University of Southampton

Faculty of Medicine

Human Development and Health

Development and Application of Powerful Methods for Identifying Selective Sweeps

By

Clare Horscroft

ORCID ID 0000-0001-5679-5912

Thesis for the degree of Doctor of Philosophy

July 2021

University of Southampton

Abstract

Faculty of Medicine

Human Development and Health

Thesis for the degree of Doctor of Philosophy

Development and Application of Powerful Methods for Identifying Selective Sweeps

by

Clare Horscroft

Identifying regions of the genome under selection is an ongoing effort in population genetics and quantitative biology. Selection creates signals within the genome that can be detected using a variety of methods. One of the major obstacles is the effect of variable recombination rates that create regions of the genome containing highly correlated variants, confounding methods that rely on correlations as indicators of selection. This thesis is mainly focussed on developing the Z_α method, which is a statistic that uses linkage disequilibrium patterns to identify sweeps while also integrating the recombination rate.

As recombination rates are an important confounder, and a key component of the Z_α method, recombination rates in human populations were studied further. The aim was to see if populations with different ancestral backgrounds have different recombination rates, and if so, at what scale. This work shows that different human populations have similar recombination patterns at the wide scale, but at the fine scale they can be quite different. This result means that individual recombination maps are required for each population when using the Z_α method.

To easily and efficiently generate the Z_α statistics, an R package was developed, published, and made freely available. R is an open source programming language, which increases reproducibility, transparency and reliability of the method and any results generated using it. Finally, the statistic was applied to the genome of the domestic dog. Firstly, the recombination map was generated, and then the new R package was used to apply the Z_α statistics including adjusting for the recombination rate. This study identified candidate regions for selection in the dog genome, both previously published and novel to this study.

Table of Contents

Table of Tables

Table of Figures

Table of Figures

List of Publications

- **Horscroft, C**., Pengelly, R. J., Sluckin, T. J. and Collins, A. (2020) *zalpha: an R package for the identification of regions of the genome under selection*. Journal of Open Source Software; **5**(56):2638.
- Jabalameli, M. R., **Horscroft, C**., Vergara-Lope, A., Pengelly, R. J. and Collins, A. (2019) *Gene-dense autosomal chromosomes show evidence for increased selection*. Heredity; **123**(6):774-783.
- Vergara-Lope, A., Jabalameli, M. R., **Horscroft, C.**, Ennis, S., Collins, A. and Pengelly, R. J. (2019) *Linkage disequilibrium maps for European and African populations constructed from whole genome sequence data*. Scientific Data; **6**(1):208.
- **Horscroft, C**., Ennis, S., Pengelly, R. J., Sluckin, T. J. and Collins, A. (2018) *Sequencing era methods for identifying signatures of selection in the genome*. Brief Bioinform; **20**(6):1997-2008.

Research Thesis: Declaration of Authorship

Print name: CLARE HORSCROFT

Title of thesis: Development and Application of Powerful Methods for Identifying Selective Sweeps

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:-

Horscroft, C., Pengelly, R. J., Sluckin, T. J. and Collins, A. (2020) zalpha: an R package for the identification of regions of the genome under selection. Journal of Open Source Software; 5(56):2638.

Jabalameli, M. R., Horscroft, C., Vergara-Lope, A., Pengelly, R. J. and Collins, A. (2019) Gene-dense autosomal chromosomes show evidence for increased selection. Heredity; 123(6):774-783.

Horscroft, C., Ennis, S., Pengelly, R. J., Sluckin, T. J. and Collins, A. (2018) Sequencing era methods for identifying signatures of selection in the genome. Brief Bioinform; 20(6):1997-2008.

Signature: ..... Date: 27th July 2021

Acknowledgements

I would like to thank my supervisors Andy, Reuben, Tim, and Sarah, for their unwavering support and unlimited amounts of patience though the entirety of my PhD. I have learnt an incredible amount from each of them, and I would not be the researcher I am today without their input.

I would also like to thank everyone who has mentored me over the years, be that by letting me demonstrate on their courses, giving advice about future careers, or supporting me in any other capacity. I truly appreciate the time and effort, and I hope to pay your kindness forward in the future.

I thank my examiners over the years for your insights and advice. I appreciate the time you took away from your own research to read and understand mine, and for making my exams an enjoyable experience!

Thank you to the University of Southampton support services: IT, Student services, the IRIDIS HPC team, and the Doctoral College, amongst others. The guidance and support given by the people involved in these services was vital to my success.

To all the participants (and their pets!) in the datasets I used in this thesis, thank you for agreeing for your personal information to be used to further scientific knowledge. Your contributions are essential and hugely appreciated.

To the Genomics Informatics team, I would like to thank you all for the fantastic support, talks and socials over the years. It is a shame in the last year there weren't quite so many chats in the kitchen with the weird kettle, but know I always appreciated catching up with everyone.

Carolina and Imogen, without you two I literally would not have gotten through these three-and-a-bit years. Thanks for always being there for advice, rants, and comradery!

Everyone involved in the Life Sciences Postgraduate Society deserves a huge thanks for the time and effort they put in to providing a safe space to network with our peers and learn for each other, as well as organising some fantastic socials. My time as a postgraduate researcher would have been very dull without you, so thank you.

Of course, thanks to my parents and to Jamie for being my absolute rocks. I appreciate the support more than I can put into words. To the rest of my family and friends, thank you for listening to me, entertaining me, and for maintaining my mental wellbeing throughout this whole process.

Definitions and Abbreviations

3P-CLR	Three Population Composite Likelihood Ratio
ABS	Ancestral Branch Statistic
ADH	Alcohol Dehydrogenase
AGVP	African Genome Variation Project
ALDH	Acetaldehyde Dehydrogenase
AODE	Averaged One-Dependence Estimator
AR	Autoregressive
ASMC	Ascertained Sequentially Markovian Coalescent
AUC	Area Under the Curve
BALLET	Balancing selection Likelihood Test
bp	base pair
BGC	Biased Gene Conversion
CEU	1000 Genomes Project sub-population: Northern Europeans from Utah
ChIP	Chromatin Immunoprecipitation
Chr	Chromosome
CL	Composite Likelihood
CLR	Composite Likelihood Ratio
cM	centimorgan
CMS	Composite of Multiple Signals
CNA	Copy Number Alteration
CNN	Convolutional Neural Network
CNV	Copy Number Variant
COI	Cone Of Influence
CRAN	The Comprehensive R Archive Network

Definitions and Abbreviations

CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CSS	Composite Selection Signals
CWT	Continuous Wavelet Transform
DAF	Derived Allele Frequency
DCMS	De-correlated Composite of Multiple Signals
DNA	Deoxyribonucleic Acid
DOF	Degrees Of Freedom
DWT	Discrete Wavelet Transform
EBP	Earth BioGenome Project
EBT	Effective Bottleneck Time
EHH	Extended Haplotype Homozygosity
EM	Expectation-Maximisation
EOS	Extreme Outlier Set
FastEPRR	Fast Estimation of Population Recombination Rates
FDR	False Discovery Rate
FISH	Fluorescence *In Situ* Hybridization
GeCIP	Genomics England Clinical Interpretation Partnership
GRC	Genome Reference Consortium
GWAS	Genome-Wide Association Study
GWSS	Genome-Wide Selection Scan
HacDivSel	Haplotype allelic class – Divergent Selection
HIV	Human Immunodeficiency Virus
HKA	Hudson–Kreitman– Aguadé test
HWE	Hardy-Weinberg Equilibrium
IBD	Identical By Descent
IBS	Identical By State

ICU	Intensive Care Unit
iHS	integrated Haplotype Score
iSAFE	integrated Selection of Allele Favoured by Evolution
iSMC	integrative Sequentially Markov Coalescent
Kb	Kilobase
kNN	k-Nearest Neighbour
kya	Thousand years ago
LD	Linkage Disequilibrium
LDA	Linear Discriminant Analysis
LDU	Linkage Disequilibrium Unit
LSD	Levels of exclusively Shared Differences
LUCA	Last Universal Common Ancestor
M	Morgan
MAF	Minor Allele Frequency
Mb	Megabase
MCMC	Markov Chain Monte Carlo
MODWT	Maximal Overlap Discrete Wavelet Transform
MRCA	Most Recent Common Ancestor
N_e	Effective population size
NGS	Next Generation Sequencing
nS_L	number of Segregating sites per Length
OOA	Out Of Africa
pAUC	partial Area Under the Curve
PBS	Population Branch Statistic
PC	Principal Component
PCA	Principal Component Analysis

PSMC	Pairwise Sequentially Markovian Coalescent
QMF	Quadrature Mirror Filter
RAiSD	Raised Accuracy in Sweep Detection
rjMCMC	reversible-jump Markov Chain Monte Carlo
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
ROH	Runs Of Homozygosity
SDS	Singleton Density Score
SFS	Site Frequency Spectrum
S/HIC	Soft/Hard Inference through Classification
SMRT	Single Molecule Real-Time
sNMF	sparse Nonnegative Matrix Factorisation
SNP	Single Nucleotide Polymorphism
snRNA	small nuclear RNA
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
SWIF(r)	Sweep Inference Framework (controlling for correlation)
TE	Transposable Element
TMRCA	Time to Most Recent Common Ancestor
TSel	Time to most recent common ancestor Selection
TSS	Transcription Start Site
VCF	Variant Call Format
VEP	Ensembl's Variant Effect Predictor tool
WGS	Whole Genome Sequencing
XP-CLR	Cross-Population Composite Likelihood Ratio
XP-EHH	Cross-Population Extended Haplotype Homozygosity

Section 1 Introduction to selection

In this section I introduce the concept of selection in genetics, the history of evolutionary thought, and describe the basics of genetic variation and inheritance. I go on to discuss methods for finding candidate regions of the genome where a selective event may have occurred. In final chapter of this section I apply a subset of these methods to simulated data to see if the methods can distinguish between neutral datasets and those containing sweeps. I go on to apply the Z_α statistic to a human dataset as a proof of concept.

Parts of this section were written up and published in the journal Briefings in Bioinformatics [1]. I would like to thank the anonymous reviewers for their comments and suggestions including running extra simulations, which greatly enhanced this work.

Chapter 1 Introduction

1.1 Aim of this thesis

This thesis aims to take the reader through the process of researching methods for identifying selection, choosing one to develop further, testing it and creating software for it, and then applying it to real world data. The method chosen, the Z_α statistic, is based upon fluctuations in linkage disequilibrium within the genome. This is highly correlated with recombination, and so a large portion of this thesis is dedicated to understanding recombination within different populations. Thus, the goals of this thesis were fourfold:

1) To investigate the methods available to researchers to identify regions of the genome under selection
2) To investigate linkage disequilibrium maps and recombination
3) To develop the Z_α statistic
4) To apply the Z_α statistic to population genomic data

Within this first chapter the motivation, history, and biology of evolutionary research will be explored. Firstly, evolution will be discussed, which will include examples of evolution in human history. After this, details of genetic variation will be covered, including how selective events manifest in a population. Finally, some past selection events will be discussed. While methods for identifying selection in the genome will be touched on here, they will be covered in much greater detail in the next chapter.

1.2 Motivations for studying evolution

Humans (*Homo sapiens*) are just one of millions of species currently inhabiting the Earth, from large animals and plants through to bacteria and other single celled organisms [2]. All these species originally came from one population of organisms, known as the last universal common ancestor (LUCA), perhaps as long as 4.5 billion years ago [3]. By studying evolution, we as humans can learn about our place in the world and how we and our fellow extant species adapted to our environments time and time again to still be here today. This context of humans as just one of many survivors, and what our purpose is on Earth within this context, is still debated and has relevance socially and culturally.

Simply by observing, one can note the incredible range of phenotypes expressed by organisms all over the Earth. One purpose for studying this variation is to group individuals into species and

then classify those into families and other taxonomies. By studying the similarities and differences between species, discoveries can be made about how and why they came to be that way. For example, animals living in colder regions tend to have shorter limbs than animals living in hot climates, which is a trait that has evolved and is observable within and between many species [4]. Vestigial structures, and other oddities such as the suboptimal length and route of the laryngeal nerve in giraffes (G*iraffa*), would be inexplicable without understanding the evolutionary history of the creature [5]. Studying evolution and genetics has led to the discovery that some creatures that were originally thought to be closely related, have in fact a much longer time to the most recent common ancestor than was previously thought, such as horseshoe crabs (*Xiphosura*) and crustaceans [6]. On the broader scale, it was discovered that fungi are more closely related to humans than plants as was assumed [7], and that archaea are a domain separate from bacteria [8]. Organising and classifying organisms in this way increases knowledge of the history of life and the ways in which organisms have adapted to the different environments that they inhabit. It can also give clues as to the selective pressures that they must have overcome in the past and hints at how species may evolve in the future.

A more basic motivation to study evolution than the drive to add knowledge is the need to feed ourselves and future generations. Identifying how crops and livestock adapt and change to different environments and consumer desires is important practically, but also financially, another very human motivation. By studying how crops and livestock evolve, breeds can be optimised to fit a niche, for example chickens bred for egg laying and broiler lines for meat [9]. This increases output in terms of calories per individual as well as maximising profitability. Studying evolution allows for predictions to be made about the future challenges given events of the past. The anticipated increase in global temperatures due to climate change could have a large impact on food availability and studying the genomes of relevant populations can elucidate how they adapted to thermal fluctuations previously [10]. Work is underway to test heat-stress in both crops [11] and livestock [12]. It is important to start breeding animals now that have the ability to cope with rising temperatures and other effects of climate change, and previous work on selection is aiding this goal [13].

Medicine is a field that has benefitted from utilising evolutionary theory to explain how and why illness and disease occurs. This can be studied from two angles: from the perspective of the sick individual, and the perspective of the disease. For humans, evolutionary medicine has helped illuminate mechanisms around immunity [14], as well as explaining why some diseases have not been selected out of existence. For example, cancer is caused by somatic mutations and so cannot be selected out by natural selection; however, techniques from evolutionary biology can be applied to the cancer cells themselves within a single individual [15]. Mechanisms such as

balancing selection, where there are trade-offs between the risk of developing a disease but also being protected from another, can explain why some diseases are still fairly common within a population despite the clear negative effect on the individual [16]. Studying evolution has also increased knowledge of how diseases such as viruses adapt and spread, which is of use to epidemiologists studying disease at a population level [17]. Antimicrobial resistance is a current threat to global health [18], so understanding how microbial populations are becoming immune to medicine like antibiotics will be vital to developing treatments in the near future [19].

1.3 Evolution and natural selection

Some of the earliest references to evolutionary thought appear to be from Ancient Greek philosophers, including Anaximander, who suggested humans came from a water-based creature [20]; Xenophanes, who theorised about gradual changes in the land and sea distribution of Earth based on fossil evidence of sea creatures found inland [21]; and Empedocles, who wrote on the origin of humans, animals, and plants [22]. Aristotle posited theories around inheritance, as well as describing one of the first taxonomic systems, classifying hundreds of animals in his books *History of Animals* and *Parts of Animals* in the 4th century BC [23]. He also inspired his student Theophrastus to classify plants, who went on to write *Enquiry into Plants* [24].

Around the same time, Chinese philosophers also had ideas about species and life. Zhuang Zhou, the author of an important Taoist text, wrote about species becoming other species, a concept now called transmutation [25]. A few centuries later, a poem by Lucretius talks of species becoming extinct if they do not possess certain traits such as speed and strength [26]. He also claims some animals would have died out had they not been useful to humans. Domestication of plants and animals was vital to sustaining human population growth, as will be discussed later in this chapter.

Several Islamic philosophers also tackled the concept of evolution. Al-Jāḥiẓ wrote his *Book of the Animals* in the 9th century, describing hundreds of species of animals similarly to Aristotle's book over a thousand years earlier [27]. Al-Jāḥiẓ also discusses the food chain and the skills animals must possess to survive. In the 10th century, Ibn Miskawayh wrote about evolution, describing species evolving into each other over time, for example minerals into plants and plants into animals [28]. Similar ideas were also expressed in the *Encyclopaedia of the Brethren of Purity* written by a secret society of philosophers around the same time [28]. In the 14th century, Ibn Khaldun wrote the *Muqaddimah* in which he describes a continuum of species, the formation of new species over time, and emphasises the close relationship between humans and monkeys [29].

In the 18th century, Swedish scientist Linnaeus published the *Systema Naturae*, a work that by its final edition contained classifications for thousands of plants and animals [30]. In this work he used binomial nomenclature to refer to species, assigning each a simple two-word name - a system that is still in use today [31]. He grouped humans with apes and monkeys, a decision that was controversial at the time [32]. He also defined human variations by anatomical but also cultural and moral differences. This was one of the earliest examples of scientific racism, which allowed for future scientists to make ideological inferences about the morality of human populations based on anatomic differences [33].

Around the same time as Linnaeus, three Frenchmen were also independently considering evolution. Maupertuis wrote about heredity, that characteristics of the parents can be passed down to the offspring, and about natural selection, suggesting that only the fittest species survived and that many others had not [34]. Buffon published the *Histoire Naturelle*, a multi-volume encyclopaedia on animals and minerals [35]. He observed that many animals and plants were distinct in different regions, regardless of how similar the climate was. He also believed that all humans had a single origin, with differences due to environmental factors; however, he also postulated that the observable differences were a degeneration from the original race of Caucasians [36]. Meanwhile, Diderot was contributing to his own *Encyclopédie*, and wrote about evolution without intelligent design - an act which saw him jailed [37].

Another Frenchman, Lamarck, noticed that animals had adapted to the environment in which they live [38]. To explain this, his idea was that species acquired traits throughout their lives then passed those down to the next generation, a theory now called Lamarckian inheritance. While this was debunked, modern studies into epigenetic phenomena such as inheritable methylation markers that alter DNA expression are sometimes referred to as a kind of Lamarckian inheritance [39]. An example of this is the Dutch famine birth cohort study, which found a difference in the size of babies born to mothers whose mothers were undernourished during pregnancy due to the famine than to controls who were unaffected by the famine [40].

Charles Darwin was born in the early 19th century to a family with a history evolutionary thought: his grandfather, Erasmus Darwin, had mused on the origin of life and had speculated that all life came from one common ancestor in his work *Zoonomia* [41]. After going on a five year voyage, studying the specimens he encountered and talking to people in many professions including animal breeders, Darwin published both a paper with Wallace [42] and his own work *On the Origin of Species* outlining his theories on evolution [43]. In this work, he describes natural selection as a process where beneficial traits are maintained in populations over time, and detrimental traits are lost. He explains that how, as different populations select for different

traits, new species are formed. He also defines artificial selection as when humans deliberately breed animals or plants with the intent of propagating desirable traits.

Darwin's theory of natural selection relied on the concept on inheritance, the mechanisms of which were not clear at this time. Mendel was a monk who performed experiments on inheritance in pea plants (*Pisum sativum*). During this work, he discovered the concept of dominance, where one trait will override another, and of segregation, where an individual will randomly pass down the traits it received from its parents [44]. This second discovery is observable when two heterozygotes for a trait with dominant and recessive forms are crossed, resulting in the recessive trait reappearing in a quarter of the offspring. He also described random assortment, the idea that each trait is inherited independently of any others. While true for some traits, genetic linkage means this often is not true, see section 1.7.

Although Mendel's work was originally published in 1866, it was not combined with Darwin's ideas until the early 20th century, when Fisher and Haldane independently used mathematics to show that natural selection could occur by way of Mendelian genetics [45, 46]. This combining of ideas and methods of researching evolution was later dubbed "modern synthesis" by Huxley [47]. He linked together the ideas of natural selection and competition with inheritance, mutation, and variation. Sewall Wright was also working on population genetics at the same time, focussing on inbreeding and genetic drift [48]. In addition, he authored the F-statistics, which will be discussed in the next chapter [49].

Dobzhansky, working with Morgan on fruit flies (*Drosophila pseudoobscura*), wrote about variation both in the lab flies and in wild flies. He presented evidence that genetic variation is common, especially between sub-populations, and that it is genetic changes in populations that motivate natural selection [50]. After Morgan had shown that genes are situated along chromosomes, Dobzhansky also wrote about gene interactions and chromosomal rearrangements causing evolutionary changes [51].

In 1942, Mayr published his book *Systematics and the Origin of Species*, within which he discusses speciation, especially by way of geographically separated sub-populations eventually becoming so genetically different that they cannot interbreed any more [52]. He also argued that genes cannot be taken in isolation in regard to natural selection, recognising that through genetic linkage, selected variants will always be connected to and influenced by other genes [53].

The modern synthesis established that evolutionary change is driven by variations in genetics, both within and between species. This was generally accepted by the mid-20th century, and saw the rise of population genetics and evolutionary biology [54]. As new technologies and methods

were developed throughout the 20th century, molecular biology became an important part of evolutionary research [55]. The rest of this chapter mainly focusses on genetic variation and the process of evolution.

1.4 Data sources

As seen throughout the last section, for most of the history of evolutionary research, ideas and classifications came from observation through study of anatomy and physiology. It was by comparing the physical features of different creatures that Linnaeus created his taxonomical system and Darwin formed his ideas on evolution. While observation was and still is useful, there are now many more ways in which evolution and selection can be studied.

While Xenophanes was using fossilised evidence as far back as the 6th century BC, it was not until the 18th century, when Cuvier studied fossils and came to the conclusion that they must belong to extinct species, that palaeontology became its own field of science [56, 57]. In 1944, Simpson published *Tempo and Mode in Evolution*, a book combining the fields of palaeontology with the studies of selection in genetics, as part of the modern synthesis on evolutionary thought [58]. He showed that the fossil record substantiated the theories of evolution and natural selection.

Palaeoanthropology, the study of early human remains, has been vital to understanding the history of the human species. By studying and dating bones discovered across the world, scientists have been able to piece together the evolution of man; from *Graecopithecus freybergi* fossils dated over 7 million years ago [59], to recent extinct subspecies of archaic humans such as Neanderthals (*Homo neanderthalensis*) [60] and Denisovans (*Homo denisova*) [61]. Some findings contained traces of ancient DNA that have been compared to modern human genomes to, among other things, calculate the amount of genetic material contributed from extinct species and determine genetic diversity both within and between species [62, 63]. Studying bones and other artefacts from around the world has enabled archaeologists to make inferences about when and where modern humans began and how they then spread across the globe [64]. More detail on this will be given in section 1.11.

DNA was first discovered in 1869 by Miescher, although it was not until 1944 that Avery *et al*. showed that DNA is the substance containing heritable genetic material, as confirmed by the Hershey-Chase experiment in 1952 [65-67]. In 1953, the double-helix structure of DNA was discovered by Franklin, Watson, and Crick, using the famous Photo 51 [68, 69].

While chromosomes were first discovered in 1842 by Nägeli [70], it took until 1956 before Tjio published that humans possess 46 chromosomes [71]. This was the beginning of cytogenetics and

karyotyping, a method for observing chromosomes. The medical implications of this included observing aneuploidy, the underlying cause of conditions such as Down's syndrome [72] and Patau syndrome [73], and conditions caused by sex chromosome differences such as Turner's syndrome [74] and Klinefelter syndrome [75]. Methods include fluorescence *in situ* hybridization (FISH) [76] and Giemsa banding [77]. While humans have 46 chromosomes, our closest relative the chimpanzee has 48, the same as the other great apes (*Hominidae*) [78]. This is due to a fusion of two different chromosomes into what is now known in humans as chromosome 2 [79]. A vestigial centromere is present (chromosomes should only possess one), and telomeric regions (usually found at each end of a chromosome) can be observed within the chromosome as well as at each end. Neanderthals and Denisovans also possessed 23 pairs of chromosomes, meaning the fusion must predate our most recent common ancestor [80].

The evolution of DNA sequencing techniques is described in generations, from first generation sequencing thorough to the current third generation. The first generation began in the 1970s when Sanger published his chain-termination method and sequenced the whole genome of a virus [81, 82]. Simultaneously, Maxam and Gilbert were also working on a sequencing technique, based on DNA chemical modification and cleavage [83]. As these sequencing methods only work for short strings of DNA, the shotgun sequencing technique was developed to allow longer fragments to be analysed by reducing them to smaller fragments and reassembling them later computationally [84, 85]. The polymerase chain reaction (PCR) method was developed in the 1980s to amplify DNA by making many copies of a DNA sample in a short amount of time [86]. This technique is still used today and is especially useful in cases where there is only a very small DNA sample to be analysed, such as ancient DNA samples [87].

Second generation, or next generation sequencing (NGS), began with the invention of pyrosequencing. This method involved observing the light emitted from synthesis of pyrophosphate [88, 89]. Pyrosequencing was further developed by the company 454 Life Sciences, who created the first automated NGS product available commercially [90]. Solexa, later bought by Illumina, created another parallel method for sequencing DNA. This method utilised a technique called "bridge" amplification, called such due to the shape made by the DNA strands in this part of the process. The first technology developed was the Genome Analyser, followed by others such as HiSeq, MiSeq, MiniSeq, NextSeq and NovaSeq, each used for different requirements around run time, reads per run, and read length and depth [91]. Sequencing by oligonucleotide ligation and detection (SOLiD) was another NGS technology, utilising a sequencing by ligation approach as opposed to the sequencing by synthesis method used by most other techniques [92]. Other NGS methods of note include DNA nanoball sequencing [93], Ion Torrent [94] and Heliscope [95, 96] (although this is sometimes categorised as third generation [91]).

There are two main technologies that are established as third generation sequencing methods. The first is single molecule real time (SMRT) sequencing [97]. This method was developed by Pacific Biosciences and works by using a zero-mode waveguide for observing signals from individual nucleotides at a time. This method does not require amplification, is fast, and can process very long reads. The second technology is nanopore sequencing, which uses an electrical field to identify bases [98]. The Oxford Nanopore Technologies' minION sequencer has the advantage of being small and portable meaning real-time sequencing anywhere in the world is now a possibility [99].

There have been some large genetic studies conducted in recent times, with unprecedented sample sizes. Examples of these include the HapMap project [100, 101], the 1000 Genomes Project [102], the African Genome Variation Project (AGVP) [103] and the 100,000 Genomes Project [104]. Biobank projects with the aim of matching genomic information with a wide range of phenotypes include the UK Biobank [105], Qatar Biobank [106], the China Kadoorie Biobank [107], and FinnGen [108]. Commercial projects such as 23andme and Ancestry have amassed large amounts of genealogical data, and have been used for medical research, such as Genentech's Parkinson's disease project using data collected by 23andme [109]. No population-level research could be undertaken without initiatives such as these that collect and provide genomic information for large samples of individuals.

This section has briefly covered ways of attaining genetic data. The next few sections will go into more detail about how the genomic variation identified using these methods can result in selection.

1.5 Genomic variation

Nearly all the diversity observable between species, and between individuals within species, is due to the differences in their genomes. By interrogating the genome researchers can attain valuable knowledge on how and why these differences occurred. This section will discuss how difference types of genomic variation can influence phenotypes.

