Tanner, J. E., Messer, P. W., & Barreiro, L. B. (2016). Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. *Genome Biology*, *17*(1), 246. https://doi.org/10.1186/s13059-016-1098-6

Sánchez-Quinto, F., & Lalueza-Fox, C. (2015). Almost 20 years of Neanderthal palaeogenetics: Adaptation, admixture, diversity, demography and extinction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1660). https://doi.org/10.1098/rstb.2013.0374

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., & Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, *507*(7492). https://doi.org/10.1038/nature12961

Sankararaman, S., Mallick, S., Patterson, N., & Reich, D. (2016). The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Current Biology*, *26*(9). https://doi.org/10.1016/j.cub.2016.03.037

Schaefer, N. K., Shapiro, B., & Green, R. E. (2021). An ancestral recombination graph of human, Neanderthal, and Denisovan genomes. *Science Advances*, *7*(29), eabc0776. https://doi.org/10.1126/sciadv.abc0776

Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, *46*(8), 919–925. https://doi.org/10.1038/ng.3015

Schilthuizen, M., Giesbers, M. C. W. G., & Beukeboom, L. W. (2011). Haldane's rule in the 21st century. *Heredity*, *107*(2), 95–102. https://doi.org/10.1038/hdy.2010.170

Schumer, M., Xu, C., Powell, D. L., Durvasula, A., Skov, L., Holland, C., Blazier, J. C., Sankararaman, S., Andolfatto, P., Rosenthal, G. G., & Przeworski, M. (2018). Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science*, *360*(6389), 656–660. https://doi.org/10.1126/science.aar3684

Schwartz, J. J., Roach, D. J., Thomas, J. H., & Shendure, J. (2014). Primate evolution of the recombination regulator PRDM9. *Nature Communications*, *5*(1), 4370. https://doi.org/10.1038/ncomms5370

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R Journal*, *8*(1). https://doi.org/10.32614/rj-2016-021

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*(6), 461–468. https://doi.org/10.1038/s41592-018-0001-7

Seehausen, O. (2004). Hybridization and adaptive radiation. *Trends in Ecology & Evolution*, *19*(4), 198–207. https://doi.org/10.1016/j.tree.2004.01.003

Sfeir, A., & Symington, L. S. (2015). Microhomology-mediated end joining: A back-up survival mechanism or dedicated pathway? *Trends in Biochemical Sciences*, *40*(11), 701–714. https://doi.org/10.1016/j.tibs.2015.08.006

Shaikh, T. H., Kurahashi, H., Saitta, S. C., O'Hare, A. M., Hu, P., Roe, B. A., Driscoll, D. A., McDonald-McGinn, D. M., Zackai, E. H., Budarf, M. L., & Emanuel, B. S. (2000). Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: Genomic organization and deletion endpoint analysis. *Human Molecular Genetics*, *9*(4), 489–501. https://doi.org/10.1093/hmg/9.4.489

Skov, L., Coll Macià, M., Sveinbjörnsson, G., Mafessoni, F., Lucotte, E. A., Einarsdóttir, M. S., Jonsson, H., Halldorsson, B., Gudbjartsson, D. F., Helgason, A., Schierup, M. H., & Stefansson, K. (2020). The nature of Neanderthal introgression revealed by 27,566 Icelandic genomes. *Nature*, *582*(7810). https://doi.org/10.1038/s41586-020-2225-9

Skov, L., Hui, R., Shchur, V., Hobolth, A., Scally, A., Schierup, M. H., & Durbin, R. (2018). Detecting archaic introgression using an unadmixed outgroup. *PLoS Genetics*, *14*(9). https://doi.org/10.1371/journal.pgen.1007641

Skov, L., Peyrégne, S., Popli, D., Iasi, L. N. M., Devièse, T., Slon, V., Zavala, E. I., Hajdinjak, M., Sümer, A. P., Grote, S., Bossoms Mesa, A., López Herráez, D., Nickel, B., Nagel, S., Richter, J., Essel, E., Gansauge, M., Schmidt, A., Korlević, P., … Peter, B. M. (2022). Genetic insights into the social organization of Neanderthals. *Nature*, *610*(7932), 519–525. https://doi.org/10.1038/s41586-022-05283-y

Smagulova, F., Gregoretti, I. V., Brick, K., Khil, P., Camerini-Otero, R. D., & Petukhova, G. V. (2011). Genome-Wide Analysis Reveals Novel Molecular Features of Mouse Recombination Hotspots. *Nature*, *472*(7343), 375–378. https://doi.org/10.1038/nature09869

Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., Ibrahim, A., Ji, Y., John, S., Lewis, E., MacArthur, J. A. L., McMahon, A., Osumi-Sutherland, D., Panoutsopoulou, K., Pendlington, Z., … Harris, L. W. (2023). The NHGRI-EBI GWAS Catalog: Knowledgebase and deposition resource. *Nucleic Acids Research*, *51*(D1), D977–D985. https://doi.org/10.1093/nar/gkac1010

Stankiewicz, P., & Lupski, J. R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends in Genetics*, *18*(2), 74–82. https://doi.org/10.1016/S0168-9525(02)02592-1

Startek, M., Szafranski, P., Gambin, T., Campbell, I. M., Hixson, P., Shaw, C. A., Stankiewicz, P., & Gambin, A. (2015). Genome-wide analyses of LINE–LINE-mediated nonallelic homologous recombination. *Nucleic Acids Research*, *43*(4), 2188–2198. https://doi.org/10.1093/nar/gku1394

Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V. G., Desnica, N., Hicks, A., Gylfason, A., Gudbjartsson, D. F., Jonsdottir, G. M., Sainz, J., Agnarsson, K., Birgisdottir, B., Ghosh, S., … Stefansson, K. (2005). A common inversion under selection in Europeans. *Nature Genetics*, *37*(2), 129–137. https://doi.org/10.1038/ng1508

Steinberg, K. M., Antonacci, F., Sudmant, P. H., Kidd, J. M., Campbell, C. D., Vives, L., Malig, M., Scheinfeldt, L., Beggs, W., Ibrahim, M., Lema, G., Nyambo, T. B.,

Omar, S. A., Bodo, J.-M., Froment, A., Donnelly, M. P., Kidd, K. K., Tishkoff, S. A., & Eichler, E. E. (2012). Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nature Genetics*, *44*(8), 872–880. https://doi.org/10.1038/ng.2335

Stringer, C. (2012). The status of Homo heidelbergensis (Schoetensack 1908). *Evolutionary Anthropology: Issues, News, and Reviews*, *21*(3), 101–107. https://doi.org/10.1002/evan.21311

Stringer, C. B., & Andrews, P. (1988). Genetic and Fossil Evidence for the Origin of Modern Humans. *Science*, *239*(4845), 1263–1268. https://doi.org/10.1126/science.3125610

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

Sudmant, P. H., Mallick, S., Nelson, B. J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe, B. P., Baker, C., Nordenfelt, S., Bamshad, M., Jorde, L. B., Posukh, O. L., Sahakyan, H., Watkins, W. S., Yepiskoposyan, L., Abdullah, M. S., Bravi, C. M., Capelli, C., Hervig, T., … Eichler, E. E. (2015). Global diversity, population stratification, and selection of human copy-number variation. *Science (New York, N.Y.)*, *349*(6253), aab3761. https://doi.org/10.1126/science.aab3761

Sümer, A. P., Rougier, H., Villalba-Mouco, V., Huang, Y., Iasi, L. N. M., Essel, E., Bossoms Mesa, A., Furtwaengler, A., Peyrégne, S., de Filippo, C., Rohrlach, A. B., Pierini, F., Mafessoni, F., Fewlass, H., Zavala, E. I., Mylopotamitaki, D., Bianco, R.

A., Schmidt, A., Zorn, J., … Krause, J. (2025). Earliest modern human genomes constrain timing of Neanderthal admixture. *Nature*, *638*(8051), 711–717. https://doi.org/10.1038/s41586-024-08420-x

Sweigart, A. L., & Willis, J. H. (2012). Molecular evolution and genetics of postzygotic reproductive isolation in plants. *F1000 Biology Reports*, *4*, 23. https://doi.org/10.3410/B4-23

Szecówka, K., Misiak, B., Łaczmańska, I., Frydecka, D., & Moustafa, A. A. (2023). Copy Number Variations and Schizophrenia. *Molecular Neurobiology*, *60*(4), 1854–1864. https://doi.org/10.1007/s12035-022-03185-8

Tassi, F., Ghirotto, S., Mezzavilla, M., Vilaça, S. T., De Santi, L., & Barbujani, G. (2015). Early modern human dispersal from Africa: Genomic evidence for multiple waves of migration. *Investigative Genetics*, *6*(1). https://doi.org/10.1186/s13323-015-0030-2

Team, R. C. (2021). R: A Language and Environment for Statistical Computing. In *R Foundation for Statistical Computing* [Computer software].