Deoxyribonucleic acid (DNA) consists of strings of nucleotide bases: adenine, cytosine, guanine and thymine, commonly abbreviated to their initial letters ACGT, as first discovered by Kossel who received the Nobel prize for this work in 1910 [110]. When one of these nucleotides is erroneously replaced by another during DNA replication, it is called a base substitution or point mutation. Some substitutions are more common than others due to the structure of the nucleotide, and these are called transitions. Adenine being substituted by a guanine and vice versa, and cytosine substituted by a thymine and vice versa, are all transitions. All other

substitutions are known as transversions [111]. In populations a point mutation is called a single nucleotide polymorphism (SNP) once it reaches a substantial concentration, for example occurring in around 1% of the individuals in the population.

Genes are regions of the genome that are functional. They are made up of multiple parts, including exons, introns, enhancers, and promotors. Exons are the part of a gene that can be transcribed into RNA, which is explained further in the next paragraph, and introns are removed during RNA splicing. Enhancers and promotors control and regulate how the gene is expressed. The human genome consists of over 20,000 protein coding genes [112].

DNA strings in exons are transcribed into ribonucleic acid (RNA), which is then translated into amino acid chains. This is known as the central dogma of molecular biology [113]. RNA is translated in sets of three bases per amino acid, known as a codon, as first established in 1961 [114, 115]. For example, the DNA bases TCG code for the amino acid serine. Sometimes a base mutation will result in a codon which codes for the same amino acid, for example the codon TCT still codes for serine. These kinds of substitutions are known as synonymous mutations, and therefore are usually considered to be neutral mutations. Base mutations that result in a change in the amino acid are called non-synonymous mutations, for example the codon ACG would now code for threonine. Some codons are signals for the translation mechanisms to stop translating; in DNA these are TAA, TAG and TGA. If a point mutation were to change a codon such as TCG into the stop-codon TAG, this would cause the amino acid chain to terminate prematurely and is known as stop-gain mutation. In general, synonymous mutations have little to no effect on the individual, whereas non-synonymous and stop-gain mutations can potentially have a considerable effect and can be terminal.

Insertions and deletions are when one or more bases are inserted or removed from the DNA string. This could have the effect of inserting or removing multiple amino acids, but also, in the case of the change not being a multiple of three bases, the whole amino acid chain from that point forward could be changed. This is known as a frame-shift mutation. A copy number variant (CNV) is where the number of copies of the section of DNA is different to the reference, due to deletions, insertions, or duplications. All these variants can have a large effect on the individual.

Other structural changes which could occur in DNA are translocations, where a part of a chromosome breaks off and joins to a different chromosome, and inversions, where a region of a chromosome is reversed. Transposable elements (TE) are regions of DNA which can move to other locations on the genome [116]. A TE was responsible for the black colouring in the peppered moth (*Biston betularia*), which is a classic example of natural selection [117].

Any kind of point mutation or structural change could have phenotypic effects on the individual, whether beneficial or deleterious. While most of the methods discussed in the next chapter use SNP data to interrogate the genome for evidence of natural selection, the role of the other types of variation should not be undervalued and may even be more important when considering evolution and speciation [50, 118, 119]. Most SNPs have two variations, or alleles, and are thus called biallelic. While triallelic SNPs do exist in human populations, they are rare and most analysis is limited to biallelic SNPs, so they are often discarded [120, 121]. For this thesis, when a variant at a locus is referred to, it can be assumed that the variant at the base pair (bp) location given is a biallelic SNP. The most common allele in the population is known as the major allele, and the other as the minor allele. A common way of quantifying this is to use the minor allele frequency (MAF), which is a proportion defined as the count of the minor allele present in the population divided by twice the number of individuals.

In a diploid species, every individual has two copies of each chromosome in each cell, one from each parent. Variants found in an individual's genome will either have been passed down from one or both parents, or will have spontaneously arisen, known as a *de novo* mutation. Throughout an individual's life their cells will replicate, and new mutations will appear. These mutations are known as somatic mutations and are not heritable, as opposed to germline mutations which can be passed down to the next generation. Germline mutations are the focus of this thesis as they are heritable.

While this section has covered some of the basic types of variation found in the genome, there are many other types of genetic variation that can have considerable phenotypic effects, such as aneuploidy and epigenetics. The aim was to illustrate that genomic variation is complicated, messy, and challenging to work with. Not even basic variation like point mutations are simple and equal, as one base substitution could have large consequences, whereas another could be completely neutral. To simplify the problem, often models will be built that ignore some kinds of variation and only consider a subset other types, such as SNPs in a population. However, it should always be kept in mind that other types of variation exist and thus the results from these models may be limited. There is still much work to be done in understanding the relationship between genomic and phenotypic variation.

1.6 Recombination

Recombination describes the process during which chromosomes crossover and swap DNA, potentially creating new combinations of alleles. This phenomenon was first observed by Creighton and McClintock in 1931 [122] while they were studying maize plants (*Zea mays*). They

crossed plants with chromosomes possessing visible physical features and saw that the next generation contained plants with chromosomes containing new arrangements of the features that could only have occurred via a crossover event. Before this crossovers had only been theorised, most notably by Janssens (translated and republished in 2012 [123]) and Morgan [124], after whom the unit of genetic linkage, centimorgans, was named [125].

While most cells in a diploid individual will contain both sets of chromosomes, the gametes (eggs and sperm) will contain only one set of chromosomes. These cells are therefore haploid, meaning they contain half the usual number of chromosomes. A haplotype refers to sets of genes or variants that are inherited together along a single chromosome. Recombination is the process in which chromosomes crossover during gamete formation [122], breaking up segments of chromosomes and shortening haplotypes. While recombination can occur during mitosis, this thesis focusses on recombination in meiosis.

Figure 1-1 shows a basic diagram illustrating recombination. The left cell is a diploid cell with a single pair of chromosomes. To the right are four potential haploid gamete cells for the same individual, two of which have experienced a single recombination event, and two which have recombined twice. This figure demonstrates how haplotypes can be broken up by recombination. Alleles that would have been inherited together are now not, and alleles that previously resided on separate haplotypes are now joined.

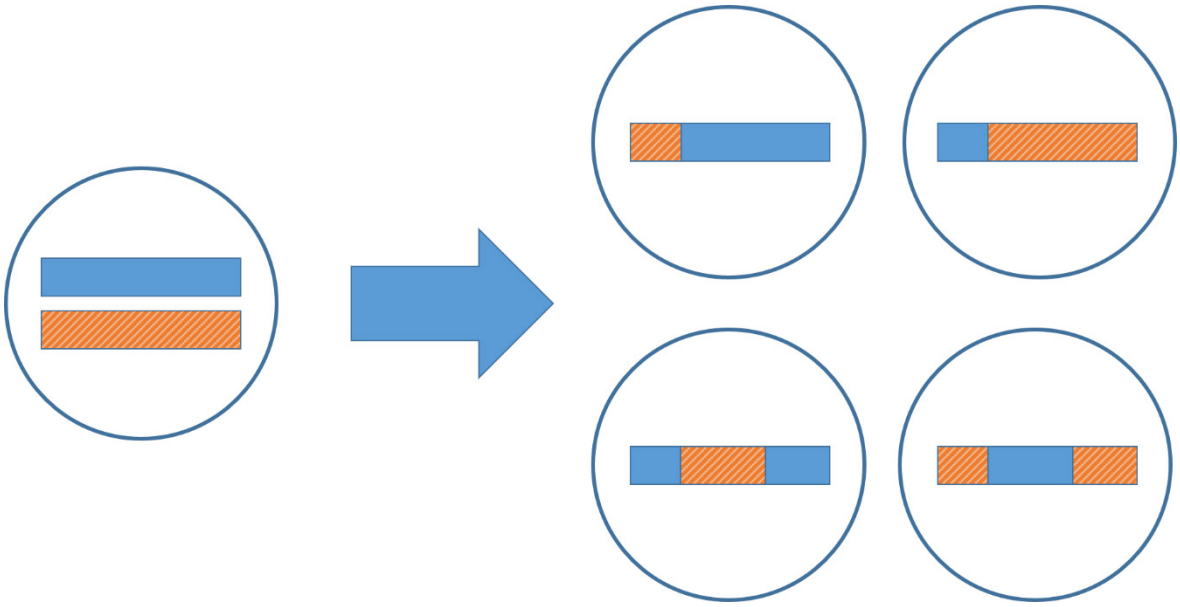


Figure 1-1 Recombination example

This figure shows a very simple example of recombination. The cell on the left has one pair of chromosomes, illustrated in solid blue and striped orange. After meiosis and recombination has occurred, four potential gametes have been formed. Recombination events have crossed-over the two original chromosomes, in the top two examples one recombination event occurred, and in the bottom two, there were two recombination events.

The rearrangement of haplotypes is one of the main evolutionary benefits of recombination. In a population without sexual reproduction and recombination, two beneficial alleles residing on separate haplotypes could never be inherited together, and thus the two alleles would end up competing against each other. This is called clonal interference and it is especially true in asexually reproducing populations [126]. The Hill-Robertson effect describes how recombination allows these two alleles to become linked on one haplotype, expediating the evolutionary process for the population [127]. Conversely, Muller's ratchet describes how, in the absence of recombination, negative mutations would build up on a haplotype [128, 129].

Recombination rates are not uniform across the genome: hotspots can be observed where recombination is much more likely to occur, and cold regions where little recombination happens [130]. Recombination is also correlated with distance from the centromere, with little recombination taking place nearby, due to the risk of improper separation of chromatids during meiosis [131, 132]. Recombination between loci is measured in centimorgans (cM), where a distance of one cM can be interpreted as the expectation of 0.01 crossovers per generation between the two loci. Measuring distance between loci in this way can be more useful than the physical distance, as it allows genes and other genetic markers to be grouped by how linked they

are in terms of coinciding inheritance. The first linkage map was created in 1913 for fruit flies (*Drosophila melanogaster*) by Sturtevant [133].

Recombination maps can be created for populations to estimate the recombination rate across the genome. Linkage maps are based on family data in which recombination events can be directly observed, usually mother-father-child trios [134] with more information coming from families with multiple children [135]. Linkage maps are low resolution and can only show current recombination events, although they do have the benefit of allowing sex-specific analysis. Linkage disequilibrium maps, or LD maps use population genetic data to estimate LD rates which are correlated with linkage maps, for example Myers *et al.* [136]. Using population data means recombination events over many historical generations are included, and the resolution is much higher compared to linkage analysis, with the resolution increasing with the ever-growing availability of whole genome sequencing (WGS) data [137]. However, comparisons with linkage maps show that LD maps may be affected to some degree by processes such as selection, genetic drift and population bottlenecks, confounding recombination estimates [138, 139]. Combining family and population-based maps using joint linkage-linkage disequilibrium analysis could be an effective way to detect recent evolutionary pressure [140, 141]. Genetic maps are discussed further in Chapter 4 where maps created using different methods and software are compared.

This has been a short introduction to recombination to give a basic understanding of how genetic material can crossover during meiosis. For more information see Chapter 4, where recombination and mapping are covered in greater depth.

1.7 Linkage disequilibrium

In an infinitely large population, with no gene flow or migration, random mating, and no selective pressures, it might be expected that variants should be random and independent. That is, the possession of one allele cannot be used to predict the possession of another. However, real-life populations never satisfy all these conditions. Variants are said to be in linkage if the alleles are correlated with each other. Linkage disequilibrium (LD) is a measure of how different the relationship between the alleles are to the equilibrium assumption of being randomly associated [142].

LD can be measured in multiple ways, the most basic being:

$$D_{AB} = p_{AB} - p_A p_B \quad (1.1)$$

p_A is the frequency of allele A at a locus, and p_B is the frequency of allele B at a different locus. p_{AB} is the frequency of haplotypes containing both A and B. Mathematically, if A and B were

independent and randomly distributed, *D* would be close to zero. However, if alleles A and B are linked somehow, *D* will be non-zero [143].

Another common measure of LD is *D'*, which divides *D* by its theoretical maximum, so that LD can be compared across many pairs of loci which may have very different frequencies [144].

$$D' = \frac{D}{D_{max}} \tag{1.2}$$

where

$$D_{max} = \begin{cases} max\{(-p_A p_B, -(1-p_A)(1-p_B)\} & where\ D < 0 \\ min\{p_A(1-p_B), p_B(1-p_A),\} & where\ D > 0 \end{cases} \tag{1.3}$$

The correlation coefficient *r*, usually squared to remove the sign, is another way of defining LD and this is the definition used for most of the following work [145]. It is defined as:

$$r = \frac{D_{AB}}{\sqrt{p_A(1-p_B)p_B(1-p_A)}} \tag{1.4}$$

D' and r^2 are both measures that will return a result between zero and one, with a result nearer one indicating higher disequilibrium. r^2 is a measure of how useful the alleles are for predicting each other. *D'* can be inflated when one allele is rare, and r^2 can also be affected by differences in the allele frequency [146]. For most analyses where r^2 is used, SNPs with small MAFs are removed. There are many other ways of measuring LD between two loci, and also between multiple loci, discussed at length in the literature [147-149].

Linkage between two loci can be affected by many factors. Recombination events can cause a breakdown of LD, resulting in the broad expectation that loci further apart are less likely to be in LD as there is a higher chance recombination will occur between them. Demographic changes such as bottlenecks can affect LD, as well as other population structures, such as those where inbreeding occurs. Selection can also be a cause of LD between loci due to hitchhiking, which will be explained in the section 1.9. Epistasis, where the effect of a genetic variant is conditional on other variants, could also cause loci to be in LD if, for example, possessing a particular combination of alleles is fatal. Genetic drift can also cause LD to increase or decrease due to random chance [143].

Patterns in LD can be exploited for a number of uses. Association mapping is a powerful way of utilising LD information to find regions of the genome that may have some involvement in a particular phenotype [150]. LD patterns can also be interrogated to find evidence for natural selection, as discussed further in section 2.2.2.

1.8 Selection

The ability of populations to survive and adapt to their environments and the process by which they achieve it is known as natural selection. This fitness to survive is at least partly genetically determined. Individuals in a population who have traits that are beneficial to them in terms of survival or reproductive success are more likely to transmit these traits on to the next generation than individuals without. Therefore, the frequency of the trait will increase throughout the whole population over time [151]. This is known as positive selection, where the trait selected for has a beneficial effect on fitness compared to individuals in the population without the trait [152]. This could be a trait which has existed in the population for generations that has become beneficial due to some change in the environment, or it could be a new trait caused by a mutation arising spontaneously.

The opposite of this is known as negative, background, or purifying selection. This is where a new mutation or current variant has a negative effect on fitness and thus should eventually be eliminated from the population by selection. While positive selection is the focus of this work, background selection is the more common type of selection [153-155]. One of the main evolutionary benefits of sexual reproduction is the ability to remove negative mutations from a population through recombination [129]. Without recombination, negative mutations would build up on a chromosome, without any mechanism to remove them. This effect is known as Muller's ratchet [128].

Balancing selection is where the heterozygote has the selective advantage. Individuals who either have both copies of the allele in question, or do not have it at all, are disadvantaged in some way. An example of this are humans living in malaria risk areas who are carriers for sickle cell anaemia. As sickle cell is a recessive disease, heterozygotes have the advantage of not having sickle cell (although there are some disadvantages to being a carrier). As well as this, possessing at least one copy of the sickle cell allele gives the individual protection from malaria. Therefore, in this scenario heterozygotes are the fittest in terms of the chance of survival and reproductive success [156].

Given these definitions, it is expected that an allele undergoing positive selection should see an increase in frequency in the population over time, a negatively selected allele should be decreasing in the population, and alleles undergoing balancing selection should persist at a stable frequency – all assuming a constant environment with that selective pressure. However, through random chance it is possible that any of these alleles could be lost completely from the population, or rise to fixation, which is where every individual in the population possesses the allele. The random component in the fluctuation of allele frequency is known as genetic drift.

When searching for signatures of selection, genetic drift can be confused for selection, or lack thereof, and thus is a confounding factor [157].

1.9 Types of sweep

When alleles are positively selected for and rise in frequency in a population over time they are said to sweep through that population. When sweeps occur, they can leave behind patterns and evidence in the genome due to recombination and hitchhiking. Hitchhiking is a process where, as an allele sweeps through a population over generations, other nearby neutral or even slightly deleterious alleles get swept along with it, as they reside on the same haplotype and thus are inherited together. This means the nearby alleles experience an increase in frequency in the population as well [158]. As an allele sweeps through the population, due to the hitchhiking effect, the population is left with regions of the chromosome containing highly correlated alleles. Eventually, recombination will break down these associations. When there are genomic regions with a non-random association between alleles at different loci, the loci are said to be in linkage disequilibrium or LD. One basis for detecting selective sweeps is in finding and identifying regions in a chromosome with this structure. However, this can be confounded by numerous factors, including the variable nature of recombination itself. Detecting sweeps located near to a recombination hotspot is challenging due to the hitchhiking effect being lost quickly due to the shortening of haplotypes caused by the frequent recombination. Conversely, areas with very low recombination rates may resemble a sweep. Therefore, correcting for recombination rates where known would be advantageous [159].

There are multiple types of sweep, depending on the starting point of the sweep or the present status. Hard sweeps are sweeps which are initiated from a spontaneous beneficial mutation arising in an individual which then rapidly rises in frequency throughout the population until it reaches fixation or near fixation [158]. Due to the mutation arising in only one individual, it is likely that, upon reaching fixation in the population, there will be only one or a small number of haplotypes containing the allele.

If the allele did not arise from a single mutation like in a hard sweep, but instead was already present in the population before some external change caused it to be beneficial, it is known as a soft sweep starting from standing variation [160]. As the variant had arisen previously in the population some time before the sweep began, it is likely to be associated with numerous haplotypes due to past recombination and mutation events.

Multiple independent beneficial mutations sweeping simultaneously in a single species, where the mutations are different but have the same effect on fitness, can also be called a soft sweep [161].

This is also known as convergent evolution within a species. When dairy farming emerged around 10 thousand years ago (kya), people who could process lactose through to adulthood, known as lactase persistence, were at an advantage [162]. Different mutations around the *LCT* gene, which produces the enzyme lactase which helps to digest lactose found in dairy products, swept through multiple populations globally, and thus is an example of this [163, 164].

The term soft sweep is therefore an umbrella term incorporating instances of selection where the origin included multiple haplotypes, each with the same phenotypic effect, sweeping through a population until they reached a moderate frequency [165]. Soft sweeps result in a different variation pattern in populations compared to hard sweeps. When applying methods for detecting selection, sometimes the regions of the genome on either side of hard sweeps, the shoulders, can be mistaken for soft sweeps [165, 166]. An example of a hard and soft sweep in a population can be viewed in Figure 1-2.

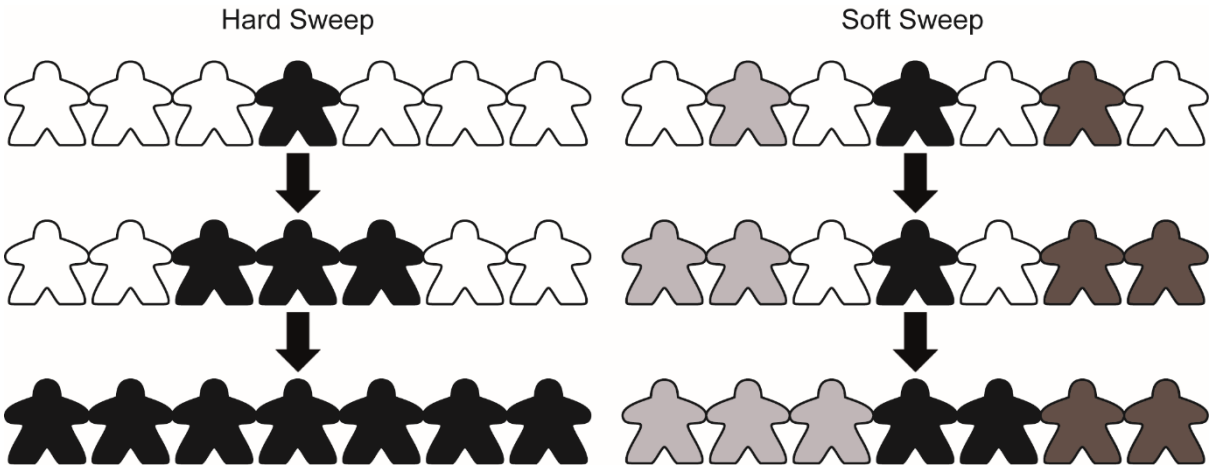


Figure 1-2 Examples of a hard and soft sweep.

The leftmost figure shows a spontaneous mutation arising in one person in the population. After a few generations, the mutation has begun to spread, and eventually everyone in the population possesses the mutation.
The figure on the right shows a population where a mutation has occurred in three different people, either the same mutation on different haplotypes, or three distinct mutations with the same phenotypic effect. The mutations become beneficial, perhaps due to a change in environment. The figure shows the three haplotypes rising in frequency until everyone in the population possesses one.

Sweeps which are still ongoing at the time the population sample was taken are known as incomplete or partial sweeps. These can be either soft sweeps or hard sweeps which have not yet reached fixation. A paper published by Williams and Pennings in 2020 shows examples of

different types of sweeps occurring in the human immunodeficiency virus (HIV) as the virus developed resistance to drugs [167].

Polygenic selection is when there are multiple alleles contributing to a phenotype under selection [168]. Human height is an example of this, as hundreds of loci contribute to height throughout the genome [169, 170]. If a polygenic trait were to be positively selected for, it is highly unlikely that one of the alleles would reach fixation, but instead all the alleles would become slightly more frequent as a group, and thus would be difficult to detect [171, 172].

Identifying and classifying sweeps by these patterns is the aim of the methods discussed in the next chapter. However, this effort is confounded by many factors such as the variable speed of sweeps, multiple origins of alleles and the impact of migration and environmental effects, amongst others [173]. The objective is to develop robust methods to detect true signatures of selection amongst the noise created by these confounding factors.

1.10 The effect of demography

Changes in the demography of a population over time can make detecting sweeps challenging. A rapid increase in population size over a short amount of time results in a population with reduced genomic variation than would be expected for a population of that size. Populations with a sudden reduction in size could present with a larger array of genomic variation than would be expected for a population of that size. Therefore, genomic variation caused by demographic changes could be conflated with patterns caused by selective sweeps. Population bottlenecks, caused by a subset of a larger population separating and developing into a new population, can also look similar to sweeps in the genome [174]. A population bottleneck could also appear to make one type of sweep look like another, for example if the population was undergoing a soft sweep consisting of several advantageous mutations and the individuals in the population surviving the bottleneck only possessed one of the mutations. This would result in the hardening of the original soft sweep [175]. As demographic changes can have such a confounding effect on the detection of sweeps, methods that can take into account demographic changes whilst modelling sweeps are required [176-178].

Instead of using the population size N in models, using the effective population size N_e is often more realistic. The effective population size includes just the population of breeding age, and adjusts N for the violation of the assumptions of random mating and constant population size in many models [179]. As an example, consider a simple population where all individuals satisfy the usual assumptions, apart from having a history of varying population sizes over m generations.

The effective population size for such a population could be found by calculating the harmonic mean, which by nature will bias towards the smaller values [180]:

$$N_e = 1/\left[\frac{1}{m}\left(\frac{1}{N_1}+\frac{1}{N_2}+\cdots+\frac{1}{N_m}\right)\right] \quad (1.5)$$

For example, imagine this population was recorded for m = 3 generations where the size of each generation was N_1 = 500, N_2 = 30 and N_3 = 400. There was a severe bottleneck in generation two, which would have resulted in a massive loss of genetic diversity from the first generation. Generation three could only contain the diversity that was retained through the bottleneck. The arithmetic mean for this example would be 310 but the harmonic mean is around 79. This reflects the genetic diversity that was lost in the bottleneck and illustrates why using the effective population size is more appropriate than the current census size.

The out of Africa (OOA) event, which is estimated to have started around 50,000 years ago when modern humans began to spread into the Eurasian continent, is an example of a population bottleneck [181, 182]. As a result, there is far less variation in the genomes of humans outside of Africa compared to populations within Africa [183]. Non-African populations also have more evidence of selection in their genomes than African populations, due to the new environments encountered when migrating outwards [184]. This means different populations currently living around the world have different estimates for N_e. It is common to use an estimate of N_e = 10,000 for European populations, and N_e = 16,000 has been used for African populations [185-187]. However, Tenesa *et al.* [188] present much lower estimates, although the estimated N_e for their African population was still higher than for other populations, and Charlesworth [189] estimates an N_e of 25,000 for humans as a whole species.

N_e is an important concept in population genetics theory. Two equations that will be referenced later in this thesis that use this concept are:

$$\theta = 4N_e\mu \quad (1.6)$$

where θ is the is the population mutation rate and μ is the population mutation rate per site per generation; and:

$$\rho = 4N_e r \quad (1.7)$$

where ρ is the population recombination rate and r is the recombination rate between adjacent sites per generation. The Watterson estimator is a way of estimating θ using the nucleotide diversity in a population [190]. This concept is used in the Tajima's D method for identifying non-

neutral evolution, as will be discussed in section 2.2.1. The recombination equation will be applied in Chapter 4.

In other species, domestication has caused changes in the demography of its populations. The assumption of random mating in these species is often false, and inbreeding can also be a factor. Identity by descent (IBD) and identity by state (IBS) are important concepts when considering populations who might have recent shared ancestry [191]. A region of the genome that is identical by state between two individuals simply means that the regions contain matching nucleotide strings. For the region to also be identical by descent, both individuals need to have inherited that region from the same common ancestor. By estimating the proportion of the genome that is IBD between two individuals, the degree of relationship between them can be ascertained.

Runs of homozygosity (ROH) refer to regions of the genome where an individual has inherited IBD segments from both parents. ROH can be used to infer the history of the population, especially when the pedigree is unknown or inaccurate [192]. They can also be indicative of selection, for example in chickens (*Gallus gallus*) where there were ROH found around regions associated with commercially-relevant phenotypes [193]. It is important when studying selection that individuals in the sample are from the same population but are not closely related. This is to avoid bias and misleading relationships between alleles that are only correlated because of the presence of closely related individuals in the sample [194].

1.11 Examples of selection in human history

Recent estimates date the split between human and chimpanzee lineages at around 6.5 to 12.1 million years ago, following a period of gradual speciation and hybridisation [195, 196]. Around 300,000 years ago, anatomically modern humans arose in Africa, eventually leaving and spreading across the entire globe during the out of Africa (OOA) dispersals around 50,000 years ago [181, 197]. The new climates, diets, and environments encountered by humans have led to new adaptations in order to thrive in these new locations. Genomic differences between modern human populations can be identified by scrutinising the DNA to find evidence of this spread across the world [182]. African populations have far more variation than those of populations anywhere else due to the founder effect [183]. As humans spread across the Eurasian continent, they interbred with Neanderthals and Denisovans, the evidence of which can be found in the genomes of modern Eurasian populations but is absent in African populations [198-200]. The genomes of modern human populations can be interrogated for signatures of selection to find evidence of adaptation and explain why and how different populations have different traits.