Telis, N., Aguilar, R., & Harris, K. (2020). Selection against archaic hominin genetic variation in regulatory regions. *Nature Ecology and Evolution*, *4*(11). https://doi.org/10.1038/s41559-020-01284-0

Turner, D. J., Miretti, M., Rajan, D., Fiegler, H., Carter, N. P., Blayney, M. L., Beck, S., & Hurles, M. E. (2008). Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nature Genetics*, *40*(1), 90–95. https://doi.org/10.1038/ng.2007.40

UK Biobank Whole-Genome Sequencing Consortium. (2025). Whole-genome sequencing of 490,640 UK Biobank participants. *Nature*, *645*(8081), 692–701. https://doi.org/10.1038/s41586-025-09272-9

Vernot, B., & Akey, J. M. (2014). Resurrecting surviving Neandertal lineages from modern human genomes. *Science*, *343*(6174). https://doi.org/10.1126/science.1245938

Vicari, S., Napoli, E., Cordeddu, V., Menghini, D., Alesi, V., Loddo, S., Novelli, A., & Tartaglia, M. (2019). Copy number variants in autism spectrum disorders. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *92*, 421–427. https://doi.org/10.1016/j.pnpbp.2019.02.012

Vilanea, F. A., & Schraiber, J. G. (2019). Multiple episodes of interbreeding between Neanderthal and modern humans. *Nature Ecology and Evolution*, *3*, 39–44.

Wei, X., Robles, C. R., Pazokitoroudi, A., Ganna, A., Gusev, A., Durvasula, A., Gazal, S., Loh, P.-R., Reich, D., & Sankararaman, S. (2022). The lingering effects of Neanderthal introgression on human complex traits. *bioRxiv*.

Weischenfeldt, J., Symmons, O., Spitz, F., & Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: Insights from and for human disease. *Nature Reviews Genetics*, *14*(2), 125–138. https://doi.org/10.1038/nrg3373

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Wetzel, A. S., & Darbro, B. W. (2022). A comprehensive list of human microdeletion and microduplication syndromes. *BMC Genomic Data*, *23*(1), 82. https://doi.org/10.1186/s12863-022-01093-3

Wigginton, J. E., Cutler, D. J., & Abecasis, G. R. (2005). A note on exact tests of Hardy-Weinberg equilibrium. *American Journal of Human Genetics*, *76*(5). https://doi.org/10.1086/429864

Witt, K. E., Villanea, F., Loughran, E., Zhang, X., & Huerta-Sanchez, E. (2022). Apportioning archaic variants among modern populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1852). https://doi.org/10.1098/rstb.2020.0411

Wu, H., Mathioudakis, N., Diagouraga, B., Dong, A., Dombrovski, L., Baudat, F., Cusack, S., de Massy, B., & Kadlec, J. (2013). Molecular basis for the regulation of the H3K4 methyltransferase activity of PRDM9. *Cell Reports*, *5*(1), 13–20. https://doi.org/10.1016/j.celrep.2013.08.035

Yair, S., Lee, K. M., & Coop, G. (2021). The timing of human adaptation from Neanderthal introgression. *Genetics*, *218*(1). https://doi.org/10.1093/genetics/iyab052

Zarrei, M., MacDonald, J. R., Merico, D., & Scherer, S. W. (2015). A copy number variation map of the human genome. *Nature Reviews. Genetics*, *16*(3), 172–183. https://doi.org/10.1038/nrg3871

Zeberg, H., & Pääbo, S. (2021). A genomic region associated with protection against severe COVID-19 is inherited from Neandertals. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(9). https://doi.org/10.1073/pnas.2026309118

Zheng-Bradley, X., Streeter, I., Fairley, S., Richardson, D., Clarke, L., & Flicek, P. (2017). Alignment of 1000 Genomes Project reads to reference assembly GRCh38. *GigaScience*, *6*(7), 1–8. https://doi.org/10.1093/gigascience/gix038

Zhou, Y., Browning, B. L., & Browning, S. R. (2020). Population-Specific Recombination Maps from Segments of Identity by Descent. *American Journal of Human Genetics*, *107*(1), 137–148. https://doi.org/10.1016/j.ajhg.2020.05.016

Zhou, Y., Duvaux, L., Ren, G., Zhang, L., Savolainen, O., & Liu, J. (2017). Importance of incomplete lineage sorting and introgression in the origin of shared genetic variation between two closely related pines with overlapping distributions. *Heredity*, *118*(3), 211–220. https://doi.org/10.1038/hdy.2016.72