As humans spread, they also affected the other species around them. One of the more direct effects was through domestication of local plants and animals for food, with the rise of agriculture occurring concurrently across the world around 11,000 years ago [201]. This was an important step in the evolution of society as the excess of food allowed for employment in activities other than food production, such as in politics, bureaucracy, and defence [202]. However, this also came at a cost as living in denser populations and working in close proximity with animals can result in disease in humans [203]. As well as in humans, evidence of selection in plants and animal populations can be found by interrogating the genome for signatures of selection [204]. Finding the genomic mechanisms for desirable traits can be beneficial to optimise breeding schemes and maximise food production and profits. Humans also domesticated animals for reasons other than food, such as for riding, pulling, carrying, herding, guarding, and companionship. The domestication of dogs is discussed in Chapter 8.

The rise of dairy farming saw the initiation of a sweep around the *LCT* gene in humans, specifically in the *MCM6* gene which is an enhancer for *LCT* [162, 205]. This allowed humans to drink milk into adulthood rather than just in childhood like other mammals, in a development known as lactase persistence. Lactase is an enzyme that breaks down lactose, a substance found in milk. Adult individuals without the lactase persistence allele report symptoms such as nausea, abdominal pain, and diarrhoea after consuming milk products. There are many theories as to why the lactase persistence allele underwent strong selection, including: the benefit of additional calories, vitamin D deficiencies and calcium absorption in regions with less sunlight, malaria protection, sexual selection, drinking milk instead of contaminated water, and as a protection against potentially deadly symptoms such as diarrhoea during famine conditions when crops were poor and dairy was the only option for sustenance [206, 207]. The sweep is hypothesised to have started shortly after agriculture arose around 10,000 years ago, although there is some evidence that the sweep was initiated more recently [208]. Globally, there are multiple different alleles that convey the lactase persistence advantage. Across Europe and some of Asia, there is one variant that has undergone a selective sweep, known as C>T-13910. In East Asia, lactase persistence is rare [209]. In Africa, at least three different alleles have arisen, all on different haplotype backgrounds [163].

Skin pigmentation in modern human populations is complex and controlled by many genes. The *MC1R* gene is highly constrained in Africa, where it is responsible for dark skin pigmentation and offers protection against UV exposure [210]. It is possible this gene underwent a selective sweep once humans developed hairlessness and needed protection from the sun [211]. As humans migrated into the Eurasian continent and further north, constraints on this gene relaxed. European populations with paler skin colours have experienced strong selection in the gene *SLC24A5* that allows for enhanced vitamin D synthesis in UV-low regions, and other genes such as

SLC45A2 (sometimes known as *AIM1*) and *TYR* also play a role [212-215]. In East Asians, in an example of convergent evolution, variants in the gene *OCA2* were selected for instead [216]. Dark skin found in some south Asian populations, for example in India and Papua New Guinea, is due to the ancestral allele in *MC1R* being constrained again by high UV environments as opposed to any new selective event [210].

There are some human adaptations that are very niche to the environment that the population lives in. An example of this is the *EPAS1* mutations found in Tibetans living on the Tibetan Plateau [217]. This is a high-altitude region with low oxygen and so inhabitants of the plateau have adapted to live in this environment without the negative effects of hypoxia. It is believed that the haplotype containing the protective *EPAS1* mutation was a result of introgression from the Denisovan genome [218]. There is evidence of convergent selection for *EPAS1* variants in multiple species of domesticated animals and livestock owned and farmed by Tibetans living on the plateau [219]. Studying adaptations to hypoxic environments could have clinically relevant implications for intensive care unit (ICU) patients with low blood oxygen levels [220].

Malaria is a disease spread by mosquitos in tropical and subtropical regions of the Earth. The disease is responsible for a large proportion of child mortality annually, and thus genes that are protective against malaria are highly selected for in populations residing in these regions [221]. As mentioned in section 1.8, malaria-protective alleles are a classic example of balancing selection, as resistance to malaria is linked to diseases such as sickle cell disease, thalassaemia and glucose-6-phosphate dehydrogenase deficiency among others [222]. A mutation in the *DARC* gene is protective against a specific malaria pathogen, *Plasmodium vivax* [223]. This mutation is very common, if not fixed, in most of Africa, resulting in the *P. vivax* pathogen being rare in these areas [224, 225].

Alcohol is metabolised in two stages: the first transforms ethanol into acetaldehyde via alcohol dehydrogenase enzymes (ADH) that are controlled by the *ADH* gene family, and the second transforms acetaldehyde into acetate by way of acetaldehyde dehydrogenase enzymes (ALDH) from the *ALDH* gene family [226]. Acetaldehyde is toxic, and a build-up can cause negative symptoms such as nausea, dizziness, and flushing. The *ADH1B*2* allele (sometimes called *ADH2*2*) causes an increase in the activity of ADH enzymes, leading to a large build-up of acetaldehyde. This mutation is rare in Europeans (MAF ~3%), but is common in some Asian populations (MAF ~70%), where it underwent a selective sweep estimated to have started around 10,000 years ago with the uptake of agriculture [227, 228], although it has also be suggested it could have started later [229]. It is hypothesised that it is protective against alcoholism, due to the unpleasant symptoms experienced by the consumer. The *ADH1B*3* mutation has a similar effect,

and is found in African populations (MAF ~20%) while being almost completely absent from others [228]. Protective mutations in the *ALDH* genes have also been found, which work by slowing down the rate of acetaldehyde to acetate conversion, causing severe symptoms [230]. The *ALDH2*2* allele is very rare everywhere but in some Asian populations, where it has been shown that *ALDH2*2* homozygotes have total protection to alcoholism as consuming alcoholic beverages is intolerable [226, 231]. These findings have proved useful in medicine. Disulfiram is a drug that can be used to treat alcoholism, and works by inhibiting ALDH production, much the same way as the *ALDH2*2* allele [226].

CCR5-Δ32 is an allele that is protective against the human immunodeficiency virus (HIV), which is an ongoing modern pandemic that first appeared in humans in the early to mid-20th century [232, 233]. This is a stop-gain mutation caused by a 32 bp deletion in the *CCR5* gene. The allele has been undergoing a selective sweep amongst Europeans populations for much longer than HIV has been present in humans, originating in north Europe from a single mutation [234]. An initial proposal for the cause was that the sweep started with the Black Death, as the original estimates of the age of the sweep were congruent and this was a wide-spread event that significantly reduced the population of Europeans [235]. However, later models suggest that smallpox outbreaks may have been the trigger for the sweep instead [236]. This is perhaps more convincing for four reasons: the timing fits revised estimates better and, while the plague disappeared from Europe in the 18th century, smallpox was present until the mid-20th century; smallpox spreads from human to human unlike plague; smallpox affects children, and thus would exert higher selective pressure than a disease that targets adults who have already produced offspring; and the biological mechanisms of poxviruses are similar to HIV. Medical interventions involving *CCR5-Δ32* protection have included a successful bone marrow transplant from a *CCR5-Δ32* homozygote donor [237, 238], and gene therapy [239]. Controversially, in 2018 a scientist attempted to use CRISPR/Cas9 gene editing to introduce this mutation into human embryos with limited success, although the resulting children appear to be healthy [240].

Susceptibility to infectious disease is a heritable trait [241]. For example, there is some evidence that possessing the ancestral allele of the *IFITM3* gene can increase the chance of death by H1N1 influenza, with a selective sweep found around this region for an allele conveying less susceptibility [242]. There is evidence of familial susceptibility during the 1918 Spanish influenza pandemic, a significant pandemic of modern times, with recent estimates suggesting a final mortality figure of 17 million people globally [243, 244]. At the time of writing, the COVID-19 pandemic has been the cause of death for over 3 million people worldwide according to the Johns Hopkins Center for Systems Science and Engineering [245]. There is some evidence that some people are particularly susceptible to severe forms of the disease due to a variant inherited from

Neanderthal ancestors, which is common in some south Asian populations and rare in Africa due to the location and timing of the admixture [246]. The virus appears to mostly affect older people, and thus this event will have a limited effect on future generations in terms of genomic signatures [247]. There is evidence of ancient selective sweeps in some East Asian populations for variants protective against coronaviruses, potentially from 25,000 years ago [248].

This section has illustrated just some of the adaptations that modern humans have acquired in recent times. Identifying the regions of the genome responsible for selected phenotypes requires mathematical methods. In the next chapter, methods that can be used for identifying selection in humans and other species are detailed and discussed.

Chapter 2 Methods for identifying selection

2.1 Introduction

Identifying regions of the genome which have been subject to selection in populations is important for understanding the history and function of those regions. To achieve this, genome-wide selection scans (GWSS) can be performed using various methods to identify regions of the genome containing signals of selection. For selection scans to be effective, three things were required. Firstly, the technology to sequence DNA at a high enough resolution. Secondly, the computing power to process the data. Lastly, a large number of samples from individuals who are representative of the population in question. These things coalesced around the turn of the millennium when genome-wide selection studies began to take off [159].

In 1973, Lewontin and Krakauer published a paper discussing selection and how it can be identified in the genome by comparing the distributions of gene frequencies between populations [249]. This paper has been credited with inspiring the selection scans of today [250]. In 2016, Haasl and Payseur [159] conducted a review of GWSS published since 1999, resulting in a list of over 100 GWSS executed over multiple species. The majority of the GWSS in the review used cross-population methods to identify candidate regions for selection. These studies provide valuable insights into past selection and current variation in phenotypes and the relationships between them.

For researchers studying selection today, there are many methods to choose from, as will be shown in the next section. Selection leaves patterns in the genome that can be identified in a multitude of ways, inspiring the creation of methods that focus on different signals [251]. Different types of sweep, for example hard or soft, recent or ancient, in progress or having reached fixation, will all be detected differently. The presence of a data set belonging to an appropriate outgroup to compare to will also factor in to whether a method is suitable or not. For all these reasons an abundance of methods have been developed, and continue to be developed, to increase the accuracy of sweep detection.

After identifying candidate regions, discovering the causal variant and the associated phenotypic trait under selective pressure is not trivial. This can be achieved by locating a variant in the region that affects function, through functional studies and detailed annotation, or through association studies, possible only where there is phenotypic heterogeneity in the population [252]. Genome-wide association studies (GWAS) have been used to show the evolutionary benefit of regions under selection and the particular beneficial variants contained within them [253]. Methods such

as iSAFE (integrated selection of allele favoured by evolution) have been developed to try to pinpoint the causal variants in candidate regions [254].

Simultaneous rises in the processing power of computers and in the quantity of available genomic data mean there is opportunity for greater understanding of selection through analysis than ever before. Necessarily, new methods and statistics have been developed to interrogate these vast datasets to achieve the fullest picture of selection processes and the structure and function of genomic regions under selective pressure. Machine learning has been an important development for the analysis of large and complex datasets and the application of such models has led to the extraction of meaningful results in many fields [255].

Many methods have been developed to interrogate the genome for signatures of selection. In 2006, Sabeti *et al.* published a review of methods that were in use at the time and common results and findings from these studies [252]. Seven years later, Vitti *et al.* performed a comprehensive review of methods available at the time of publishing [256]. This chapter is focussed on methods published subsequently to the latter review and is a comparison of the methodology and relative ability of the methods to detect signals over different scenarios. Firstly, some classic methods are presented before thirty recently published methods are reviewed in sections 2.2.2 to 2.2.8. These methods are summarised in Table 2-1.

Table 2-1 An overview of the thirty recently published methods.

This table gives an overview of methods that were published after the review paper by Vitti *et al.* [256].

Name	Method	Input	Output	Software/web link	Key findings / Author benchmarking
Z_α [257]	Linkage Disequilibrium	Phased SNP data and genetic map	Sweeps	None See Chapter 7 for new software	Outperforms Kelly's Z_{nS} [258] and ω_{max} [259] and other novel statistics across a range of allele frequencies, with or without recombination rate variation. Better or comparable when considering an Out of Africa model and when starting from standing variation.
H12 [260, 261]	Haplotype homozygosity	Phased SNP data, although see the G12 statistics [262]	Hard and soft sweeps	None	Compared to iHS (integrated Haplotype Score) [263], H12 has increased power for detecting recent soft sweeps. H12 is as good as, or better, at detecting recent hard sweeps.
nS_L (number of Segregating sites by Length) [264]	Haplotype homozygosity	Phased SNP data including whether each allele is ancestral or derived	Sweeps	https://github.com/szpiech/selscan	Outperforms iHS [263], EHH (Extended Haplotype Homozygosity) [265], Tajima's D [266] and Fay and Wu's H [267], over a range of scenarios, improved results comparable only to iHS. Method loses power when the selection coefficient is low, and as the allele frequency nears fixation, especially for soft sweeps.

Name	Method	Input	Output	Software/web link	Key findings / Author benchmarking
SDS (Singleton Density Score) [268]	Haplotype method	Phased SNP data including ancestral/derived status	Recent sweeps	https://github.com/yairf/SDS	Compared to iHS performs better when the selection coefficient is sufficiently strong, and selection began ~100 generations ago or is continuous. iHS performs better when selection stopped 100 generations ago.
X_{MD} (Comparative Haplotype Identity Statistic) [269]	Haplotype-based method	Two populations, phased SNP data	Sweeps (especially soft and partial sweeps)	https://github.com/jeremy-lange/CHI-Statistic	Compared to F_{ST} [270] and XP-EHH [271]. XP-EHH outperformed χ_{MD} where population bottlenecks or migration were operating, but χ_{MD} outperformed XP-EHH in most other scenarios, especially in partial sweeps and soft sweeps, in particular where the effective population size (N_e) was low.
HacDivSel (Haplotype allelic class - Divergent Selection) [272]	Haplotype or outlier method	Two populations, phased or unphased SNP data	Sweeps	http://acraaj.webs.uvigo.es/	$nvdF_{ST}$ (normalised variance difference F_{ST}) was compared to SvdM [273] and EOS (Extreme Outlier Set) to BayeScan [274]. Both new methods were more powerful for detecting sweeps over a range of scenarios.

Name	Method	Input	Output	Software/web link	Key findings / Author benchmarking
TSel (Time to most recent common ancestor Selection) [275]	Pairwise TMRCA and Anomaly Detection	Phased SNP data	Sweeps	http://blogs.cornell.edu/clarklabblog/clark-lab/software/	TSel was compared to four other statistics including nS_L. In each case TSel was better or equivalent. The difference is especially notable where selection intensity is low where other methods perform no better than random. When compared to HKA (Hudson–Kreitman– Aguadé test) [276], TSel was also shown to be good at identifying recent strong balancing selection.
SweepFinder2 [277, 278]	Composite Likelihood Method	Allele frequency, recombination map and B-value map [279]	Sweeps	http://degiorgiogroup.fau.edu/sf2.html	When compared to HKA [276] and other composite likelihood methods such as in the original SweepFinder [280], SweepFinder2 is superior given strong background selection, with other methods returning almost 100% false positive rates.
CLR (Composite-Likelihood Ratio test) [281]	Composite likelihood method	Phased SNP data including ancestral/derived status	Incomplete sweeps	None	Vy's composite likelihood method was compared to iHS, where it outperformed in all scenarios tested, and nS_L, where it was almost always better. Correctly identified many locations of sweeps which were not detected by iHS or nS_L.

Name	Method	Input	Output	Software/web link	Key findings / Author benchmarking
DCMS (De-correlated Composite of Multiple Signals) [282]	Composite Method	Two populations phased SNP data	Sweeps	None	DCMS was compared to CSS and meta-SS [283] as they are all combining methods. DCMS outperforms both over a range of parameters, except where the frequency of the selected allele is low, or the interval distance is very high.
CSS (Composite Selection Signals) [284]	Composite Method	Two populations, phased SNP data. Ancestral/derived status if available	Sweeps	None	No comparison to another model. Application to cattle and sheep data showed clusters of extreme CSS values in candidate regions.
evoNet [285]	Machine learning – deep learning, neural networks	Phased SNP data	Classification: neutral, hard, soft, or balancing, plus estimate of population sizes	https://sourceforge.net/projects/evonet/	No comparison model for the selection part, but when tested on simulations the model is good at specifying the correct class, especially with pre-training.

Name	Method	Input	Output	Software/web link	Key findings / Author benchmarking
S/HIC (Soft/Hard Inference through Classification) [286]	Machine learning - extremely random trees	Phased SNP data, including ancestral/derived status. Subsequently published the diploS/HIC method for unphased data [287]	Classification: hard, hard-linked [near a hard sweep], soft, soft-linked, neutral	https://github.com/kern-lab/shIC	Outperforms seven other methods, distinguishing regions with a sweep from those which are neutral or linked to sweeps.
Hierarchical Boosting [288]	Machine learning - boosting	Phased SNP data	Classification: complete or incomplete sweeps, ancient or recent	http://hsb.upf.edu/	Compared to evolBoosting [289], CMS (Composite of Multiple Signals) [290, 291] and SFselect [292] – had the highest sensitivity under every scenario simulated.
ASMC (Ascertained Sequentially Markovian Coalescent) [293]	TMRCA	Phased SNP data	Recent sweeps	https://github.com/PalamaraLab/ASMC	Compared favourably to iHS and SDS statistics in simulations, especially for recent, strong selection.
Relate [294]	Genealogical tree-based method	Phased SNP data including ancestral/derived status	Sweeps	https://myersgroup.github.io/relate/	Resulted in a higher statistical power when compared to SDS, a tree-based extension to SDS [295], and iHS statistics over a range of selection coefficients.

Name	Method	Input	Output	Software/web link	Key findings / Author benchmarking
SWIF(r) (Sweep Inference Framework (controlling for correlation)) [296]	Machine learning - averaged one-dependence estimator (AODE)	Phased SNP data from two populations	Sweeps	https://github.com/ramachandran-lab/SWIFr	Outperformed SweepFinder and CMS under a range of scenarios, and also evolBoosting and evoNet after being altered to be window-based.
Trendsetter [297]	Machine learning - Multinomial Regression with Trend Filtering	Phased or unphased SNP data	Classifies hard and soft sweeps	http://degiorgiogroup.fau.edu/trendsetter.html	Performs comparably to other classifiers: evolBoosting, S/HIC and diploS/HIC [287]
ImaGene [298]	Machine learning - deep learning	Phased data, ideally deep sequencing data	Classifies sweeps by selection strength	https://github.com/mfumagalli/ImaGene	No comparison to other methods. Performs well in simulations under mis-specified models.
pcadapt [299, 300]	PCA-based	SNP data from multiple populations	Sweeps	https://github.com/bcm-uga/pcadapt	When compared to BayeScan, hapflk[301], outflank [302], and sNMF (sparse Nonnegative Matrix Factorisation) [303], pcadapt shows greater statistical power, especially for models including admixture.

Name	Method	Input	Output	Software/web link	Key findings / Author benchmarking
EigenGWAS [304]	PCA-based	SNP data from multiple populations	Sweeps	https://github.com/gc5k/GEAR/wiki/EigenGWAS	Results were highly correlated with the F_{ST} statistic. Recovered expected hits in humans such as *LCT* in European populations.
ibd-ends [305]	Haplotype IBD method	Phased SNP data	Sweeps	https://github.com/browning-lab/ibd-ends	Finds expected candidate regions when applied to a British sample.
VolcanoFinder [306]	Composite Likelihood Method	Allele frequency data, ideally with ancestral/derived status	Adaptive introgression	http://degiorgiogroup.fau.edu/vf.html	Outperformed SweepFinder2 and BALLET (Balancing selection Likelihood Test) [307], although neither of these methods are designed to find adaptive introgression. Robust to balancing selection.
3P-CLR (Three Population Composite Likelihood Ratio) [308]	Composite Likelihood Method	Allele frequency with ancestral/derived status, recombination map	Sweeps	https://github.com/FerRacimo/3P-CLR	Performed comparably to XP-CLR [309] after adjusting for the fact that XP-CLR only uses two populations.
LSD (Levels of exclusively Shared Differences) [310]	MRCA tree-based method	Local gene trees for each locus, and overall population tree	Sweeps	https://bitbucket.org/plibrado/lsd/src/master/	When compared to PBS (Population Branch Statistic) [217] the method performed similarly or better.

Name	Method	Input	Output	Software/web link	Key findings / Author benchmarking
ABS (Ancestral Branch Statistic) [311]	Tree-based method	Allele frequency data from four populations	Ancestral sweeps	http://degiorgiogroup.fau.edu/abs.html	Performs well when compared with 3P-CLR and performs comparably to LSD.
RAiSD (Raised Accuracy in Sweep Detection) [312]	Composite method	Phased or unphased SNP data	Hard sweeps	https://github.com/alachins/raisd	Performs well when compared to SweepFinder2, SweeD [313] and OmegaPlus [259]. Can pinpoint the location of the swept variant to a greater accuracy.
F_c [314]	SFS and linkage disequilibrium	Phased SNP data	Sweeps	https://sites.google.com/site/sattalab/software	Tested on both neutral and selected simulations and applied to putatively selected regions in the human genome.
D_u [315]	Tree-based method	Phased SNP data	Recent sweeps	https://zenodo.org/record/835226#.YIDJeuhKg2w	Reduced false positive rate when compared to Tajima's D, Fu and Li's D test [316] and Fay and Wu's H.
McSwan (Multiple-collision coalescent Sweep analyser method) [317]	Machine learning - linear discriminant analysis (LDA)	Phased or unphased SNP data	Hard sweeps	https://github.com/sunyatin/McSwan	Performed well when compared to SweeD. Comparisons were made with XP-CLR and Hierarchical Boosting when applied to the *LCT* region.

2.2 Methods

2.2.1 Classic methods

One basic way of identifying regions of the genome that may be under selection is to look at the synonymous and non-synonymous mutation rates [318]. The statistic $\omega = d_N/d_S$ (also known as K_a/K_s) is defined as the number of non-synonymous mutations in a genomic region over an amount of time (d_N) divided by the number of synonymous mutations (d_S) in the same region and time period. The idea here being that if the ratio is greater than 1, there may be some evidence that the non-synonymous changes in this region are beneficial. However, this method can return false negatives if there was also background selection in the same region. The method can also be biased due to some base pair changes (A <-> G or C <-> T) being more likely [111].

The McDonald-Kreitman test [319] builds on this to consider synonymous and non-synonymous changes both within and between species. Firstly, a two by two contingency table is generated, see Table 2-2.

Table 2-2 McDonald-Kreitman Test contingency table

This table shows the contingency table for the McDonald-Kreitman test, where the numbers represent counts of mutations between or within species that are either synonymous or non-synonymous in a region.

	Between (fixed)	**Within (polymorphic)**
Synonymous	D_S	P_S
Non-synonymous	D_N	P_N

The ratios D_N/D_S and P_N/P_S can then be calculated. If there is no selection, the ratios would be expected to be equal. In the presence of positive selection D_N/D_S would be higher than P_N/P_S, and vice versa for negative selection. The McDonald-Kreitman test can be affected by slightly deleterious mutations, recombination, and complex demography [320, 321].

Nucleotide diversity (π) is a concept that can be utilised to compare populations by assessing the count of differences in the nucleotide bases between individuals within a population. It is defined as:

$$\pi = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j \pi_{ij} \qquad (2.1)$$

where x_i and x_j are the frequencies of sequences *i* and *j* respectively, *n* is the number of sequences, and π_{ij} is the number of nucleotide base differences between sequences *i* and *j* over the number of nucleotide bases [322]. Nucleotide diversity is expected to be reduced in populations that have experienced selective pressure [323].

In the 1920s Wright formulated his F-statistics for measuring heterozygosity in subpopulations. One of these, F_{ST}, is useful for comparing genetic differentiation between groups, which can be used to identify selection when applied to regions of the genome [49, 324]. F_{ST} can be estimated using the formula:

$$F_{ST} = 1 - \frac{\pi_w}{\pi_b} \tag{2.2}$$

where π_w is the average pairwise number of base differences within a population and π_b is the average pairwise number of base differences between the populations [270]. F_{ST} requires the sample sizes to be large and similar between populations to avoid bias, otherwise adjustments should be made [325]. F_{ST} is a popular method that is commonly used when there are data from multiple populations available to be compared [159].

Tajima's D is another popular method for identifying selection [266]. This method utilises the Watterson estimator for the population mutation rate θ, which assumes neutrality [190]. Deviations from this could therefore mean that there has been some selective pressure. The basic formula is:

$$Tajima's\ D = \frac{d}{\sqrt{Var(d)}} \tag{2.3}$$

where

$$d = \pi - \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}} \tag{2.4}$$

π is the average number of differences between each pair in the sample, *S* is the number of segregating sites in the sample, and *n* is the sample size. A basic interpretation is that if Tajima's D is negative, that could indicate a selective sweep or a population expansion, and if it is positive then there may be balancing selection or a population contraction. It is hard to determine significance for Tajima's D due to demography, so generally extreme values are reported as sites of interest [256].

The methods described in this section are only a few of the many methods that have been utilised to identify areas of the genome under selective pressure. The review paper by Vitti *et al.* [256]

gives a thorough overview of methods developed since these classic methods, and see also the review of methods by Pavlidis and Alachiotis [326], which focuses on software and tools. The next few sections detail methods that have been developed recently. Note that while the methods have been categorised discretely, in practice many of the methods could fall into more than one category.

2.2.2 LD-based methods

As described in the previous chapter in section 1.7, SNPs are said to be in linkage disequilibrium (LD) when they are non-randomly associated. In the context of selective sweeps, the hitchhiking effect of nearby alleles with the beneficial allele itself causes LD. LD methods exploit this by finding areas of the genome that contain SNPs that are correlated with each other. Jacobs *et al.* [257] published the Z_α method that works by averaging the squared correlations between SNPs on each side of a target SNP, within a given window. LD is expected to be high on either side, but not necessarily between the two sides. This is because recombination events occur independently on each side, creating different patterns of allelic association as the recombination breaks down the LD left behind by the sweep [327], see Figure 2-1. However, this pattern also resembles the pattern found around a recombination hotspot. LD maps made from population data at high resolution could be used to make the distinction [136]. A caveat on this is that although there is a clear relationship between the "true" yet low resolution linkage maps and higher resolution LD maps, the nature of the differences between them are unclear [138].

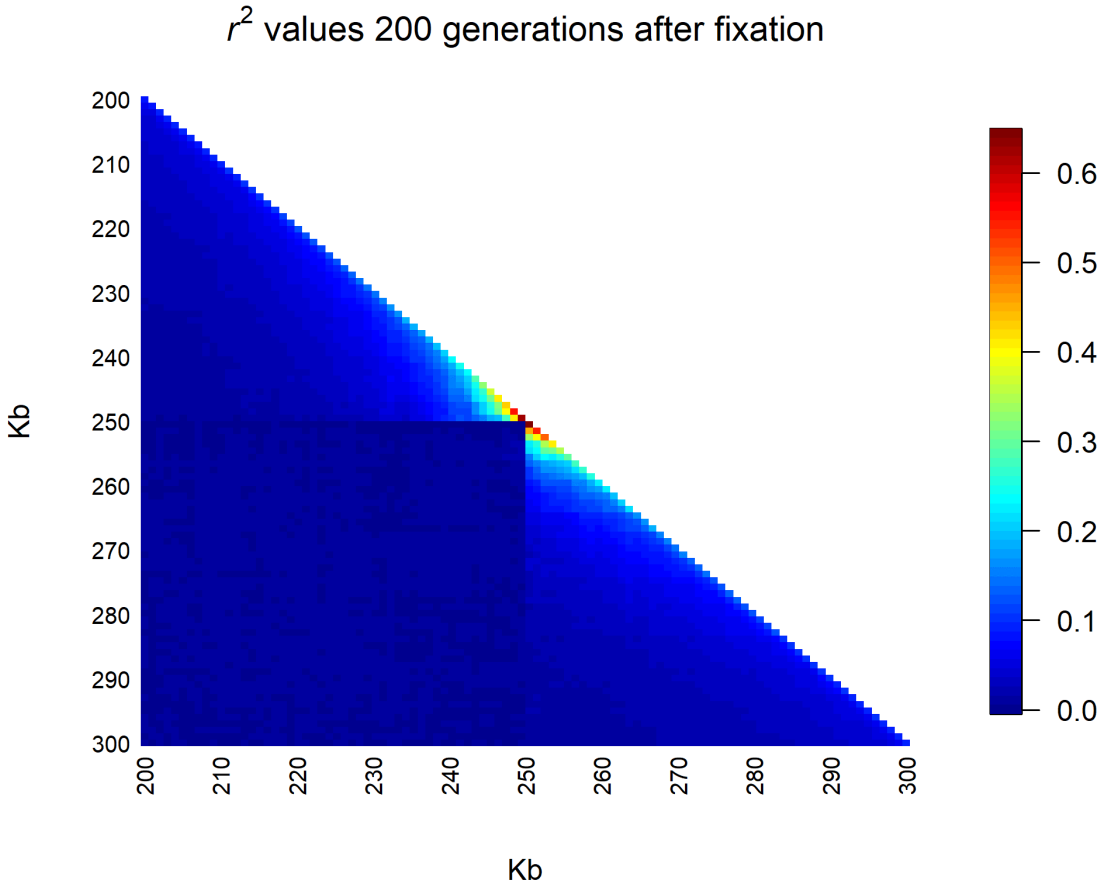


Figure 2-1 r^2 values 200 generations after fixation

This image shows the average squared correlation between SNPs in a simulated dataset. There was a selective event at position 250 kilobases (Kb) that reached fixation 200 generations ago. Regions either side and nearby the site of selection show correlation, whereas regions across the site of selection are not correlated. Details of the simulation software can be found in section 3.2.1 in the next chapter. For this figure, bins of 1000 base pairs (bp) were used to calculate average squared correlations.

2.2.3 Haplotype methods

When hitchhiking occurs, haplotypes become longer than in areas of the genome where no selection has occurred. Haplotypes are also more uniform and so each haplotype will be more frequent. It is these patterns which haplotype methods are designed to identify. Variable recombination rates are confounding factors, as difficulties can be caused by the similarity between regions with less recombination and regions containing sweeps and hitchhiking [209, 328]. Figure 2-2 illustrates the rationale for many haplotype methods.

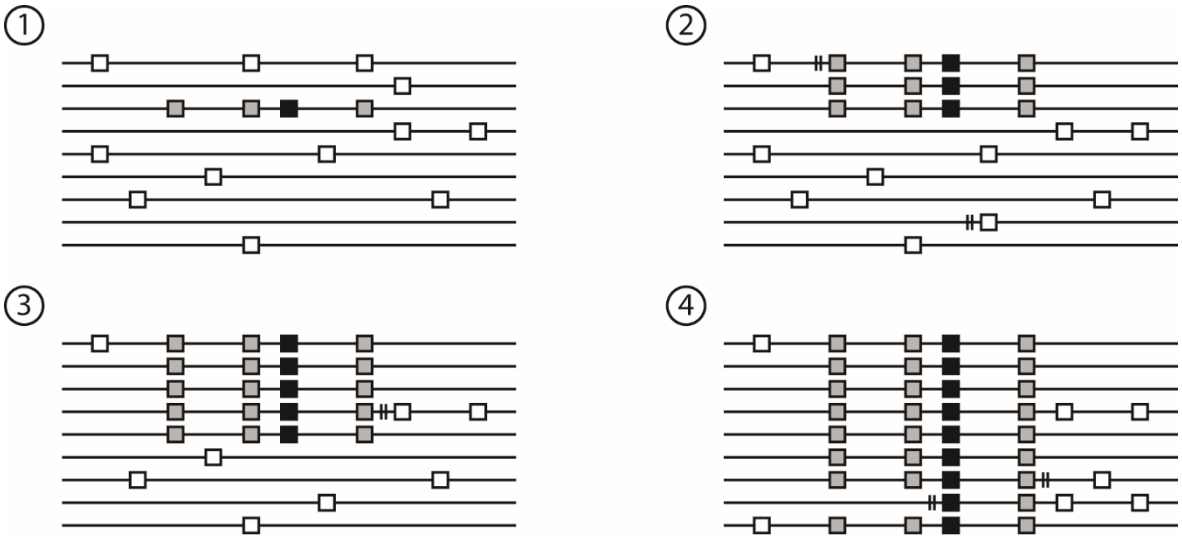


Figure 2-2 An illustration of selection in a population.

This figure shows the impact of selection and hitchhiking in a population consisting of nine individuals with a single chromosome each. Step 1 shows the third individual developing a new favourable mutation (represented by a black square). In step 2, after a few generations, the mutation is starting to spread through the population. Nearby variants (grey squares) are hitchhiking alongside. Recombination events, represented by a double line, are also occurring in the population. Step 3 shows the mutation over halfway towards fixation in the population. Note that, for the individuals possessing the beneficial allele, the singleton mutations (those not found in any other individual) are further away from the selection site than the singleton mutations in the rest of the population. The statistics nSL and SDS are effective for detecting sweeps at this stage, using the rationale that haplotypes are longer when they contain the beneficial allele and shorter when they do not. Finally, step 4 shows the fixation of the beneficial allele in the population. H12 is calculated by considering the squared frequency of the two most common haplotypes. In step 1 where there was no sweep, H12 was equal to 0.14, and by step 4 H12 had increased to 0.48.

The method H12 identifies sweeps by searching the genome to find areas with high haplotype homozygosity [260]. The expectation during a hard sweep is that the haplotype containing the beneficial allele will be high in frequency. During a soft sweep, the beneficial allele may be on more than one haplotype, so this method also considers the second most frequent haplotype as well as the most frequent. There is a further statistic, H2/H1, that can be calculated to attempt to distinguish soft sweeps from hard sweeps using this method. While sweeps can cause long haplotypes, regions of low recombination also present as unusually long haplotypes. The authors tested the method using *Drosophila melanogaster* data, avoiding the issue of low recombination rates being confused for selection by removing these areas using recombination maps generated

by Comeron *et al.* [329]. Avoiding the issue of false positives in this way is clearly only viable given a reliable genetic recombination map has been produced for the species being examined. Subsequently to publishing the H12 statistics the authors published the G12 statistics, for use when phased data is unavailable [262].

ibd-ends is a method that defines segments as identical by descent (IBD), and especially focusses on specifying where the ends of each IBD segment is [305]. The method then finds the regions that have the highest proportion of haplotypes that are IBD. This would imply that a variant in the region is beneficial to have been passed down over many generations, and thus is a candidate region for a selective sweep.

nS_L (number of segregating sites by length) is a method designed to detect sweeps using the length of haplotypes, where the expectation is that selection will lead to longer haplotypes due to the hitchhiking effect [264]. The method considers haplotypes containing the derived allele and those containing the ancestral allele separately, counting segregating sites and taking ratios between the two groups of haplotypes. Longer haplotypes in the derived group, assuming selection on the derived allele, implies fewer segregating sites than in the ancestral group. The method identifies hard and soft sweeps by considering the distribution of the counts of segregating sites between pairs of chromosomes. The method was tested and declared robust in the presence of misspecified population parameters and recombination rates.

The singleton density score (SDS), similarly to nS_L, utilises the idea that haplotypes undergoing a selective sweep should be longer than other haplotypes [268]. Singleton mutations are expected to be rarer on haplotypes containing the selected allele. To apply this method, the distance to the nearest singleton mutation on each side of the target locus is measured for each individual in the sample.

Another method developed using the rationale of longer haplotypes as evidence for a selective sweep, is χ_{MD} (comparative haplotype identity; median denominator) [269]. A reference population is required alongside the population of interest for this statistic. The lengths of unusually long identical haplotypes shared between each pair of chromosomes are summed for both populations and compared.

The final haplotype method reviewed here is HacDivSel (haplotype allelic class – divergent selection), and is actually two methods designed to detect selection by comparing two diverging populations, one developed for when phased data is available and the other for when it is not [272]. The first method $nvdF_{ST}$ (normalised variance difference and F_{ST}) combines a haplotype method with the F_{ST} method described by Ferretti *et al.* [330]. This method splits the population

into haplotypes containing the major and minor allele, finds the minor variant frequencies and compares the variances, the idea being that lower variances would be expected in the haplotypes containing the major, selected for, allele due to other major alleles hitchhiking along. The second method, where haplotypes cannot be ascertained, is the extreme outlier set test (EOS), which combines a k-means classification algorithm with F_{ST} to identify outliers of interest.

2.2.4 Time to Most Recent Common Ancestor (TMRCA) and genealogical trees

The TSel (TMRCA selection) method uses pairwise TMRCA distributions to detect selection [275]. This TMRCA method estimates the time that has passed since the most recent common ancestor for each pair of chromosomes at each non-recombining locus. Calculating TMRCA is not trivial, but methods have been developed for estimating it, such as the pairwise sequentially Markovian coalescent (PSMC) [331]. Statistics such as the mean, variance, maximum and median are then calculated from these distributions, creating a vector of factors for each locus. These vectors are then run through an anomaly detection algorithm to find distortions across regions which could indicate a selective sweep.

The ASMC (ascertained sequentially Markovian coalescent) method estimates TMRCA and then utilises the statistic DRC_T (density of recent coalescence within T generations) to find regions that have an unusually high density of SNPs with recent pairwise coalescent estimates [293].

Relate is another tree-based method [294]. This method works by creating the genealogical tree using a hierarchical clustering method to ascertain the order in which the haplotypes in the dataset coalesced. This information is then used to calculate the probability that a mutation in the current population would have reached its current allele frequency given the historical number of lineages it was present in. Smaller probabilities indicate that a sweep may have occurred.

The statistic D_u is determined based on the lengths and number of segregating sites on different branches of an unbalanced tree [315]. This method avoids the problem of misspecification of ancestral and derived variants, but the authors caution against using this method for genome-wide scans due to multiple-testing issues, and instead only using it on small regions.

LSD (levels of exclusively shared differences) is a method that is designed to analyse multiple populations simultaneously [310]. So long as the overall population tree is specified correctly, it can identify regions of the genome that have been selected for across the different populations by utilising allele frequency changes. ABS (ancestral branch statistic) is another method requiring multiple data sources [311]. ABS is designed to find selection that began in the ancestral population to two “sister” populations who are directly descended from the ancestral population.

This method requires exactly four datasets and works by calculating branch lengths based on F_{ST} calculations at each locus.

2.2.5 PCA-based methods

Principal component analysis (PCA) is used to cluster data and reduce the dimensionality of data, which is especially useful in genetics where there may be thousands of SNPs. This is useful as for many genetic analyses, it is important that individuals in a sample are all from the same population and that population stratification is realised [332]. In sections 4.2.1 and 8.2.3, PCA will be used to cluster the individuals in real-life datasets to assure that the samples used going forward in the analysis are all from the same population. This method can be leveraged to identify differences in populations at individual loci, and thus can be used for identifying selection. pcadapt is a method that uses this theory to identify outliers that could be indicators of a selective event [299, 300].

EigenGWAS is another method that uses this concept [304]. Eigenvectors and eigenvalues are a feature of matrix decomposition used in PCA and are used here to identify signals that SNPs could be under selection. Both PCA-based methods described here are useful when the population structure of the individuals in the dataset is unknown or if there is admixture.

2.2.6 Composite Likelihood Methods

The site frequency spectrum (SFS) is the distribution of derived allele counts in a population over a genomic region [333]. Distortions in the SFS can be indicators of a selective sweep, due to hitchhiking bringing either high or low frequencies of derived alleles along with the selected allele, depending on whichever was associated with the beneficial allele. Composite likelihood methods exploit this pattern to detect sweeps. The composite-likelihood ratio (CLR) method published by Vy and Kim is one such method [281], as is Sweepfinder2, which also includes a correction for background selection [277]. Background selection can reduce the amount of genomic diversity in a region, which can look like a positive sweep, so false positive rates should be reduced by making this adjustment. To achieve this, B-value maps need to be created, which provide estimates for background selection across the genome [279].

VolcanoFinder is a method that was built off the back of SweepFinder2 [306]. However, this method was specifically designed to identify adaptive introgression as opposed to classic sweeps. Adaptive introgression can occur when two species or two distantly related populations of the same species reproduce. Regions of the genome that were previously not present in one of the populations can be beneficial and are then subject to selection. This can look different to a traditional hard or soft sweep, as discussed by the authors of VolcanoFinder [306]. This method

has the advantage of only needing the genetic data for the benefitting population. Other methodologies also search for adaptive introgression, such as the range of k-nearest neighbour (kNN) approaches discussed by Pfeifer *et al*. [334], although these rely on having data for multiple populations.

3P-CLR (three population composite likelihood ratio) [308] is another composite likelihood method, based on the XP-CLR (cross-population CLR) method published by Chen *et al.* in 2010 [309]. This method identifies sweeps by comparing two populations, but then also uses a third outgroup population to identify sweeps that began before or after the split between the two populations.

2.2.7 Composite methods

Composite methods combine multiple sweep-detecting statistics to gain the benefits and information from all of them. The DCMS (de-correlated composite of multiple signals) method is an example of this, which works by finding the p-values for eight different statistics (including the original SweepFinder's CLR method), and then combines them whilst weighting for correlations between the different statistics [280, 282]. The CSS (composite selection signals) method ranks three different statistics across all SNPs in a genomic region, and then identifies clusters of SNPs with extreme scores [284].

The μ statistic employed by the RAiSD (raised accuracy in sweep detection) method uses a combination of signals to identify sweeps, including local LD patterns, changes in the SFS and reduced variance [312]. These signals are combined into single statistic, calculated using a sliding window across the genome. Another statistic, F_c, is derived using information about both the SFS and LD [314].

2.2.8 Machine Learning

Supervised machine learning methods are systems that are trained using a training dataset containing true inputs and outputs. During training the parameters of the system are set in such a way that the resulting outputs are as closely matched to the true outputs as possible, while also avoiding over-fitting. When the system is then fully trained it can then go on to process data where the output is unknown. Seven machine learning methods are reviewed here. The first, evoNet, is a method invoking deep learning, a form of machine learning developed from studies in neural networks [285]. This method classifies regions by the type of selection it judges to have taken place, and further reports an estimate of the demographic history of the genome as a whole. The method works by leading data through a network containing layers of nodes where at each node a data transformation of some kind is applied, until the final classifications are decided.

To train this method, datasets with different selection types and demographic histories were simulated, and 345 selection statistics were calculated, including the H12 statistic. During the training, the data transformations at each node are finalised and the weights on the edges between nodes are established.

ImaGene is another deep learning method [298]. It generates images from genomic data and uses a convolutional neural network (CNN) to analyse them. First, training and testing sets are simulated using models representing the population in question. Second, the images are generated from this data using the population or individual for the row, the locus for the columns, and the allele frequency as the colour in the third dimension. These images are then analysed using the CNN to estimate the selection strength at each point. Flagel *et al.* [335] have also used CNNs to infer a range of population genomic parameters including selection, introgression, recombination rate and population size.

S/HIC (soft/hard inference through classification) is a classification method which utilises an extremely random trees machine learning method [286]. It is trained using a simulated dataset. This method works by generating many decision trees that all place a vote at the end of the process for which classification of selective sweep they believe the genomic region in question has undergone. Trees are generated semi-randomly, and use multiple statistics as input data, including the H12 method. At each node in a tree, one of the statistics is examined and the next branch chosen, until a leaf at the end of the tree is reached with the final classification.

SWIF(r) (sweep inference framework (controlling for correlation)) uses an averaged one-dependence estimator (AODE) machine learning method to classify regions of the genome as neutral or adaptive [296, 336]. It does this by calculating probabilities of adaptation in each region separately, using multiple statistics: F_{ST}, XP-EHH (cross-population extended haplotype homozygosity), iHS (integrated haplotype score), and difference in derived allele frequency ΔDAF. As most of these statistics are cross-population statistics, it is necessary to have a comparison dataset for the population in question.

Trendsetter classifies regions as neutral, soft sweeps or hard sweeps [297]. It uses a multinomial regression with a trend-filtering method using multiple statistics: the mean pairwise nucleotide sequence difference π, r^2, the number of haplotypes, H1, H12 and H2/H1. These can be switched out for alternative statistics when the data are unphased.

The McSwan (multiple-collision coalescent sweep analyser) method uses a linear discriminant analysis (LDA) method on the SFS to find evidence of hard sweeps [317]. It allows for multiple

populations to be used, if available, and can accept unphased data. It also estimates the TMRCA for the ancestries possessing the selected variants.

A hierarchical boosting method is used in the final machine learning method reviewed here [288]. Like the others, it is trained using a simulated dataset, and classifies regions by completeness of sweep and age. This method works by calculating multiple summary statistics, and then leading them through successive linear regression functions until a final classification is made.

2.3 Data

Methods by themselves are useless without data to apply them to. While simulation studies can be used to test the theory behind the methods and apply them over a range of scenarios, the ultimate goal is to use them on real datasets for real populations to identify regions of the genome that have experienced or are experiencing selective pressure.

Technology and cost have been the barriers to sequencing, and many of these methods work most reliably when the population sample is large, and the marker density is high. It is well established that the more SNP markers present in a dataset, the ability to analyse linkage disequilibrium for the purposes of disease-causing variant detection or selection analysis increases [137, 337]. Only having access to sparse markers means that regions of the genome with high rates of recombination are not resolved fully. Array-based panels such as the HapMap project are an improvement on this, with phase II density recorded as one SNP per kilobase (Kb) on average [100, 101]. Furthermore, technological advancements led to whole genome sequencing becoming viable, with the first almost complete human genome being sequenced in 2004 [338]. Since then, costs have reduced and projects such as the 100,000 Genomes Project have become feasible [104]. Technologies for achieving this include Illumina sequencing, SOLiD (sequencing by oligonucleotide ligation and detection), SMRT (single molecule real-time) sequencing, and portable Nanopore sequencing technology [91, 339].

Many of the methods require multiple populations to compare to each other, where in practice this might not be possible. Single population statistics have the advantage in this case as they do not require sequencing or acquiring data for more populations. Examples of relevant additional populations include: populations of the same species separated by physical distance, for example human populations from around the globe; comparison of different yet closely related species, for example humans and chimpanzees; or even populations separated by time, using ancient DNA to compare old variation to new variation. While many species have been sequenced, for example the Ensembl database [340] lists over 300 species as of November 2020, this does not mean that population data is available. Most species have not been sequenced and do not have a reference

sequence. The Earth BioGenome Project (EBP) aims to sequence all the eukaryotic species within the next decade, which should eventually lead to more population data becoming publicly available [341]. Ancient DNA is limited by the amount that has been discovered and the quality of the remains [342]. Nevertheless, even with only few samples, an abundance of information has been uncovered about Denisovans and their relationship to modern humans [80, 343, 344].

Another potential issue is that many of the methods, especially haplotype and LD-based methods, work best with phased data. This is genetic data where the haplotypes have been estimated given the genotypes. Most sequencing methods will not provide haplotype data without also having to sequence family members of the individuals in the sample, which would be costly and impractical in many cases. There are many statistical methods and software for inferring haplotypes from genotype data, including but not limited to: the Expectation-Maximisation algorithm [345], PHASE [346], and BEAGLE [347].

2.4 Discussion

The methods presented here cover a diverse range of rationales and approaches to the problem of detecting sweeps. While classic methods were concerned primarily with finding evidence of a sweep, some recent methods go one step further to identify specific types of sweep. Examples of this include: nS_L [264], which was designed to identify incomplete sweeps; TSel [275] and McSwan [317], which seek out hard sweeps; and SDS [268], which targets recent sweeps. Some methods are adaptations of classic methods with the aim of reducing false positive rates, such as in HacDivSel [272].

There are many challenges to applying these methods. Several rely on cross-population analysis and therefore require detailed data from multiple populations, so clearly these methods will not be appropriate for use when the only available data is from a single population. Methods such as VolcanoFinder [306] were deliberately designed to work in the absence of multiple populations. Conversely, some methods were designed for use when data are abundant and it would be prudent to use all the information available, such as 3P-CLR [308] and LSD [310]. Phased genotype data is required for many of the methods, especially those which focus on haplotype analysis, and this is not always available. HacDivSel [272] and Trendsetter [297] address this by offering a second solution given an unphased dataset. Other authors have subsequently released methods to allow for unphased data, such as G12 (instead of H12) [262] and diploS/HIC [287]. Knowledge of the ancestral or derived status of a SNP is sometimes required but is often unknown, with the CSS method's authors proposing an alternative statistic when this is the case [284].

Some methods rely on genetic linkage maps, which present a particular difficulty. Recombination rate variation can be a confounding factor for many methods. Z_α's performance is improved in the presence of a variable recombination rate once there has been an adjustment by a map [257]. Linkage maps are not of high enough resolution to use in this case, although methods to create high resolution maps based on population data, such as LDhat, may generate a good enough approximation [136, 348]. However, these maps may be confounded by factors which linkage maps are not, such as drift, mutation, population bottlenecks and selection effects [138, 328].

One of the most glaring problems with the application of some of the statistics is the inability to state whether a signal is statistically significant or not, due to the lack of neutral distribution to compare it to [349]. Simulations are used by some methods to gain an approximation of the distribution under neutrality; however, some methods just report the most extreme values, for example investigating just the top 1% of results in the TSel analysis [275]. This can be problematic, as there will always be outliers in an empirical distribution and it is not guaranteed that, if a sweep has occurred, it will fall into the extreme outlying set of results [350]. It should be expected to find false positives present in the results when employing outlier methods [351].

Background selection can have a confounding impact on the performance of some methods and is often ignored in neutral demographic models used for comparison with selection. Jenson *et al.* [250] suggest this problem should be addressed by including or adjusting for background selection in models. SweepFinder2 [277] is a method which explicitly corrects for background selection by requiring a B-value map; however, these maps are currently only available for humans [279] and *D. melanogaster* [352], and are themselves subject to confounding factors and can over-correct or under-correct in different areas. In 2020, Schrider [353] published a paper suggesting that background selection may not be as much as a confounder as previously thought and that it does not manifest the same as hitchhiking, meaning most sweep detection methods should still be effective even in the presence of background selection.

Commonly, methods focus only on selection involving SNPs or single mutations. However, other evolutionary important changes can occur, such as transposable element insertions, which is when regions of DNA move location in the genome [354]. nS_L [264] and H12 [260] have both been used to detect such changes, and χ_{MD} [269] has been suggested as also being appropriate to use in this context [355, 356]. Mutations that have reached fixation, and thus are not polymorphic in present day populations, can also be hard to detect for some methods that rely on the ancestral allele to be extant in the population [357].

Polygenic selection is particularly challenging to detect as it does not follow the usual patterns of typical soft and hard sweeps [168]. CSS [284, 358], HacDivSel [272] and Relate [294] are methods

presented here which have shown some promise of detecting polygenic sweeps. There are other methods that have been developed or are in development which specifically look for evidence of polygenic selection [171, 359]. GWAS have been utilised to find evidence of selection of polygenic traits in current populations [360]. SDS was used to find evidence of polygenic selection for height in humans [268]; however, these results were not replicated when a different underlying GWAS was used [361]. A coalescent tree-based extension to SDS was developed specifically to assess the selective history of polygenic scores [295]. Methods have also been developed to identify selection of epistatic traits, which is where alleles affect traits in a non-additive way [362]. Epistasis can affect the results of some methods, as can pleiotropy, which is where an allele is associated with more than one trait and thus may be subject to multiple selective pressures [363, 364].

While out of scope for this thesis, there have also been efforts to find evidence of balancing selection. Classically, Tajima's D or the HKA (Hudson–Kreitman– Aguadé) test could be used for this purpose [307]. TSel and evoNet both showed evidence of being able to detect balancing selection [275, 285]. A few recent methods that have been published explicitly to find evidence for regions under balancing selection include the T1 and T2 statistics [307], the non-central deviation statistic [365] and the β statistic, inspired by SFS methods [366].

Demographic changes are an important confounding factor to consider when assessing evidence for selection. Significant changes in a population can occur rapidly, for example population bottlenecks, which will leave signatures in the genome and can frustrate attempts to identify selection [367]. The founder effect is an example of this, where a small sample of a large population physically migrates to a separate location and then expands from there. The new population will be less genetically diverse than the original population as only the variants in the founders will be present to pass on [368]. Many models assume that populations persist at a stable size or that members of the population always have an equal chance of mating, whereas in reality these assumptions are rarely likely be true. Therefore, testing the robustness of statistics over a range of demographic scenarios is important [369]. Some methods can perform badly under specific demographic scenarios and having an unknown or misspecified demographic history can confound things further. nS_L [264], TSel [275] and S/HIC [286] were tested under scenarios of uncertain demography and are considered robust. evoNet demonstrates that it is possible to detect demographic changes along with sweeps concurrently [285].

The H12 and S/HIC methods have been used to detect soft sweeps in fruit flies and humans respectively [260, 370]. These findings were criticised in depth by Harris *et al.* in 2018, who argued that soft sweeps are in fact rare and that these analyses were lacking in testing of realistic

demographic histories and prone to error [371]. For H12, they suggested it is unable to accurately distinguish hard from soft sweeps using H2/H1, and that the fixed window size could be problematic given recombination effects, which would lead to an overestimation of soft sweeps. This criticism was also levied by Vy *et al.* in their 2017 paper where they used CLR and nSL to find evidence of more hard than soft sweeps in African fruit flies [372]. S/HIC was criticised for having a high false discovery rate (FDR) and for not being robust to uncertain demography as claimed. Both authors have published rebuttals, with Garud *et al*. revisiting their work on fruit flies and the H12 method using complex demographic models and reporting the same results: that the sweeps are true and that they are largely soft [373]. Schrider and Kern debunked the high FDR criticism of their S/HIC analysis but conceded that the model could be confounded by complex demography [374]. However, they argue that all methods can be confounded, and while it is important to report the limitations of models, this does not mean the methods have no use in the identification of swept regions.

Machine learning methods can be challenging to implement. The time and processing power needed to train a model can be considerable and it is becoming necessary to utilise high performance computing. The evoNet deep learning model reportedly took hundreds of hours to train and test, including simulating the data and computing the statistics [285]. Machine learning methods have a stigma attached, that the internal workings are mysterious and are compared to a "black box" where data goes in and results come out with no real understanding of what happened inside. However, the S/HIC [286] and hierarchical boosting [288] methods reviewed here have endeavoured for transparency by reporting the contributions of each statistic and their relative ranks. The flexibility of these methods is an advantage as many different statistics could be used instead of the defaults with little change to the method structure. This is also true of the composite methods. As new statistics are developed, it may be worth revisiting these methods to see if they can be improved upon. Finally, machine learning methods should also be used with caution, as the results can only be as good as the training set used, so care must be taken to supply them with high quality, accurate training sets [375].

Reproducibility is one of the key components to good scientific research, however, irreproducibility remains a problem and selection signal research is no exception [376, 377]. The reasons for this are wide ranging, and include: values for parameters either not being defined or being arbitrarily assigned values with no or little reasoning, software bugs or changes with no documentation, or differing results depending on computer architecture such as floating-point errors [377, 378]. There is also a temptation to create a narrative around results by searching the literature for any reason there may have been a selective advantage as a means of validation [379]. Pavlidis *et al.* [377] investigated candidate genes identified from a neutral simulation and

created plausible accounts for why these genes may have been selected for, to show how "storytelling" is not an appropriate way to verify results.

Various papers recommend using multiple statistics in combination with others, as different methods search for different signals and therefore more signals of selection can be identified more accurately. Vy and Kim [281] show this in their CLR paper by using their method with the nS_L method, demonstrating that while both methods detected a similar number of true signals, only a fraction of the results overlapped between methods. Using complementary statistics and information to detect sweeps is the rationale behind many composite and machine learning techniques, with many papers suggesting that the best way to identify sweeps is to combine methods [380-382]. While agreeing that they can be powerful, Lotterhos *et al.* warn that composite measures should be interpreted with care given the relative strengths and weaknesses of the incorporated statistics [383].

This area of research is exciting and important, as evidenced by the development of numerous and wide-ranging methods, from relatively simple statistics to sophisticated machine learning models. The future looks particularly bright for the development of machine learning models to be applied to an increasing amount of data with more complexity [255]. There is further work to be done in the field to separate signals of selection from the effects of demographic changes, as it is recognised that demography is a confounding factor and most published methods are tested under a range of demographic scenarios for robustness. Also important in future work is the application of these methods to non-human species and other organisms, which will aid in understanding the genome and the selection that has occurred within it. The development of recombination maps for these species will also be important as patterns of recombination are different in different species, for example hotspots in chimpanzees are in different places than in humans, despite having a close evolutionary relationship [187]. It is exciting to see what new knowledge will be created in evolution and natural history as new methods and statistics are developed and applied in the future.

Chapter 3 Application of recent methods

3.1 Introduction

The previous chapter discussed many methods that can, in theory, identify regions under selection in the genome. In this chapter, some of the methods were applied to simulated data to establish that these recently developed methods could indeed detect selection. This served the purpose of testing the methods to see if they could be applied, either from the software or coded from scratch, and increased the understanding of how they worked. Z_α, H12, nS_L and TSel were the four methods which were chosen for additional study, all of which are described in the previous chapter in section 2.2. These methods were chosen as they are all relatively simple methods to implement that only require a dataset from a single population. They all identify different selection signals: Z_α uses LD fluctuations, H12 looks for high frequency haplotypes, nS_L detects unusually long haplotypes, and TSel finds regions with a smaller TMRCA. nS_L and TSel both have their own software, so implementing external software could be compared to statistics coded manually (Z_α and H12). Further, TSel is provided as an R package, allowing comparison of methods both within and outside the R environment.

As a proof of concept, it was decided to further apply Z_α to human genomic data, in addition to the simulated data. The target was the *LCT* gene, which is widely recognised as a gene in the genome having undergone selection in relatively recent human history [162, 163]. The *LCT* gene is associated with lactase persistence, which is the ability to process lactose into adulthood instead of losing this ability after weaning [205]. Around 10 kya, humans started to farm animals, and also around this time, a selective sweep began for lactase persistence [384]. It is hypothesised that being able to drink the milk of dairy animals was beneficial to humans due to the additional calories, or even that the vitamin D found in milk was beneficial, especially in North-Western Europe [207]. Another hypothesis is that during famine conditions, those who could tolerate milk products were more likely to survive than those who were already malnourished and then also suffering from gastrointestinal problems such as diarrhoea after consuming milk products [206].

The aim of this piece of work was to see if the four methods chosen here could distinguish between neutral and selected regions of the genome, and furthermore, to see if Z_α would identify signals around the site of a mutation affecting the *LCT* gene.

3.2 Methods

The code for this chapter can be found at https://github.com/chorscroft/PhD-Thesis/tree/main/Chapter_3.

3.2.1 Simulation

When testing methods for selection, it is important to show that the methods can not only detect selection, but do not detect selection where there is none; that is, they should minimise false positives. Therefore, neutral simulations should be generated as well as simulations containing selection so the differences between the two can be compared.

Initially, the simulation software used were ms [385] for the neutral simulation and mbs [386] for selection as used in the nS_L paper [264]; however, they were found to be slow [387]. The software msms [388] contained all the functionality needed for this particular piece of work, including being able to generate both neutral and selected scenarios, and was quicker so this software was used to generate the final simulations.

Each scenario consisted of 100 simulations. Each simulation was comprised of a population of 100 chromosomes, 500 Kb in length. In all scenarios, a mutation rate of 10^{-9} per site and a recombination rate of 10^{-8} between adjacent base pairs per generation was assumed, based on parameters from the methods' papers. For selection scenarios, the selective advantage for the homozygote was set to 0.012 and an additive model was applied with the degree of dominance set to 0.5. The selected site was always in the centre of the region. One neutral scenario was generated, and four selected scenarios: partial sweeps with final allele frequencies set to 0.3, 0.5 and 0.9, and one with a sweep that reached fixation 200 generations ago. The neutral scenario and the scenario with a partial sweep to a frequency of 0.5 were the focus for the rest of the analysis.

3.2.2 Selection methods

The methods Z_α and H12 were implemented using R v3.4.2 [389]. The Z_α statistic was calculated using SNPs on average 1 Kb apart and a window size of 200 Kb. H12 was calculated using a window size of 400 SNPs in intervals of 20 SNPs. Only using some of the data significantly sped up calculations without losing any information as there was only an incremental difference between adjacent SNPs.

nS_L was implemented in its own software, downloaded on April 19th 2018 from the website linked in the original paper [264]. The default maximum window length of 200 SNPs was used, and the

output was normalised by allele frequency bins with the frequency bin increments set at 1%. The normalisation was implemented in R.

TSel was implemented using the R package tsel v0.5 [390]. Initially, TSel was calculated using the alternate diversity features of π, Watterson's θ and Tajima's D as they are easily extracted from the msms software and therefore were quick and easy to test the TSel software with. For final analysis the coalescent trees were extracted from the msms software and then manipulated by the ape R package v4.1 [391]. From this point R was used to calculate the features using the pairwise TMRCA for non-recombining loci; the features calculated were the average, maximum, minimum, median, variance, fraction of pairs equalling the maximum, the first quartile and the third quartile. TSel scores were calculated for every 20th SNP in the simulated dataset for efficiency reasons.

3.2.3 The Wellderly Study

The Wellderly Study was launched in 2007 by researchers at Scripps Translational Science Institute, with the aim of discovering a genetic cause for long, healthy lives [392]. The study is ongoing and has so far recruited over 1,400 people. Participants in the study are aged between 80 and 105 years old and must not have ever developed any chronic medical conditions or diseases such as cancer, stroke, Alzheimer's, Parkinson's disease, Diabetes, or heart attack. The dataset used from the Wellderly study contained over 60 million variants for 597 individuals, using the GRCh37/hg19 build.

The data were first filtered for chromosome 2, with a minimum minor allele frequency (MAF) set to 0.01 and a Hardy-Weinberg Equilibrium p-value threshold set at 0.001, using VCFtools v0.1.15 [393]. This reduced the number of variants to 761,113.

The data were then filtered for the 3 megabase (Mb) area from 135,000,000 to 138,000,000 around the *LCT* gene located at 136,545,410 to 136,594,750 on the reverse strand, reducing the number of variants to 8,371. VCFtools was then used to transform the data into the .tped file format. Finally, SNPs were then selected so that the average distance between them was 2,500 bp. This was executed in R, by randomly removing SNPs and recalculating the average distance until the threshold was met.

Ideally, the data would also be filtered for ancestral background to remove confounding effects caused by including people from different populations, but for these initial investigations this was not completed. Leaving individuals from multiple ancestral background in the dataset could dilute the effectiveness of the Z_α statistic. This is because historically they will have experienced different selection pressures, and so regions of the genome under selection in one population

that would usually have high LD have variation introduced by individuals whose ancestors did not experience that selection pressure. Patterns of recombination are different between ancestral backgrounds, which could also affect the associations between SNPs.

The Wellderly dataset was not phased, meaning the data for an individual had not been split into haplotypes. This meant that for each pair of SNPs where both alleles were present, it was unclear which allele was paired with which. To estimate the haplotypes from the genotypes given, the Expectation-Maximisation (EM) algorithm by Hill [345] was employed. This algorithm estimates haplotype frequencies for alleles at two loci, given the genotype frequencies at both loci. This is achieved by initialising a frequency for one of the combinations of alleles, and then iterating through an algorithm until the frequency is stable. The algorithm was coded using R and applied to each pair of variants in the filtered Wellderly dataset.

Z_α is one of many statistics designed and published in the paper by Jacobs *et al.* [257]. It is calculated over a window size *x* bp where the target locus is at position *l* in the centre of the window. It is defined as:

$$Z_\alpha = \frac{\binom{l}{2}^{-1} \sum_{i,j \epsilon L} r_{i,j}^2 + \binom{S-l}{2}^{-1} \sum_{i,j \epsilon R} r_{i,j}^2}{2} \quad (3.1)$$

where r^2 is the squared correlation between SNPs *i* and *j* [145], *S* is the count of all SNPs in the window being calculated, *L* is the set of SNPs to the left of the target SNP and *R* to the right. For the Wellderly analysis, Z_α was applied using R, and a window size of 200 Kb was used for the calculations.

3.2.4 Graphs

The statistics were calculated for each simulation. The results for each simulation were then averaged for each 1 Kb bin in the simulated chromosomes. The absolute values for nS_L were taken before averaging. Only the centre 200 Kb region of the simulations were considered in order to avoid any edge effects. The midpoint of each simulation was collected, and a Mann-Whitney U test applied to see if there was a significant difference between the neutral and selected simulations [394]. The Mann-Whitney U test was used as no specific distribution could be assumed.

Receiver operating characteristic (ROC) curves were calculated using the pROC R package v.1.11.0 [395]. ROC curves are used to assess the ability of a model to predict results, in terms of sensitivity and specificity. For each statistic, the maximum absolute value in the 200 Kb central region for each simulation was considered, apart from for TSel where the median figure was used.

The area under the curve (AUC) is acceptable to use as a summary measure in this case as the two groups (neutral and selected) contain an equal number of data points. Partial AUC at specificity greater than 95% is also considered as it is important not to return false positives.

3.3 Results

To look at the general shape of the statistics, the results over each simulation were averaged for each 1 Kb bin and plotted. The plots can be seen in Figure 3-1. The midpoint of each simulation was assessed to see if there was a significant difference between the neutral and selected simulations. In all cases the difference was statistically significant (p-value < 2.3e-19). The p-values were still significant after Bonferroni multiple test correction [396]. The plots show that the average values for each statistic all stayed relatively flat across the region for the neutral model, whereas for the selected model they all deviated upwards towards the centre of the plot around the region under selection.

For most of the graphs there was a clear peak in the centre of the region at the site where the selected SNP was located. However, the Z_{α} graph is wider and reduces to neutral levels less rapidly on both sides of the selected site than the other statistics. This could be because the methods used window sizes that were not directly comparable, for example Z_{α} used window sizes in terms of Kb whereas H12 and nS_L were in terms of number of SNPs. Also, this simulation assumed a constant recombination rate, which is unrealistic in humans. If a realistic recombination map had been used, the hotspots encountered on each side of the selection site would have rapidly reduced the associations between SNPs, constraining the largest Z_{α} values to around the site of selection without stretching across the whole region. Because the methods are all very different, it could be useful to use more than one of them when attempting to ascertain the site of selection.

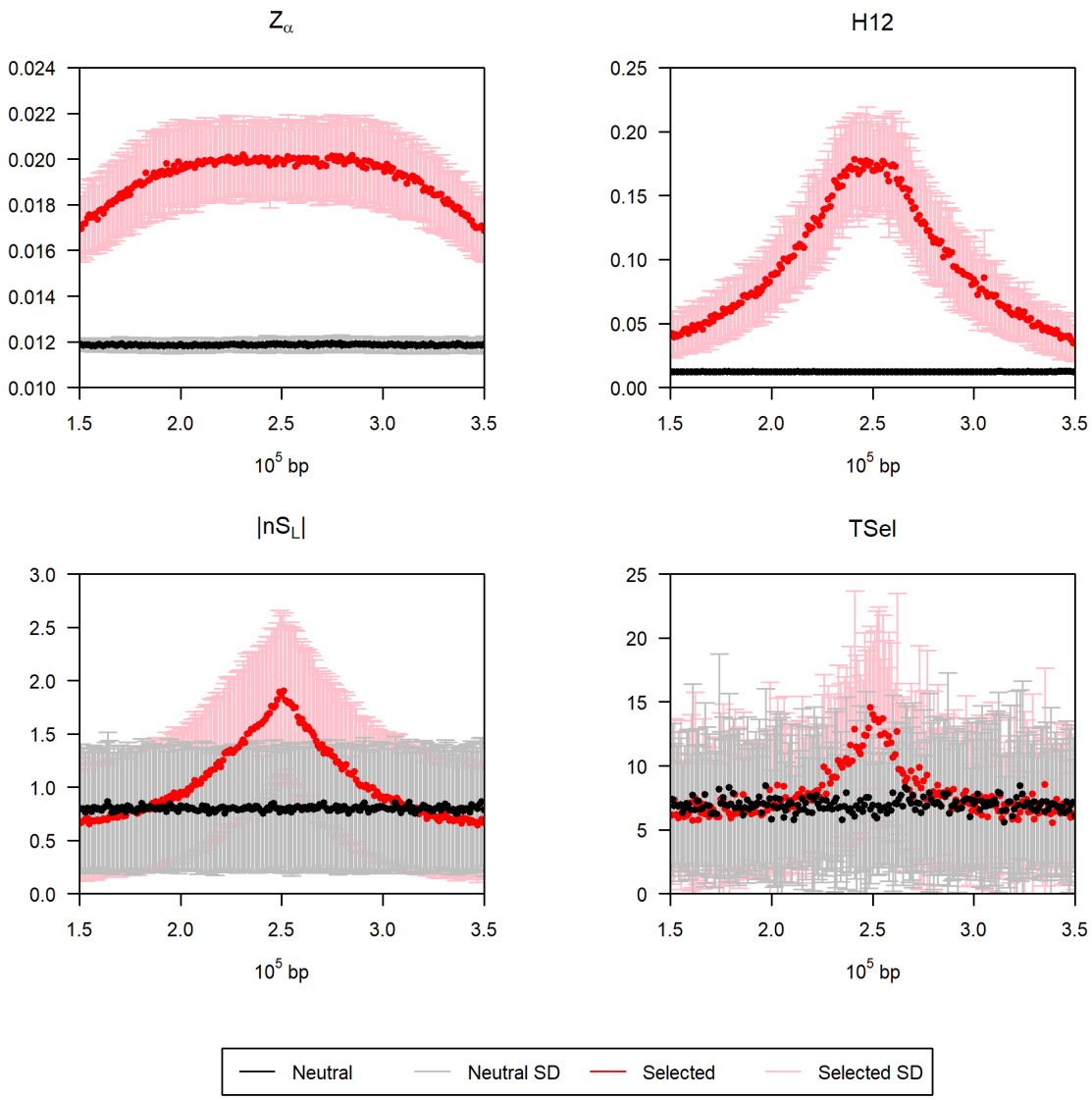


Figure 3-1 The average values of the statistics across simulations for four methods

The plots show the average value of each statistic in 1000 base-pair bins across the simulated region, for both neutral and selected scenarios. For the selected scenarios, the SNP under selection was in the centre of the region. Bars on each plot represent one standard deviation away from the mean in each bin. The value for the SNP closest to the site of selection across each of the 100 simulations containing selection are highly significantly different from the same points in the neutral simulations by Mann-Whitney U test, for each of the four methods.

ROC curves were used to compare the four methods, calculating the area under the curve (AUC) and also emphasising the partial AUC (pAUC) segment where the false positive rate was less than 5% [397, 398]. The central regions from each of the simulations were compared by finding the maximum values (median for TSel) for both models, neutral and selected. Figure 3-2 shows the

ROC curves for each of the methods. Both the Z_α and H12 statistics show a perfect trajectory, being able to distinguish neutral from selected simulations 100% of the time. The nS_L statistic performed less well, with a pAUC of just 9%.

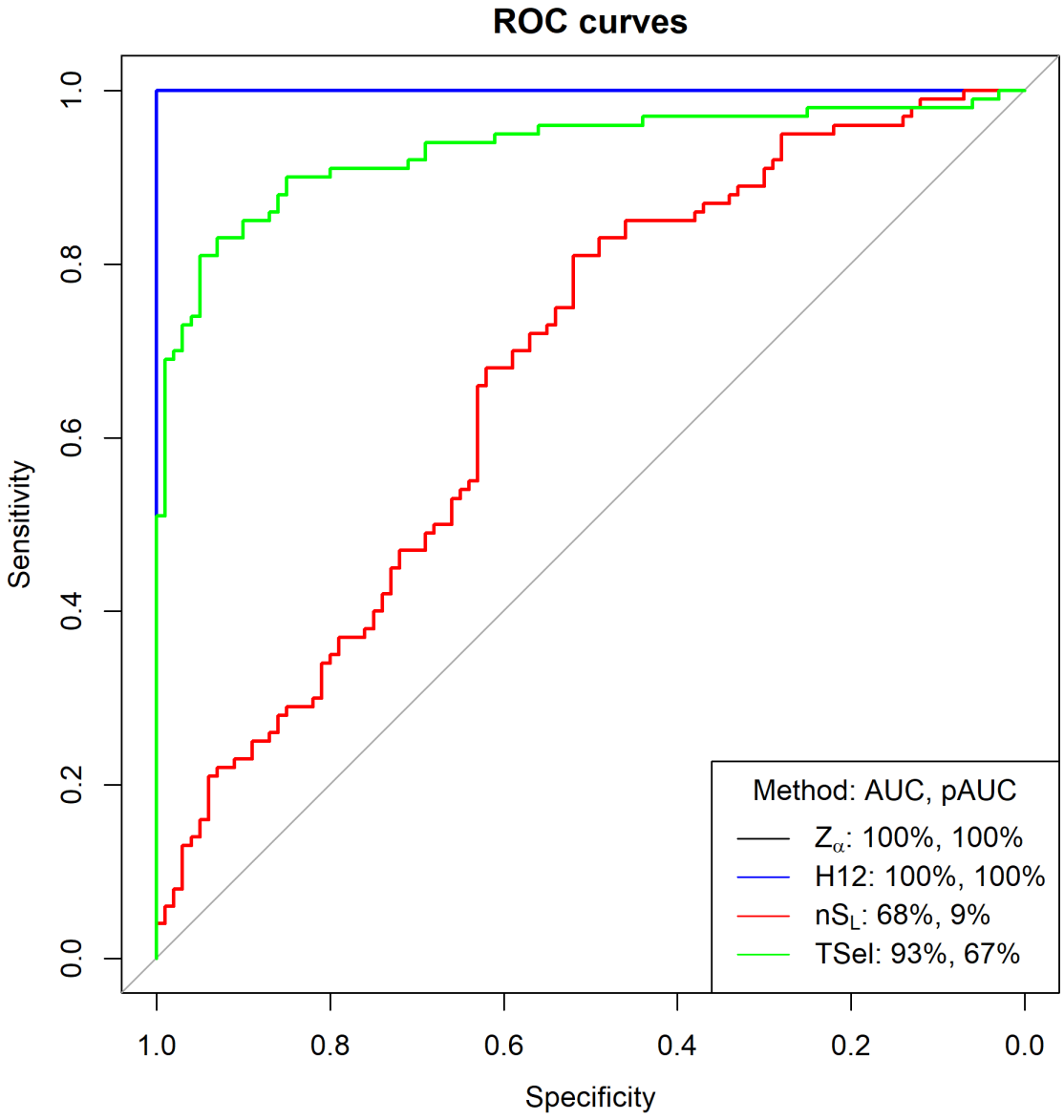


Figure 3-2 ROC curves for the four statistics

Receiver Operating Characteristic (ROC) curves created for each of the four methods using neutral and selected simulations. For each method the area under the curve (AUC) is given, as well as the partial AUC (pAUC) where the specificity is greater than 95% (equivalent to the false positive rate being less than 5%). Note the Z_α curve has been completely covered by the H12 curve as they follow the same trajectory.

Both Z_α and H12 performed very well, and Z_α was chosen as the statistic to develop further due to the lack of software and the innovative method for taking into account recombination, see Chapter 7. The Z_α statistic was then applied to a real-life human dataset for a region on

chromosome 2 as a proof of concept. The final results can be observed in Figure 3-3. While no definitive conclusions can be drawn from this graph, it is interesting and encouraging to see the statistic appear to react to something near the rs4988235 variant that is widely regarded to have been positively selected for in relation to lactase persistence in European populations [162, 163, 205]. The variant is referred to as C>T-13910 in the literature, so this is used here for consistency.

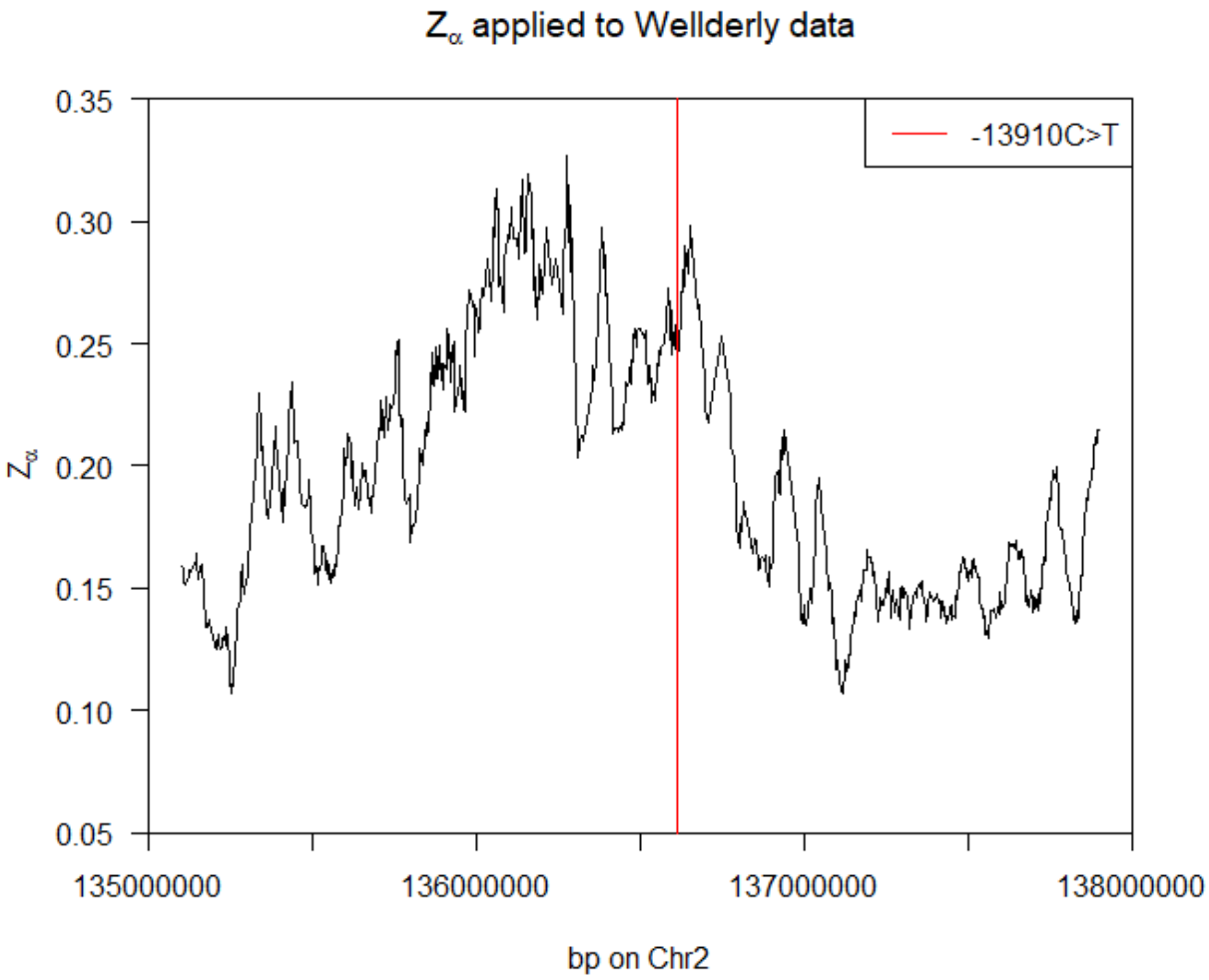


Figure 3-3 Z_α applied to the Wellderly study data centred around the *LCT* gene

The variant C>T-13910 has been highlighted as it is associated with the lactase persistence trait that is widely recognised as having been positively selected for in European populations.

3.4 Discussion

The results showed that the individual statistics were able to distinguish neutral simulations from those with selection. The averaged graphs clearly showed a different shape for each statistic in each selected scenario when compared to the neutral scenario, which was encouraging. The fact

that the differences in the midpoint values for the neutral and selected simulations were statistically significantly different was also encouraging, but may not be that helpful in application to actual population genomic data as in reality it is unknown exactly where the selection took place before applying the methods.

The ROC curves showed that Z_{α} and H12 were the most effective at detecting selection compared to neutral, followed by TSel with nS_L the least effective, for this particular selection scenario. This was unexpected as nS_L was designed to detect partial sweeps as was simulated here. The averaged graph for nS_L shows that the nS_L values change in the presence of a sweep, however, the statistic considers outliers, and the presence of many high absolute values in the neutral simulations impacts the ability of nS_L to discriminate between these and those caused by selection. Using, for example, the median value instead of the maximum value may be more effective for detecting selection, and indeed using the median in this case would have increased the AUC up to 100%. The aim of this simulation study was not to compare the methods to find the best, as many more scenarios (real and simulated) would have to be tested, but merely to test that they do work and to gain and understanding of the complexities of running simulations and applying methods. In any case, to get the best understanding of any signature of selection that might be found, corroboration through the application of many methods should occur.

The further application of Z_{α} to the Wellderly study data was again encouraging. This initial, proof of concept investigation has shown some evidence that Z_{α} can indeed identify regions of the genome under selection, in both simulated and real-world data. More work now needs to be undertaken to investigate the Z_{α} statistic with adjustments for variable recombination rates, and further development of the statistic itself.

Section 2 Recombination and human populations

In this section I look at the role of recombination in the genome. I use human populations to compare how recombination not only varies across the genome but varies within and between human populations across the world. Recombination is intrinsically linked with selection, so it is important to understand it, and how it differs between populations and evolves over time.

I use wavelet analysis to analyse the changes in recombination rates along the genome between human populations. Wavelet analysis involves some complex mathematics, so I have introduced it with a worked example to show how it works and how to interpret the results.

The final chapter in this section is about gene density and effective bottleneck time. This was part of a larger piece of work, where I was asked to create and perform some simulations for a paper comparing gene density and bottleneck times across chromosomes in different human populations [399]. Here, I present the simulations that I ran and the conclusions that were drawn from them.

Chapter 4 Analysis of recombination rate

4.1 Introduction

Selection and recombination are genetic processes that are intrinsically related. Where selection causes genetic variants to become linked, recombination is the only process that can undo that linkage. Therefore, when studying selection, it is imperative to understand the recombination landscape of the population being analysed.

Recombination is a process during meiosis in which two homologous chromosomes crossover to make a new chromosome containing genetic information from both original chromosomes [122]. This occurs specifically during the prophase I phase of meiosis I, with at least one crossover per chromosome per meiosis [400]. There is a difference in the number of recombination events occurring during meiosis in human females and males, with female rates being greater by a factor of around 1.6 [401, 402]. This is due to the fundamental differences in spermatogenesis and oogenesis – the former completing in a matter of days compared to oogenesis, which undergoes meiotic arrest during prophase I for around 12 to 50 years when ovulation restarts the process [403]. Chiasmata, the points at which recombination occur, are vital for stabilising the chromosomes during this time [404].

Sexual reproduction in general is inherently costly in terms of the resources an individual must spend on finding a mate; however, the costs are outweighed by the benefits allowed by recombination [404]. There are two main evolutionary benefits to recombination: the first, is the ability to join multiple beneficial mutations existing on different haplotypes together so that the individual can benefit from them all, and the second, to avoid the build-up of negative mutations, the so-called “Muller’s ratchet” effect [128, 129]. Without recombination, in an asexual species, a slightly deleterious mutation would be subject to genetic drift and may spread through the population with no way to remove it other than another spontaneous mutation in the same genomic region. Over time, multiple negative mutations could build up, eventually leading to the extinction of the species.

Recombination can be measured in two ways: directly using family or sperm analysis, or by using large population datasets to infer past recombination events [139]. There are advantages and disadvantages of each method. Family-based methods need large numbers of participants in order to measure the recombination events at an adequate resolution as there are relatively few recombination events over one generation. Population methods allow for much higher resolution as they contain many recombination events across the history of the population. Population

methods therefore cannot be used for sex specific analysis, whereas family-based analysis can be used to show the difference between male and female recombination rates. Although at low resolution, family-based methods can be sure to measure recombination events exactly, whereas population methods are affected by genetic drift, mutations, selection, and population events such as migration, admixture, and bottlenecks [405]. Maps generated using family-based methods are called linkage maps. An example of this are the Bherer maps, made from an amalgamation of previously published pedigree studies to maximise resolution [406]. LDhat is a piece of software designed to estimate population recombination rates across a genomic region [348]. This is the software that will be used in this chapter. Finally, LDMAP is a piece of software designed to assess LD in a population and creates LD maps [407]. These maps include recombination, like LDhat, but will also map LD created by other processes like selection. Table 4-1 shows a comparison of the main features of linkage maps, LDMAP and LDhat output.

Table 4-1 Feature comparison of linkage maps, LDMAPs and LDhat

Feature	Linkage Map	LDMAP	LDhat
Data	Families or sperm study	Population	Population
Detects	Recombination	LD including recombination, selection, drift, mutation & demography	Recombination
Resolution	Low (few meioses and low density of markers)	High	High
Sex Specific?	Yes	No	No
Time frame	Only current recombination events	Recombination events over many generations	Recombination events over many generations
Units	cM	LDU	ρ (= $4N_e r$)

Recombination events are not distributed uniformly across the genome, instead tending to occur in small 1-2 Kb wide regions known as hotspots [408]. There tends to be less recombination around the centromeric region, with human males especially exhibiting an increased rate around the telomeres [159, 401]. Other correlations have been shown between recombination rates and different genetic occurrences, for example a relation to GC content, the 13-mer sequence CCNCCNTNNCCNC [409], and areas enriched with trimethylation of histone H3 lysine 4 (H3K4me3) [410]. Recombination is also somewhat associated with promoter regions of genes in humans but is suppressed over transcribed regions [101].

PRDM9 is a protein heavily involved in the formation of recombination hotspots in mammals [411]. It contains a zinc finger domain that binds to particular DNA motifs and a SET domain that then trimethylates nearby H3K4. It also trimethylates H3K36, which may be important in recombination alongside H3K4 [412, 413]. The biological mechanisms during meiosis that lead from H3K4 trimethylation to crossovers is complex, however a brief overview follows. Double strand breaks (DSB) in DNA almost always occur at H3K4 trimethylated by PRDM9, as opposed to H3K4 trimethylated by other means, due to PRDM9 binding rearranging the nucleosome in such a way that the DSB machinery is guided to this region [414]. These DSBs are then either repaired using the homologous chromosome via double-Holliday junctions resulting in a crossover, or are repaired without any crossover [415].

In humans, it has been shown that the PRDM9 protein binds to different DNA motifs depending on the PRDM9 variant. Europeans most commonly possess the A and B types, which bind to the 13 bp CCNCCNTNNCCNC motif [416]; however, in West Africa around 36% of the PRDM9 alleles are neither the A nor B types, and have a higher frequency of C type variants which are rare in Europeans [417]. C types have been found to bind to the 16 bp motif CCNCNNTNNNCNTNNC [138], and the 17 bp motif CCCCaGTGAGCGTtgCc (where lowercase indicates a weaker signal) was found in West African populations [418]. West African populations therefore experience recombination at a more uniform rate than Europeans due to the increased opportunity for crossovers.

The gene that encodes the PRMD9 protein, *PRDM9* in humans and *Prdm9* in mice, is one of the fastest evolving genes in the genome and is at least partially responsible for speciation within the Mammalian class [419, 420]. While PRDM9 is important for regulating crossover distribution, a study with *Prdm9* knockout mice showing little change in DNA breaks could imply some independence from PRDM9 [421, 422]. Additionally, the *Canid* genome contains an inactive version of *Prdm9* and thus hotspot distribution must be controlled by some other factor [423]. While loss of *Prdm9* has been shown to cause sterility in mice [422, 424], a woman with no functional *PRDM9* genes still gave birth to healthy children [425].

An interesting phenomenon is the recombination hotspot paradox, where if both the hotspot allele and a non-hotspot allele exist in a population, through recombination and biased gene conversion (BGC), the hotspot allele is replaced by the other, non-hotspot allele [426]. This should mean that hotspots would be eradicated overtime, yet hotspots do still exist in modern populations, and hence the paradox. This can be explained by observing that when hotspots are removed in this way, different motifs become selected for via evolution of PRDM9 [402, 421]. Indeed, when comparing humans with our closest relative the chimpanzees, it has been shown

that we share very few, if any, of our hotspots, implying hotspots do evolve and move quickly [187, 427].

This phenomenon of hotspot extinction with new, younger hotspots reforming was originally introduced by Jeffreys *et al.* [428]. The fluidic nature of hotspots is known as the Red Queen theory, where hotspot motifs are lost to BGC and new motifs are then targeted [429]. However, this model requires an unlimited supply of potential recombination motifs. Úbeda *et al.* [430] propose a more modest requirement, suggesting a death and resurrection model for hotspots. Studying ancient DNA has shown that while the CCNCCNTNNCCNC motif common in modern Europeans was already being targeted before the human and Denisovan split, Denisovans and modern humans did not have overlapping hotspots [429].

It is essential to take into consideration recombination when searching for evidence of a selective sweep. Where selective sweeps will increase the linkage disequilibrium around the site of selection, recombination will achieve the opposite. A sweep located in a region of the genome under intense recombination pressure will therefore manifest differently in the data to a sweep located in a recombination cold spot. Thus, to study sweeps in a population, it is important to first understand the underlying recombination map for that population.

The aim of this chapter is to create recombination maps using LDhat for different human populations, ultimately to compare them. For this piece of work, four different datasets were run through the software: CEU, a population from the 1000 Genomes Project [102]; data from the Wellderly study [392]; and the Baganda and Zulu populations from the African Genome Variation Project (AGVP) [103]. Chromosome 22 was selected for analysis for efficiency. Two European datasets (CEU and Wellderly) and two African datasets (Baganda and Zulu) were selected for analysis to allow for comparison between African and European genomes whilst also being able to compare within the groups to benchmark the expected variation.

4.2 Methods

The code for this chapter can be found at https://github.com/chorscroft/PhD-Thesis/tree/main/Chapter_4.

4.2.1 Wellderly dataset

The Wellderly dataset [392] described in section 3.2.3 was used after the following cleaning. Firstly, the dataset was filtered for poor quality SNPs. SNPs with a minor allele frequency (MAF) of 0.05 or less were removed, as were SNPs with 10% or more of the data missing. SNPs were also removed if the p-value for the exact Hardy-Weinberg Equilibrium test was below 0.001 [431]. This

was achieved using VCFtools (v0.1.13) [393]. The original Wellderly dataset of 61,945,009 SNPs was reduced to 5,181,192 SNPs.

The Wellderly dataset contains people from multiple ancestral backgrounds, and so to make sure that further analysis is carried out on a dataset as homogenous as possible it is important to remove individuals who are very unlike the others. To do this, principal component analysis (PCA) and clustering analysis was applied.

The first step in the process is to prune the dataset of SNPs which are highly correlated as they will not be helpful in determining the differences between individuals. This was done using a pairwise LD-based pruning method implemented in PLINK v1.90beta [432]. The parameters were chosen such that the algorithm would consider a window size of 50 SNPs and prune out any pairs of SNPs with an LD (r^2) value greater than 0.5, and then move the window along 5 SNPs, repeating until reaching the end of the file. After pruning the dataset contained 722,568 SNPs.

Principal component analysis (PCA) is a way of reducing a dataset from having many variables to having only a few dimensions, or principal components (PC). PCA was performed on the data in PLINK, defaulting to 20 principal components, using a variance-standardised relationship matrix [433]. Eigenvalues show how much of the variation is described by each PC. This graph in Figure 4-1 shows the variances explained by each new PC. The idea therefore is to pick the smallest number of PCs whilst giving the most information. Using the elbow method on Figure 4-1 gives four PCs.

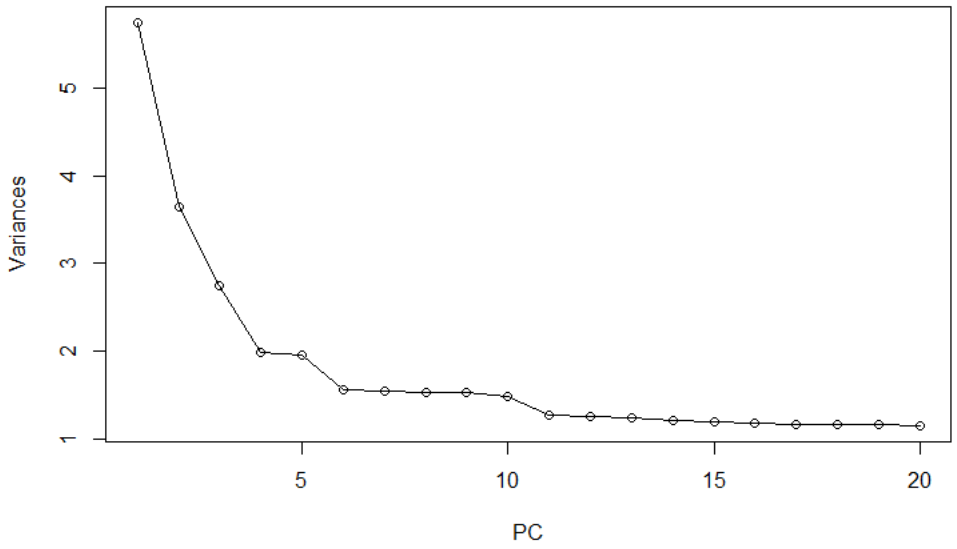


Figure 4-1 Variance explained by each PC in Wellderly data

This figure shows the amount of variance explained by each principal component. The figure is limited to 20 PCs.

k-means clustering is a method for grouping observations into *k* clusters, based on the mean distance to the centre of the cluster. The four PCs calculated above are the input data for each of the individuals. To cluster the data the k-mean clustering algorithm was implemented in R, using the Hartigan-Wong algorithm, found in the stats package (v3.4.1) [434]. The number of starting random sets was increased to 25 and the maximum number of iterations was increased to 1,000 to increase the chance of a stable output between runs. To decide how many clusters to use, the k-means algorithm was run for multiple cluster sizes, calculating the within groups sum of squares and using the elbow method as before to pick the smallest number of clusters without too much loss of information. Plotting the results in Figure 4-2, it is clear to see the elbow is at 5 clusters.

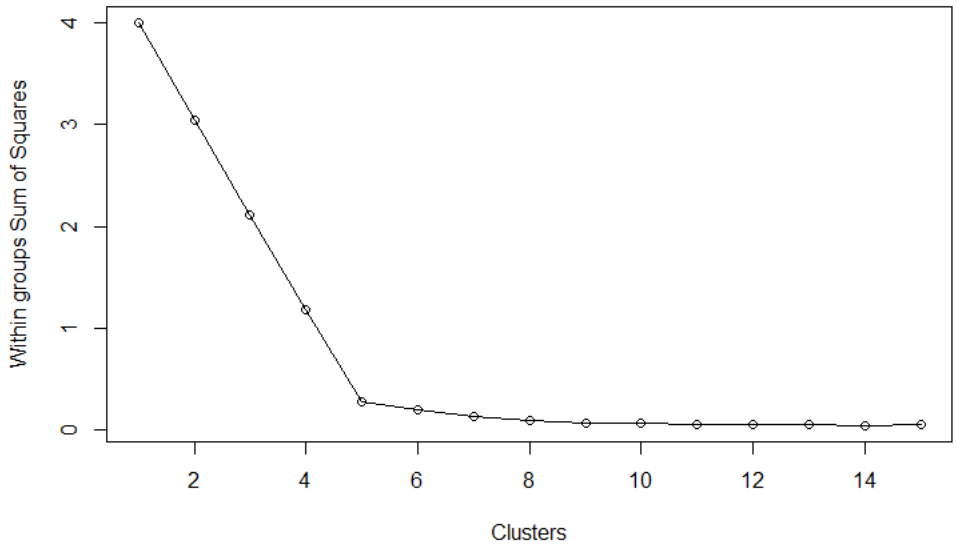


Figure 4-2 Within groups sum of squares for each cluster count

This graph shows the within groups sum of squares for each cluster count for the Wellderly dataset using k-means clustering on the four PCs.

The data were then assigned to five clusters using the k-means algorithm using the first four PCs. The final graph, plotted in two dimensions, is in Figure 4-3.

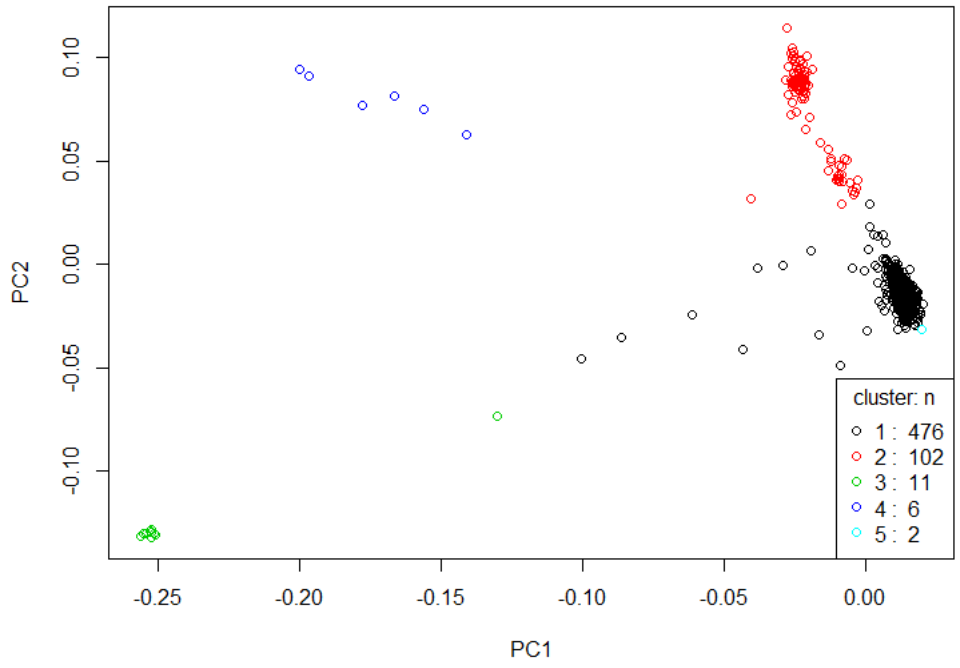


Figure 4-3 PCA of the Wellderly data

The final PCA plot of the Wellderly data, plotted in two dimensions. The largest cluster is considered to be of European ancestry.

The individuals that were kept in the data are the 476 individuals belonging to the largest cluster. The final clean Wellderly dataset contains 476 individuals and 5,181,192 SNPs. These individuals are considered to have European ancestry, based on the self-reported ancestry of the cohort.

For the LDhat analysis, the data were then filtered for just biallelic SNPs in chromosome 22. The dataset contains 476 individuals with 73,166 SNPs.

4.2.2 Other data

The CEU data contains the genotype data for 96 individuals across chromosome 22, consisting of 66,725 SNPs. It refers to Utah residents with northern and western European ancestry from the 1000 Genomes Project [102]. The CEU data were supplied pre-cleaned as described in Pengelly *et al.* [137]. This, and the other three datasets described, use the GRCh37 (hg19) build. For the rest of this document, whenever the term “European datasets” is used it refers to the CEU and Wellderly data.

The Baganda data is from the African Genome Variation Project (AGVP) and is comprised of 100 individuals from Uganda [103]. Before cleaning, the data contained 468,667 variants on chromosome 22, stored in .bed, .bim, and .fam files. The data were converted to VCF (variant call

format) and cleaned using PLINK [432] to remove SNPs where the MAF was less than 5%, missing data were more than 10% or the Hardy-Weinberg Equilibrium test p-value was less than 0.001. After filtering 116,838 SNPs remained.

The Zulu data is also from the AGVP and contains 100 individuals from South Africa. Before cleaning the data contained 468,667 variants. After cleaning as in the Baganda data, there were 121,282 variants. The Zulu and Baganda datasets are collectively referred to as the African datasets.

4.2.3 LDhat

For genotype data, LDhat requires two input files: *sites* and *locs*, where the *sites* file contains the genotype data per sample, and the *locs* file contains the physical location of each of the SNPs recorded in the sites file. All the data were stored in VCF files, which were converted into the LDhat input files using R. The R code also removed any duplicated SNPs, leaving 73,101 SNPs in the Wellderly dataset.

LDhat requires a lookup table containing the coalescent likelihoods for each combination of pairs of loci, across a range of recombination rates. The likelihoods are calculated using an importance sampling method [435]. These tables can be created using the LDhat software given a sample size n and mutation rate θ per site. As it is very computationally costly to create these tables, pre-made tables are available as part of LDhat ranging from n = 50 and θ = 0.5 to n = 192 and θ = 0.001, where n is the number of chromosomes. As the data are diploid unphased genotype data, this lookup table is applicable for a sample size of $n/2$ = 96. θ is the population mutation rate per site, and can be estimated using the formula:

$$\hat{\theta}_w^* = \left(\sum_{i=1}^{n-1} \frac{1}{i}\right)^{-1} \ln\left(\frac{L}{L-S}\right) \qquad (4.1)$$

where L is the length of the sequence, S is the number of segregating sites and n is the number of gene sequences [436]. The closest pre-generated lookup table for the data is the table where n = 192 and θ = 0.001.

The LDhat software contains a program called *convert*. This program was used for the Wellderly, Baganda, and Zulu datasets to randomly select the 96 sequences to use for further analysis, and to create the new *sites* and *locs* files accordingly. The CEU dataset already contained 96 individuals and so did not need to be filtered. This program was also used to remove any SNP locations consisting of more than 10% missing values, resulting in 71,585 SNPs for the Wellderly dataset, 112,115 SNPs in the Baganda data, and 115,777 SNPs in the Zulu data.

The *interval* program is the main program in LDhat. It generates an estimate of recombination rates across the given genomic region using a Bayesian reversible-jump Markov Chain Monte Carlo scheme (rjMCMC) [348, 437]. Briefly, at each step the algorithm considers one of four possible changes to the current state (initialised randomly): changing the rate of a block of SNPs; adding an adjacent SNP into the region; splitting a region into two parts; and merging two adjacent regions. Thus, an rjMCMC scheme is used over a basic MCMC model because of the changing dimensionality of the problem: if a region is split into two parts then the number of dimensions will be increased by one, and an MCMC model will only work under fixed dimensions. The algorithm will accept the change based on a calculated acceptance probability. The outputs from the *convert* program were used as inputs, as well as the pre-calculated look up file with n = 192 and $\theta = 0.001$. The number of iterations was set to 10,000,000, with sampling every 5,000 iterations, and a block penalty of five as advised in the LDhat manual.

The *stat* program extracts the information from the *interval* program's output files. It returns the population estimated recombination rate between each SNP location in the original file. The first 100,000 iterations of the interval algorithm are discarded as recommended in the manual.

LDhat returns the estimated population recombination rate ρ per Kb where ρ was defined in equation (1.7). N_e is the effective population size (see section 1.10 for a full description) and r is the recombination rate between the two adjacent SNPs per generation.

To compare to the LDMAP and the sex-averaged linkage map created by Bherer *et al.* [406], the LDhat output was converted to cM. This was achieved by finding the total map length from the linkage map for the region in question (60.67 cM) and finding the total map length in cumulative ρ, and adjusting every value in the LDhat output by the ratio of these values [438].

4.2.4 LDMAP

LDMAP is a piece of software used to calculate linkage disequilibrium in linkage disequilibrium units (LDU) [439, 440]. It is designed to capture all LD, and not just that attributed to recombination. Therefore, it will be affected by factors such as drift, selection, and bottlenecks.

The calculation of LDU between two SNPs is based on the Malécot equation, defined as follows:

$$\rho = (1 - L)Me^{-\sum \varepsilon_i d_i} + L \qquad (4.2)$$

where ρ is the association, M is the association probability at a distance of zero, L is the association at a large distance, d_i is the physical distance between pairs of SNPs and ε_i represents the breakdown of association between SNPs as d_i increases [147, 441]. The LDU between each pair of adjacent SNPs i is defined as:

$$LDU_i = \varepsilon_i d_i \tag{4.3}$$

One LDU describes the distance in which LD drops to background levels. If SNPs have a short LDU distance between them, it means the LD is high. If the LDU distance is long, it means there is less LD. LDU are additive, meaning that if there are three SNPs in order *i*, *j* and *k* along a chromosome, the LDU between SNPs *i* and *k* is equal to the LDU between SNPs *i* and *j* plus the LDU between SNPs *j* and *k*. $\sum \varepsilon d$ is equivalent to θ*t* where *t* is the number of generations since the founder population and θ is the frequency of recombination [147]. *εd* is used as *d* is known and *ε* can be estimated, whereas *t* is unknown and θ would have to be estimated using a linkage map, which would be of too low a resolution to calculate [442]. Nevertheless, this relationship is useful to state as it can be used to estimate the effective bottleneck time, which will be discussed in Chapter 6.

The association statistic *ρ* is defined specifically in this case as:

$$\rho = \frac{D}{Q(1-R)} \tag{4.4}$$

where *D*, *Q,* and *R,* are defined in Table 4-2.

Table 4-2 Table of haplotype frequencies

The table shows haplotype frequencies for a pair of biallelic SNPs, arranged so that *Q* < 1-*Q*, *Q* < *R*, *Q* < 1-*R* and *ad* > *bc*. *D* is defined as *ad-bc.*

Haplotype Frequencies	**SNP_2 allele 1**	**SNP_2 allele 2**	**Frequency of SNP_1 alleles**
SNP_1 allele 1	a	b	**Q**
SNP_1 allele 2	c	d	**1-Q**
Frequency of SNP_2 alleles	**R**	**1-R**	

The LDU map is constructed via an iterative process. Firstly, an initial estimate of the parameters is completed. Then, an algorithm is run where ε_i is estimated and adjusted at each step until a convergence threshold has been met [439].

The final LD map is the distance between each SNP in LDU, presented cumulatively along a chromosome. The LD map for CEU presented in this chapter was taken from Pengelly *et al.* [137].

4.3 Results

4.3.1 Recombination maps

Recombination maps of chromosome 22 were built using LDhat for each of the four datasets and the results were analysed in R. The estimated ρ/Kb value between each SNP was plotted for each dataset, see Figure 4-4.

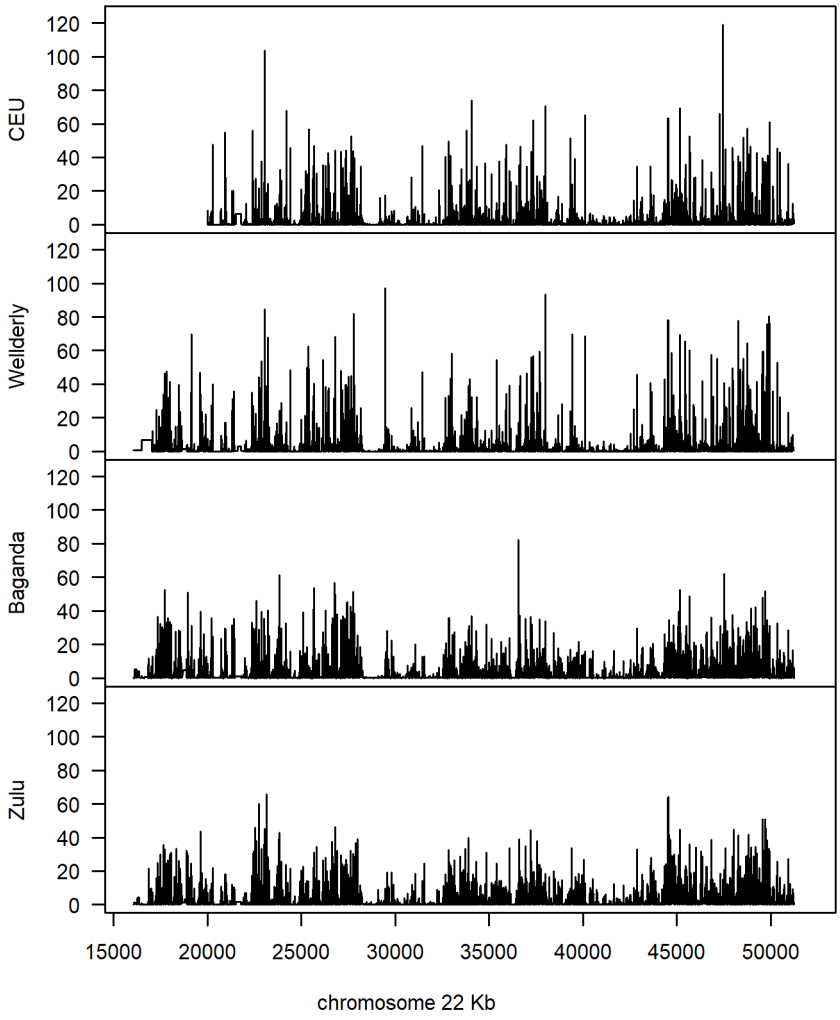


Figure 4-4 LDhat recombination rate (ρ) estimate comparison

This figure shows the estimated $\rho = 4N_e r$ per Kb values from LDhat for each of the four datasets across chromosome 22.

The graphs look similar in shape, for example all the graphs have regions of low recombination around 28,000 Kb and 42,000 Kb. Chromosome 22 is acrocentric, meaning the centromere is towards the start of the chromosome, so all the rates plotted here are on the same arm: 22q. The CEU data did not contain SNPs centromeric of 20,000 Kb, hence there is no recombination rate recorded there. The African recombination hotspots look less intense than the European

hotspots. The maximum rates for the European datasets are 119 ρ/Kb and 97 ρ/Kb for CEU and Wellderly respectively, whereas the maximum rates for Baganda was 82 ρ/Kb and Zulu was 66 ρ/Kb. The 1000 Genomes Project reported less intense hotpots in West Africans compared to Europeans and East Asians, due to Africans having more potential recombination sites available to them and thus less concentrated hot spots [102, 418].

The African datasets and the European datasets show higher correlation when compared to each other than to the other sets, as is expected, see Figure 4-5. All datasets are positively correlated with each other (p-value < 2.2e-16).

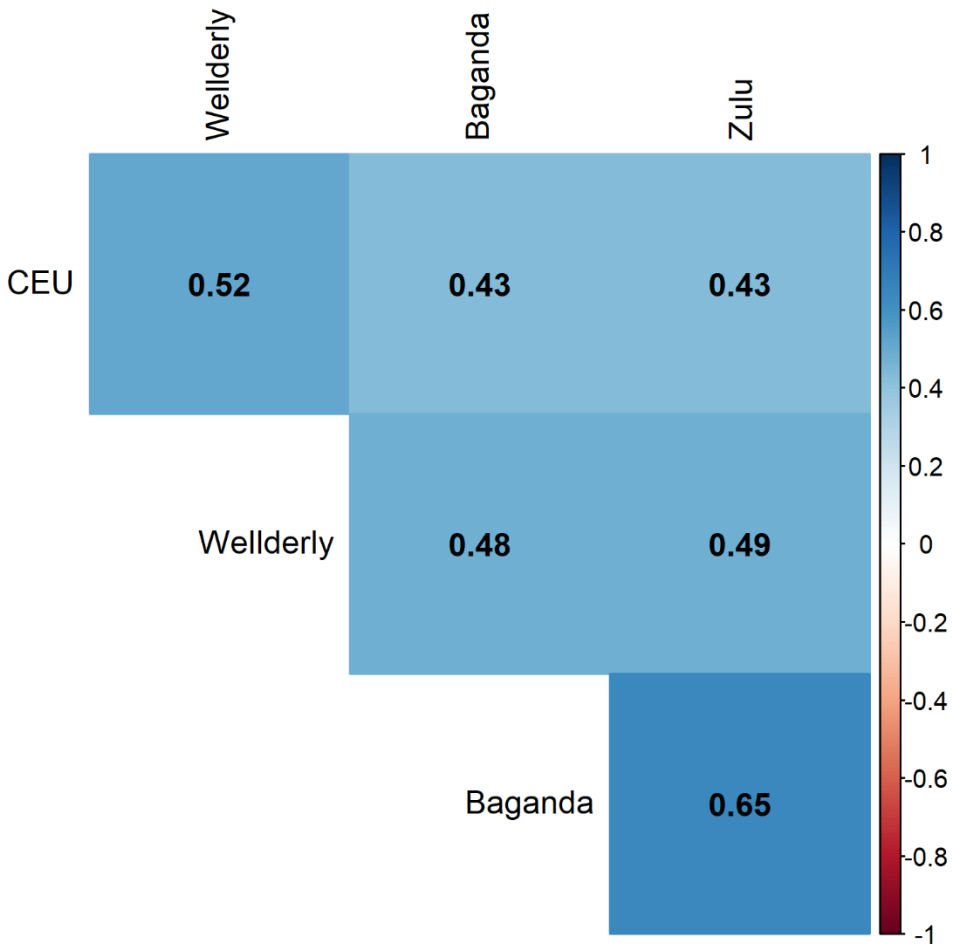


Figure 4-5 LDhat correlation comparison

This figure shows Kendall's Tau between all the datasets for the SNPs common to all datasets, where n = 34,470 across all comparisons. All coefficients are significant with p-values < 2.2e-16. All datasets show a positive correlation with each other. When comparing across the populations, the correlation is less than when comparing within populations (i.e. CEU and Wellderly, and Baganda and Zulu).

4.3.2 Comparison of maps

The LDhat output for the CEU dataset was compared to the LD map created by Pengelly *et al.* [137]. The sex averaged European linkage map created by Bherer using a MCMC method has also been included in the comparison [406]. The results can be seen in Figure 4-6. The outputs look remarkably similar, with areas of low recombination and linkage disequilibrium clearly shown on all maps. Areas with high estimated recombination rates by LDhat have a steeper incline than the estimates in the other two maps, although there is usually a rise in the other maps too. There is an interesting increase in the LDhat output around 40 Mb which is not seen in the LDMAP output or the Bherer map, indicating that there might be something else going on in that region to confound the output, such as a selective event.

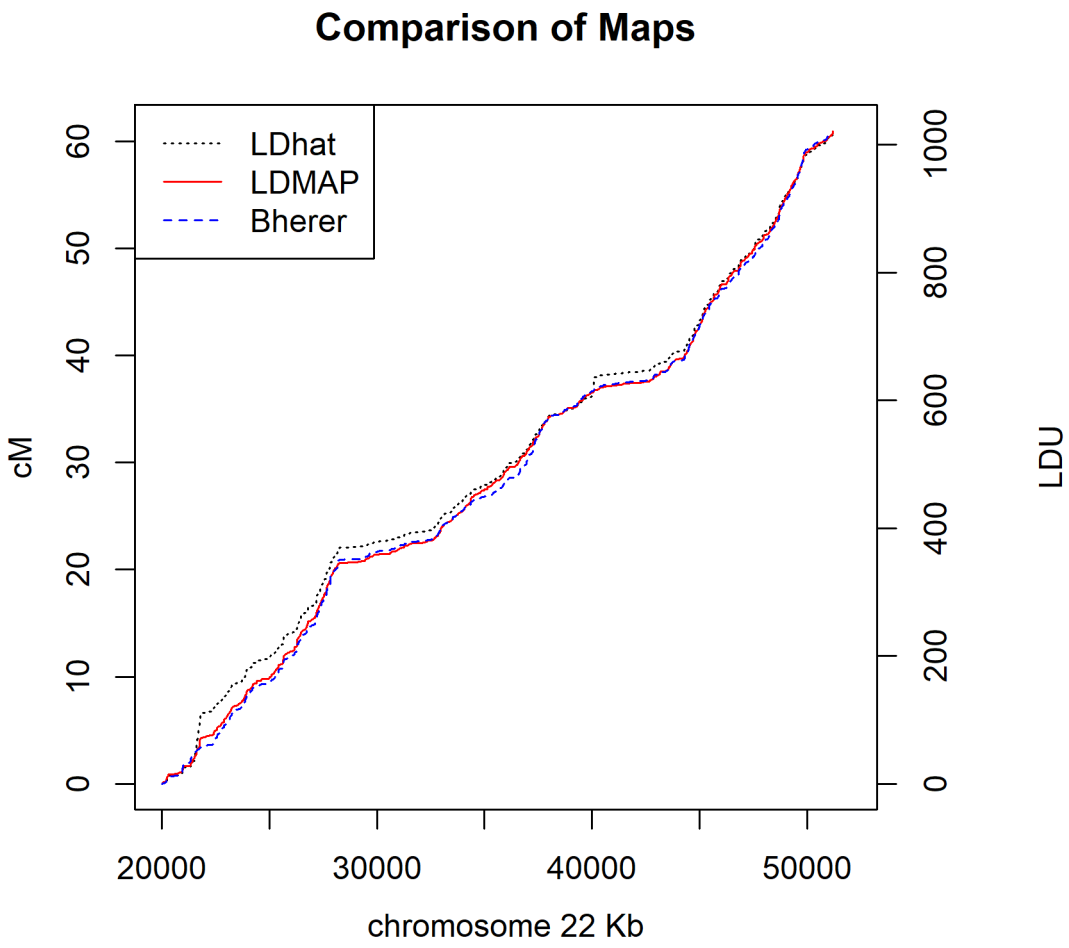


Figure 4-6 Comparison of maps: LDhat compared to LDMAP and Bherer map

This graph shows the CEU LDhat output (dotted black) and the linkage map created by Bherer (dashed blue) in terms of cumulative cMs, and LDMAP output (solid red) in cumulative LDUs. Flat regions represent areas of low recombination/high linkage disequilibrium whereas sharp jumps could indicate a recombination hotspot.

4.4 Discussion

LDhat is a widely used tool used to estimate recombination rates in human populations and other species [443]. The method was shown to be robust to a variety of demographic scenarios and SNP properties using simulation [348]. However, Dapper and Payseur [444] report an increase in false positives when demographic history is complex. The provision of pre-calculated look up tables for various *n* and θ values has allowed others to use the software without the massive computational overhead it would usually require, although these are calibrated towards humans. However, there is the possibility to interpolate between tables to obtain a more specific estimate, and a program

is included within the software to generate new tables if the parameters are very different to those of the tables provided.

When compared to a linkage map and to an LD map created using LDMAP, although all were very similar, the linkage map and the LD map seem to show a closer correspondence than to the LDhat output. This is surprising given that, while the linkage map and LDhat maps only look at recombination, the LD map returns all LD, including that caused by drift, mutation, selection etc. This was also seen in Tapper *et al.* [445] where the reasons ventured were that the differences were due to the assumptions of coalescent models, or as an artefact of the smoothing process of LDhat.

The authors of LDhat acknowledge that it is a difficult problem to obtain the pure recombination rate from the confounding factors of selection, mutation, drift, and demography [348, 446]. It has been found that some software for estimating recombination rates can be confounded by positive selective sweeps, causing false-positive hotspot detection [447]. It has been argued that the coalescent-based method used by LDhat is not affected by sweeps and at worst will produce a small decrease in recombination estimates around a strong sweep [448, 449]. However, simulations performed by Chan *et al.* [443] show some evidence for false positive hotspot detection around positive sweeps using LDhat.

rhomap is an extension of the LDhat software which claims to improve upon the LDhat output when hotspots are present [450]. However, when compared to other software, sequenceLDhot [451] outperformed both rhomap and Haploview [452] at hotspot recognition [453]. LDhot is a piece of software designed to be run after the estimating of recombination rates via software such as LDhat, to identify hotspots that are statistically significant given the background recombination rates [348].

FastEPRR (Fast estimation of population recombination rates) is a new method for estimating ρ, using a regression-based machine learning method [449]. It claims to be faster than LDhat and performs similarly to LDhat when applied to simulated data and to 1000 Genomes Project data. FastEPRR has the ability to process much larger sample sizes than LDhat, which is a clear benefit as sample sizes get larger, for example the 100,000 Genomes Project data. FastEPRR is coded in R, which makes integrating it with other processes much more streamlined than stand-alone software. pyrho is another method for estimating recombination rates, this time developed in Python [454]. Again, it is faster than LDhat and shows an improved performance, especially at the fine scale. The IBDrecomb method uses IBD to estimate recombination rates, especially of the recent past [455]. This method can also accept much larger datasets than LDhat, and the performance was comparable to both LDhat and pyrho, which in turn performed similarly to each

other in this study. While most methods require large sample sizes, iSMC (integrative sequentially Markov coalescent) is a method that can estimate recombination rates with just a single pair of genomes, which could be pertinent for very small populations or ancient populations with very few samples [456]. iSMC and IBDrecomb are both contained within their own software, coded in C and C++ respectively.

The LDhat recombination maps for chromosome 22 created here showed similar trends across the European and African datasets. Crude correlation analysis showed that the two European datasets and the two African datasets were more correlated to each other than across populations. However, a lot of information was lost by just considering SNPs the datasets had in common. Further analysis is required to find evidence of similarities and differences in rates within and between populations. For example, whether the changes in recombination rates are consistent across the chromosomes, and whether there are trends across the region, and if so, at what scale.

The CEU map created here was mostly consistent when compared to maps for Europeans created using other methods. Therefore, the results from LDhat reported here will be used for further analysis on recombination rates in different human populations in the next chapter.

Chapter 5 A wavelet analysis of European and African recombination

5.1 Introduction

Wavelet transforms are a mathematical method, related to Fourier analysis, for the processing of signals [457]. While traditionally used for time series data, they can also be applied to any information that exists along a continuum, for example depth of the sea, or more relevant to this work, distance along a chromosome. Wavelets can be used to extract information about the signal at different scales, from very short-term changes to overarching long-term trends. Wavelets can be incredibly useful for signal compression and smoothing and have been widely used for image and sound processing [458-460].

Wavelets can be applied to genomic data to analyse signals at different scales along a chromosome. To conceptualise this, it might be helpful to think about moving averages. Clearly, picking one scale (window size) for a moving average, for example 5 Kb, would return different results and potentially different conclusions about the data than picking another scale, say 50 Kb. One of the attractive things about wavelet analysis is the ability to look at fine scale and wider scales simultaneously, without having to choose arbitrary window sizes that could skew results. Recombination hotspot estimation has been shown to be sensitive to window size [461]. The contribution of each scale to the overall variance of the signal can be ascertained, and this is known as the power spectrum. This allows for easy comparison between multiple datasets across different scales.

Wavelets have been applied in the field of genomic research for many years, as evidenced in the review by Liò from 2003 [462]. More recently, there are many examples of wavelet techniques being applied in a range of genomic applications, including CNV/CNA detection [463-465], microarray de-noising [466], and downsizing of genetic data for more efficient analysis [467]. Wavelets have been used for analysing signals, for example in analysing the features found around indels [468], features of admixture in human populations [469], and dependencies between different genetic motifs [470].

Most relevant to this research are analyses involving recombination rates. Paape *et al.* [471] investigated recombination in *Medicago truncatula*, using wavelet analysis to find a negative relationship between recombination rates and the distance from the centromere in scales up to 512 Kb. In humans, wavelet analysis has been used to show that genetic diversity is only correlated with recombination rate at the very fine scale [472], and GC content correlates with recombination rates on the mid-scale (8 Kb - 512 Kb) [130]. Bherer *et al.* [406] used wavelets to

investigate sexual dimorphism in recombination rate, finding that male and female rates differed most at the fine scale, and that the power spectrum for both rates showed the majority of variance in recombination rates was found at the mid-scale between 16 and 64 Kb. Finally, Chan *et al.* used wavelet analysis to assess the difference in recombination rates between two populations of *D. melanogaster*, reporting that correlations between the populations increased with scale [443].

The next section of this chapter will go into more depth about how wavelets work and includes a worked example of three simple signals to show the theory behind them and how to interpret the results. The books by Percival and Walden [457] and Nason [473] are the main sources for the methods and descriptions in the rest of this section. After this, wavelet analysis will be applied to the recombination datasets for European and African populations created in the previous chapter. The aim of this analysis is to assess if there are differences between European and African recombination rates, and at what scale. By analysing two populations from each region, the baseline differences in recombination in different samples of the same population can be compared to the differences in samples from different populations, to give an expected level of correlation across scales and positions.

Wavelet analysis will be used in this chapter to investigate four factors. Firstly, the power spectrum will provide information on the scale within which most of the variance is contained. This will give an indication of potential periodicity in the data and will identify at what scale the biggest changes are occurring. The results are expected to be similar to the analysis by Bherer *et al.* [406], in which they found most of the variance was contained in the mid-scales due to the distribution of hotspots across the genome. Secondly, the correlations between the power spectrums of the four datasets will be compared to ascertain the similarities in the change in recombination rates across scales. If they are similar across all scales, then it may be valid to use a recombination map built for one population as a proxy for a recombination map of the other. However, if they are only similar for some scales or not similar at all, then use of the recombination map for populations other than the one it was built for may not be valid.

Thirdly, the results of the continuous wavelet transform will show the spatial distribution of the wavelet power for each of the datasets. This will show how consistently the power is distributed across the chromosome for all scales and may highlight regions that have many changes and regions that are stable. Finally, wavelet coherence analysis will show how correlated the recombination rate changes are across the chromosome between each pair of datasets. This will highlight any regions that are very similar between the datasets, and also regions that are

different, and will give further insight into the validity of using recombination maps for multiple populations.

5.2 An introduction to wavelets

5.2.1 Wavelets

Wavelets are functions, usually denoted by the Greek letter ψ, which are conditioned on two equations:

$$\int_{-\infty}^{\infty} \psi(s)\, ds = 0 \tag{5.1}$$

$$\int_{-\infty}^{\infty} \psi^2(s)\, ds = 1 \tag{5.2}$$

The second equation means that the function cannot be infinite and cannot be a flat line and the first equation means that the function must cross the x-axis at least once. This will result in a short, wave-like function, known as a wavelet. The simplest and oldest wavelet, the Haar wavelet [474], is defined as follows:

$$\psi(s) = \begin{cases} 1 & for\ 0 < s \leq \frac{1}{2} \\ -1 & for \frac{1}{2} < s \leq 1 \\ 0 & otherwise \end{cases} \tag{5.3}$$

Figure 5-1 shows the Haar wavelet, which clearly satisfies equations (5.1) and (5.2). Figure 5-1 also shows how the Haar wavelet can be stretched, shifted along the axis, and flipped, whilst maintaining adherence to the conditions. Further analysis in this chapter uses the formulation of the Haar wavelet shown in Figure 5-1d where $\psi(s) = -\frac{1}{\sqrt{2}}$ when *s* is on the interval (-1,0] and $\psi(s) = \frac{1}{\sqrt{2}}$ for *s* is in (0,1] as is implemented in the wmtsa R package [475].

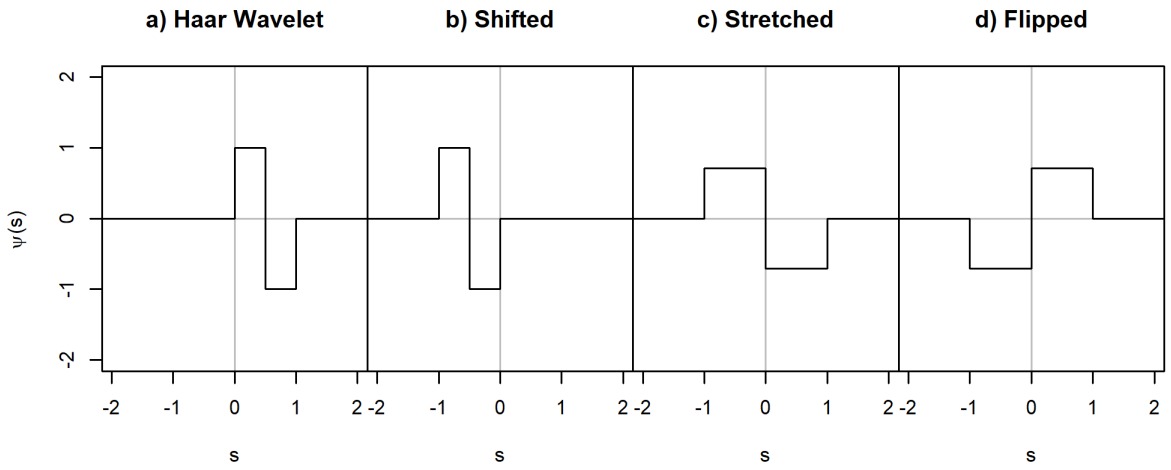


Figure 5-1 The Haar wavelet

This figure shows the Haar wavelet and various transforms. Figure a) is the Haar wavelet, in b) it has been shifted left along the x-axis. Figure c) shows the wavelet after it has been stretched, with $\boldsymbol{\psi(s)} = \frac{\mathbf{1}}{\sqrt{\mathbf{2}}}$ and $-\frac{\mathbf{1}}{\sqrt{\mathbf{2}}}$ when it is not 0. Figure d) shows the same as figure c) but mirrored at s = 0. All these transforms satisfy the wavelet conditions.

The Haar wavelet is an example of a discrete wavelet. Other discrete wavelets include the Daubechies wavelets [476] and the Coiflets [477]. Continuous wavelets can be both real and complex valued, with the Mexican hat wavelet [457] an example of the former and the Morlet wavelet [478, 479], which will be discussed later in section 5.2.5, an example of the latter.

5.2.2 Wavelet analysis example

Consider the three data signals shown in Figure 5-2. The first signal consists of only random noise, and the other two, while also containing random noise, also incorporate a wave with a period of 64 Kb. These example signals will be used to illustrate how wavelet analysis works and give some clarity to how the results of the analysis should be interpreted. For the purposes of this example, the x-axis represents a chromosome with the units in Kb, and the y-axis is some measure observed along the chromosome in 1 Kb intervals.

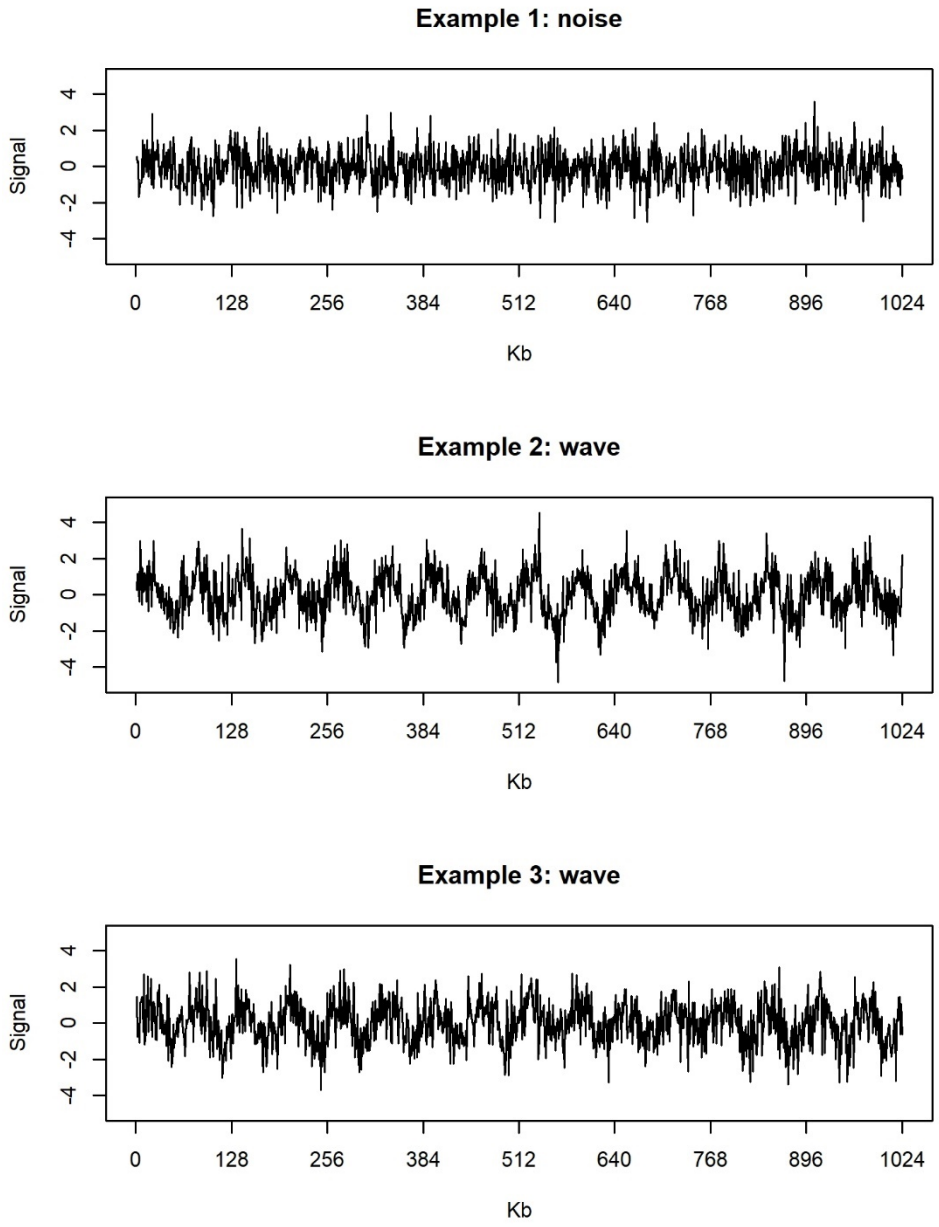


Figure 5-2 Example signals

This figure shows three example signals. The x-axis represents a chromosome and is measured in kilobases. The y-axis represents some signal that is measured along the chromosome. The first example just contains random noise drawn from a normal distribution with mean zero and standard deviation one. The second and third examples also add a sine wave with period 64 Kb to random noise.

The data for the examples were generated in R version 3.6.0 [480]. For each example 1,024 random normal values were generated using a mean of zero and a standard deviation of one – that is each value is *iid* (independent and identically distributed) and the signal can be described as Gaussian white noise. The second and third examples added a sine curve to the random values,

by taking the sine of the *x* position multiplied by π/32. The wavelet analysis in the next sections utilised the R packages wmtsa 2.0-3 [475] and biwavelet 0.20.19 [481].

5.2.3 Discrete wavelet transform

There are two types of wavelet transform: the discrete wavelet transform (DWT) and the continuous wavelet transform (CWT). The CWT will be discussed further in section 5.2.5, after first introducing the DWT as it is easier to conceptualise for this example. Another advantage of the DWT is that it is more efficient than the CWT. Where the CWT will calculate transforms for every scale, which will often differ only slightly between neighbouring scales, the DWT will only consider scales that are powers of two, i.e. 2 Kb, 4 Kb, 8 Kb, 16 Kb etc.

The DWT breaks down the original signal $x = (x_1, x_2, \ldots, x_N)$ into J levels, where the original data signal is $N = 2^J$ data points in length. Each level j in 1 to J corresponds to scale 2^j, up to a maximum 2^J. At each level the transform will generate two sets of coefficients: detail coefficients and approximation coefficients, where the first level will have half as many data points as the original signal, and each subsequent level will have half as many as the scale before that. Approximation coefficients are proportional to the averages of the original signal, whereas detail coefficients are proportional to the changes between averages.

To get the first level of detail coefficients, the wavelet filter is applied to each set of observations along the series, where the set size depends on the length of the chosen wavelet filter. For the Haar wavelet, the wavelet filter is $\left\{h_1 = -\frac{1}{\sqrt{2}}, h_2 = \frac{1}{\sqrt{2}}\right\}$. In this case, the filter is of length two and will be applied to x_1 and x_2, then x_3 and x_4 etc. The first level detail coefficients for the Haar DWT can thus be written:

$$d(x)_i^{(1)} = \frac{x_{2i} - x_{2i-1}}{\sqrt{2}} \; for \; i = 1 \; to \; \frac{N}{2} \tag{5.4}$$

The approximation coefficients are calculated similarly, but the wavelet filter is transformed by reversing the order and multiplying every second number by -1 to get the scaling filter. More formally, this transform is known as the quadrature mirror filter (QMF) and is defined as:

$$g_l = (-1)^{l+1} h_{L-l+1} \tag{5.5}$$

Where h_l are the wavelet filters, g_l are the scaling filters, and L is the length of the wavelet filter.

For the Haar Wavelet, this results in scaling filters of $\left\{\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right\}$. The first level of approximation coefficients are thus defined as:

$$a(x)_i^{(1)} = \frac{x_{2i} + x_{2i-1}}{\sqrt{2}} \; for \; i = 1 \; to \; \frac{N}{2} \tag{5.6}$$

Note that this is very similar to calculating the average between the two values, but instead is just proportional to the average. Scaling in this way means that no energy is lost in the from the original signal, as will be discussed further later in this section.

For subsequent levels, the same filters are applied to the previous level's approximation coefficients:

$$d(x)_i^{(j)} = \frac{a_{2i}^{(j-1)} - a_{2i-1}^{(j-1)}}{\sqrt{2}} \; for \; i = 1 \; to \; \frac{N}{2^j} \tag{5.7}$$

$$a(x)_i^{(j)} = \frac{a_{2i}^{(j-1)} + a_{2i-1}^{(j-1)}}{\sqrt{2}} \; for \; i = 1 \; to \; \frac{N}{2^j} \tag{5.8}$$

It should be noted that no information is lost by making these transformations and that the original signal can be reconstructed given the approximation coefficients for a level, the wavelet used and the detail coefficient for the level and each previous level. For the Haar wavelet, the approximation coefficient for a level (or indeed the original signal if *j = 0*) is found as follows:

$$a(x)_i^{(j)} = \frac{a_k^{(j+1)} - d_k^{(j+1)}}{\sqrt{2}} for \; i = 1, 3, \dots, \frac{N}{2^j} - 1; k = \frac{i+1}{2} \tag{5.9}$$

$$a(x)_i^{(j)} = \frac{a_k^{(j+1)} + d_k^{(j+1)}}{\sqrt{2}} for \; i = 2, 4, \dots, \frac{N}{2^j}; k = \frac{i}{2} \tag{5.10}$$

For other wavelets, especially those with a wavelet filter length greater than two, this calculation becomes more complex to write out by hand, but the principle is the same.

Figure 5-3 shows the wavelet decomposition for the signal in example 2, illustrating both its detail and approximation coefficients. Appendix A.1.1 shows the decomposition for all three examples. It is clear to see the sine wave pattern from the original signal in the approximation coefficients down to the 32 Kb scale.

As a simple example of how the reconstruction works, take the $a(x)^{(10)}$ value of 0.0722 from this example and the $d(x)^{(10)}$ value of -0.4326. Plugging these values into the above formulae give:

$$a(x)_1^{(9)} = \frac{0.0722 - (-0.4326)}{\sqrt{2}} = 0.357$$

$$a(x)_2^{(9)} = \frac{0.0722 + (-0.4326)}{\sqrt{2}} = -0.255$$

These are the values for $a(x)^{(9)}$. Continuing up the scales, eventually the original signal will be reached.

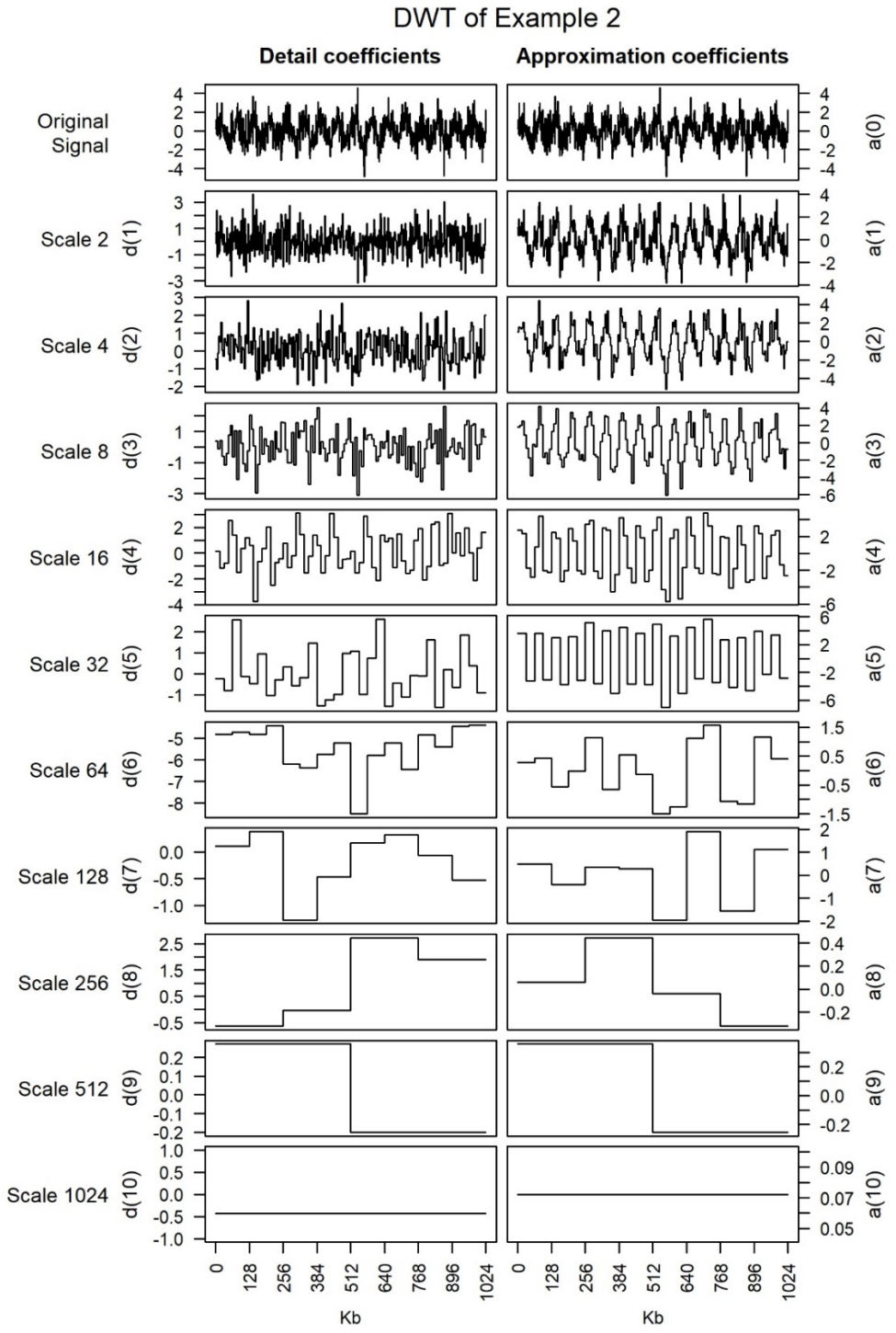


Figure 5-3 Wavelet decomposition of example 2

This figure shows the wavelet decomposition of the example 2 signal using the Haar DWT. The top graphs on each side show the original signal, with each graph below showing the detail and approximation coefficients for that scale. Approximation coefficients are proportional to the averages of the original signal, and detail coefficients are proportional to the changes in the averages. Note the y-axis is scaled for each graph individually so the detail can be seen.

As briefly mentioned before, a property of wavelet transformation is the ability to decompose the energy in the original signal to each of the scales [482]. The energy of the original signal is defined as the sum of the squares of each of the observations. This can be found from the transform by summing the squares of all the detail coefficients of every level, plus the sum of the square of the approximation coefficients for the final level. The energy decomposition is thus written as follows:

$$\sum_{i=1}^{N} x_i^2 = \sum_{j=1}^{J} \sum_{i=1}^{N_j} \left(d_i^{(j)}\right)^2 + \sum_{i=1}^{N_J} \left(a_i^{(J)}\right)^2 \qquad (5.11)$$

Where N_j is the number of coefficients at level *j*, and *J* is the number of levels, i.e. $log_2 N$. While this is not immediately useful, the relationship between this and the variance of the original signal is. From this equation, it can now be written (see page 62 of Percival and Walden [457] for proof):

$$var(x) = \frac{\sum_{j=1}^{J} \sum_{i=1}^{N_j} \left(d_i^{(j)}\right)^2}{N} \qquad (5.12)$$

That is, the variance of the original signal can be found from the average of the squared detail coefficients. From this, the power spectrum can be calculated.

The power spectrum is a highly useful feature that can be extracted from the detail coefficients. This calculation allows one to ascertain what proportion of the variance from the original signal is attributable to each scale, and only that scale. It is calculated using the formula:

$$PS(j) = \frac{\sum_{i=1}^{N_j} \left(d_i^{(j)}\right)^2}{\sum_{k=1}^{K} \sum_{i=1}^{N_k} \left(d_i^{(k)}\right)^2} \qquad (5.13)$$

Where *PS(j)* is the proportion of variance for scale *j*, and the denominator is equivalent to the numerator in the previous equation (5.12).

More simply put, this formula sums the squared detail coefficients for a scale, and divides by the sum of the squared detail coefficients for every scale. Hence, $\sum_j PS(j) = 1$. Figure 5-4 shows the power spectrum for the three example signals. It is clear to see that for example 1, the signal containing only noise, the highest proportion of variance is at the fine scale (2 Kb) and decreases as the scale increases. For the other two examples containing the sine wave, again there is a large proportion of variance at the fine scale due to the noise, and a spike at 64 Kb corresponding to the periodicity of the sine wave in the original signals.

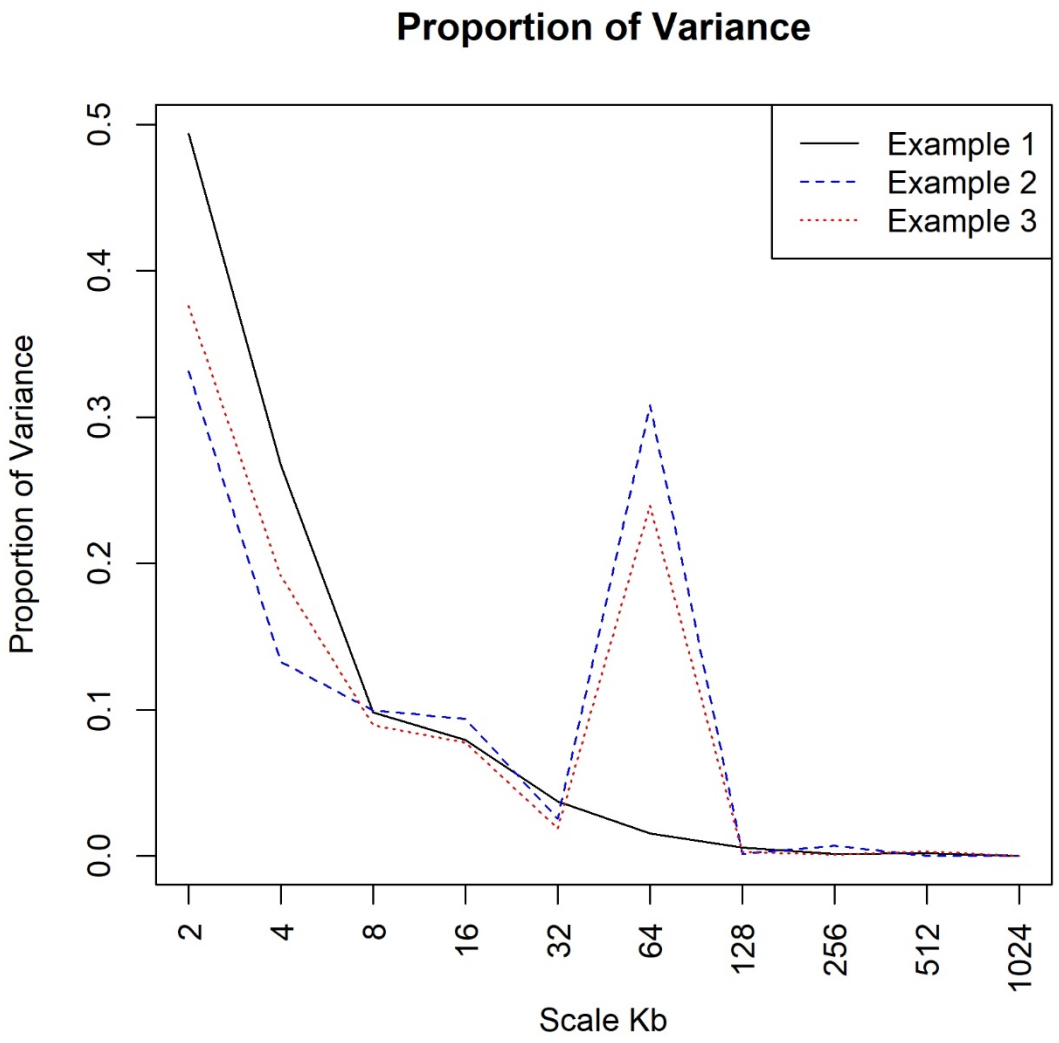


Figure 5-4 Power spectrum of the three examples

This figure shows the power spectrums for the three examples. Example 1 (black) contains only noise and thus shows the greatest proportion of variance at the finer scales. The other two examples (blue dashes and red dots) show a spike at 64 Kb corresponding to the period of the sine wave in their original signals.

One of the major drawbacks of the DWT is that a different decision on where to start the sample signal will result in different detail coefficients and potentially different interpretations. It is clear to see from equation (5.4) that while the difference between x_2 and x_1 is calculated, and x_4 and x_3, there is no calculation on the differences between x_3 and x_2 or x_5 and x_4. To illustrate this, the example 2 series was rotated by 15 data points by moving the first 15 observations to the end of the signal and the DWT recalculated. A rotation of 15 was chosen as now the first data point coincides with the peak of the underlying sine curve.

Figure 5-5 shows the new decomposition of example 2. Comparing to the original decomposition in Figure 5-3, the decompositions look very different. For example, there is clear periodicity in the detail coefficients at scale 32 Kb that was not present in the original decomposition, and conversely the periodicity that was seen in the approximation coefficients at the same level has been lost.

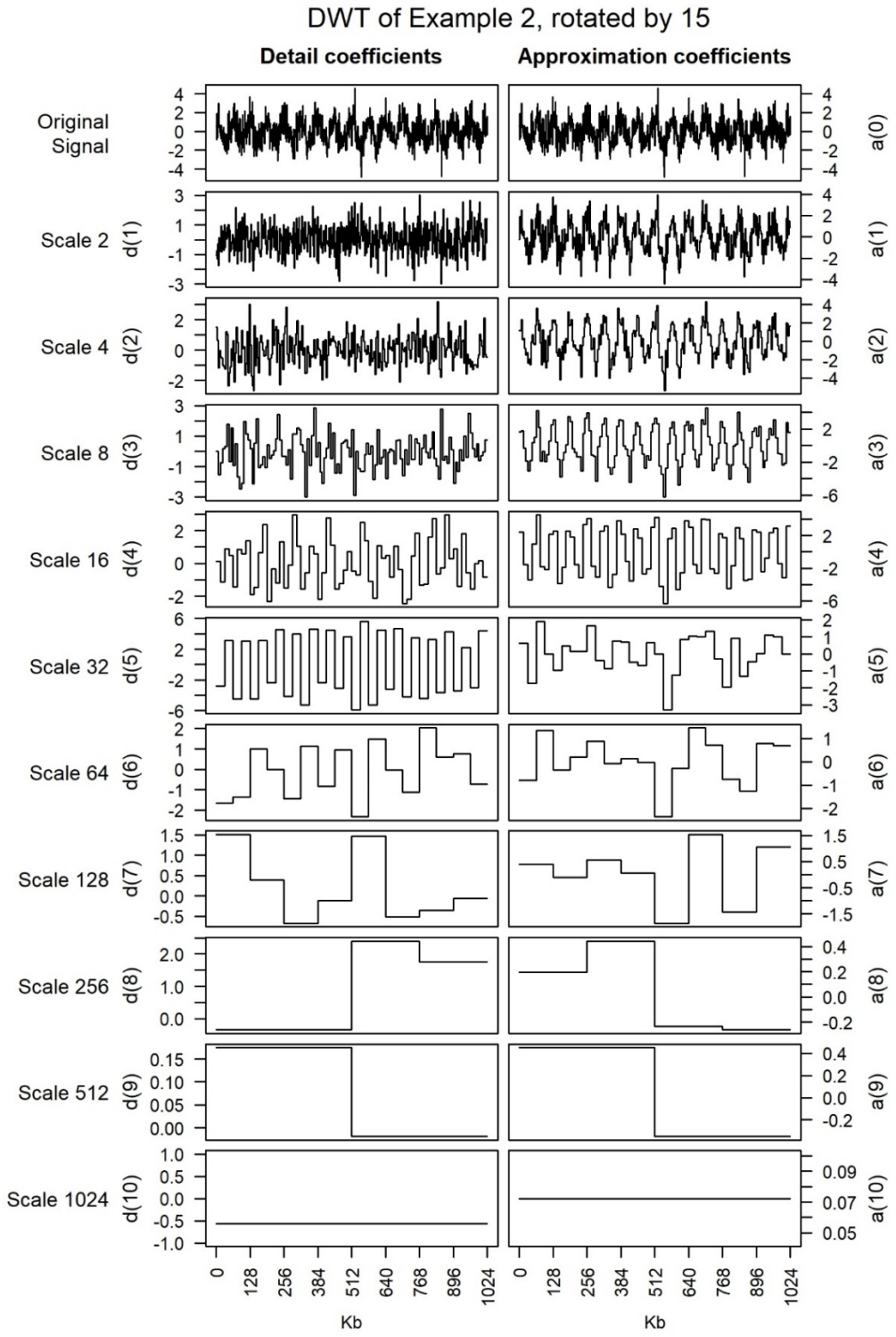


Figure 5-5 Wavelet decomposition of example 2 rotated by 15 data points

This graph shows the wavelet decomposition of the rotated example 2 dataset. The top two graphs are the original signal (rotated by 15). The left-hand graphs are the detail coefficients and the right-hand graphs are the approximation coefficients for each scale. The patterns in the graphs are noticeably different than the original decomposition in Figure 5-3.

Figure 5-6 shows the power spectrums for the original example 2 signal and the new rotated signal. The peak at 32 Kb shows that, by choosing the original starting point, information was lost about energy at the 32 Kb scale. Similarly, by choosing the rotated starting point, information was lost about the 64 Kb scale. This graph illustrates one of the main drawbacks of the DWT in that different conclusions can be drawn from the data depending on the decision of where to start the data series. To avoid this problem, the maximal overlap DWT can be utilised.

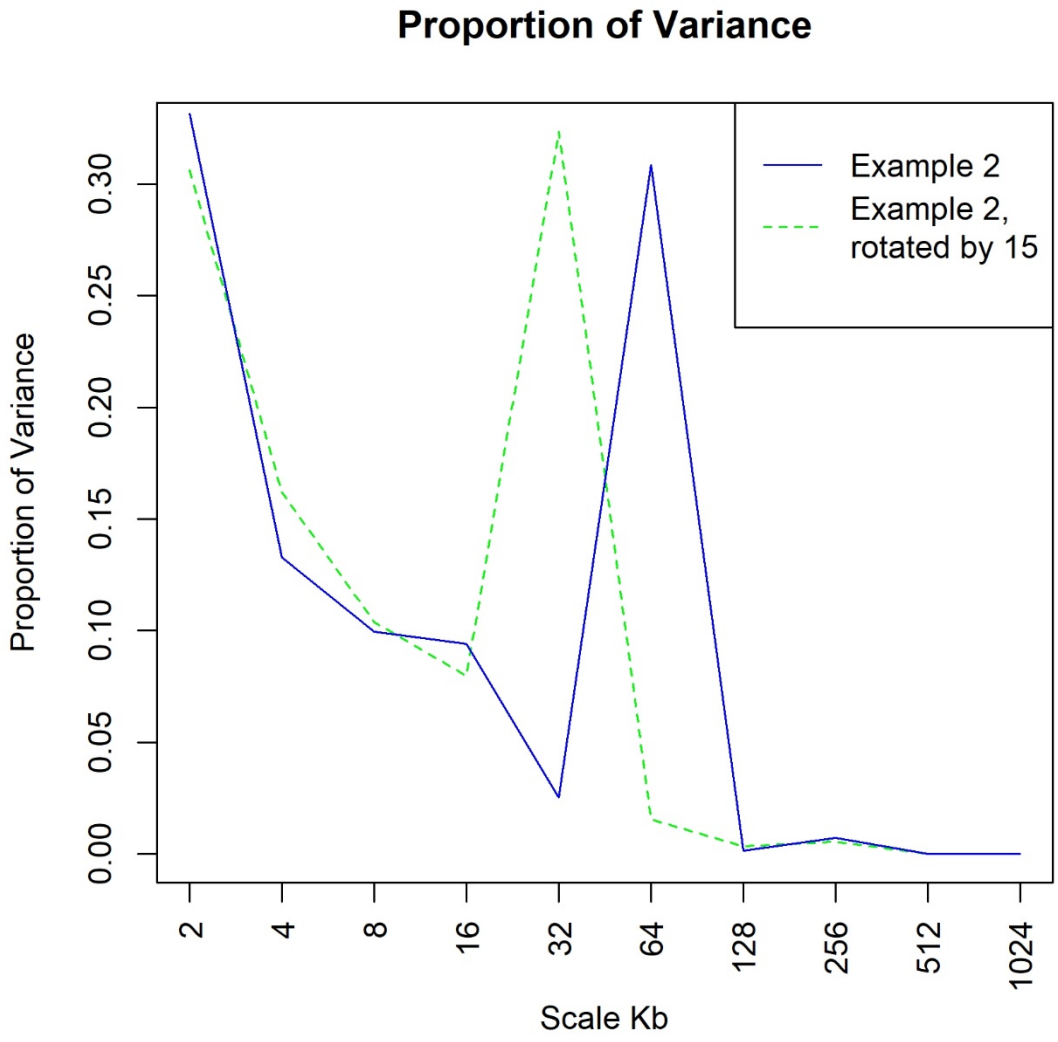


Figure 5-6 Power spectrum of example 2: original and rotated

This figure shows the proportion of variance contained in each scale for example 2. Rotating example 2 by 15 data points has resulted in a different power spectrum (in green dashes) than before (in blue). While the peak originally was at 64 Kb, the new decomposition now shows a higher proportion of variance at the 32 Kb scale.

5.2.4 Maximal Overlap DWT

The maximal overlap discrete wavelet transform (MODWT), while being less efficient than the DWT, has many benefits. As described at the end of the previous section, one of the benefits is that it is not affected by the choice of starting points, and no information will be lost due to this. Another benefit is that it does not require the series to have exactly $N=2^j$ data points. To describe the MODWT, again the Haar wavelet and the three examples will be used.

For the MODWT, instead of there being $N/2$ wavelet coefficients in the first level and half as many in the next level and so on, there are N wavelet coefficients calculated for every level. This makes the MODWT less efficient than the DWT; however, the number of calculations is still acceptable.

Firstly, the wavelet and scaling filters need to be rescaled by $\frac{1}{\sqrt{2}}$ so that the energy will still be conserved across scales in the wavelet coefficients. For the Haar wavelet, this means the wavelet and scaling filters are now:

$$\{\tilde{h}_l\} = \begin{Bmatrix} \tilde{h}_1 = -\frac{1}{2} \\ \tilde{h}_2 = \frac{1}{2} \end{Bmatrix}, \qquad \{\tilde{g}_l\} = \begin{Bmatrix} \tilde{g}_1 = \frac{1}{2} \\ \tilde{g}_2 = \frac{1}{2} \end{Bmatrix}$$

Where $\tilde{h}_l$ and $\tilde{g}_l$ are the MODWT rescaled versions of h_l and g_l from the DWT and can still be calculated from each other via the QMF. The first level detail and approximation coefficients can now be calculated using these formulae:

$$\tilde{d}(x)_i^{(1)} = \frac{x_i - x_{i-1}}{2} \; for \; i = 1 \; to \; N \tag{5.14}$$

$$\tilde{a}(x)_i^{(1)} = \frac{x_i + x_{i-1}}{2} \; for \; i = 1 \; to \; N \tag{5.15}$$

Where $\tilde{d}(x)$ and $\tilde{a}(x)$ are the detail and approximation coefficients for the MODWT using the rescaled filters. Note the circularity in these formulae: for $i = 1$, both formulae call term "x_0" which corresponds to x_N, and in the same way if the formula called for x_{-1} this would refer to x_{N-1} etc. In general, x_{i-k} corresponds to $x_{((i-k-1) \bmod N)+1}$. There is an inherent assumption in using the MODWT that it is valid for the data to analysed circularly in this way, which should be considered before implementing the MODWT.

Subsequent wavelet and approximation coefficients are calculated using the following formulae:

$$\tilde{d}(x)_i^{(j)} = \frac{a_i^{(j-1)} - a_{i-2^{j-1}}^{(j-1)}}{2} \; for \; i = 1 \; to \; N \tag{5.16}$$

$$\tilde{a}(x)_i^{(j)} = \frac{a_i^{(j-1)} + a_{i-2^{j-1}}^{(j-1)}}{2} \; for \; i = 1 \; to \; N \qquad (5.17)$$

Note that for each subsequent level, instead of just comparing the neighbouring pairs, the calculation is executed on pairs further and further away by a factor of 2.

The MODWT decomposition for example 2 is found in Figure 5-7. In a comparison with the original DWT in Figure 5-3, it is clear to see the periodicity up to scale 64 Kb in the detail coefficients. The MODWT decomposition for all three examples can be found in Appendix A.1.2.

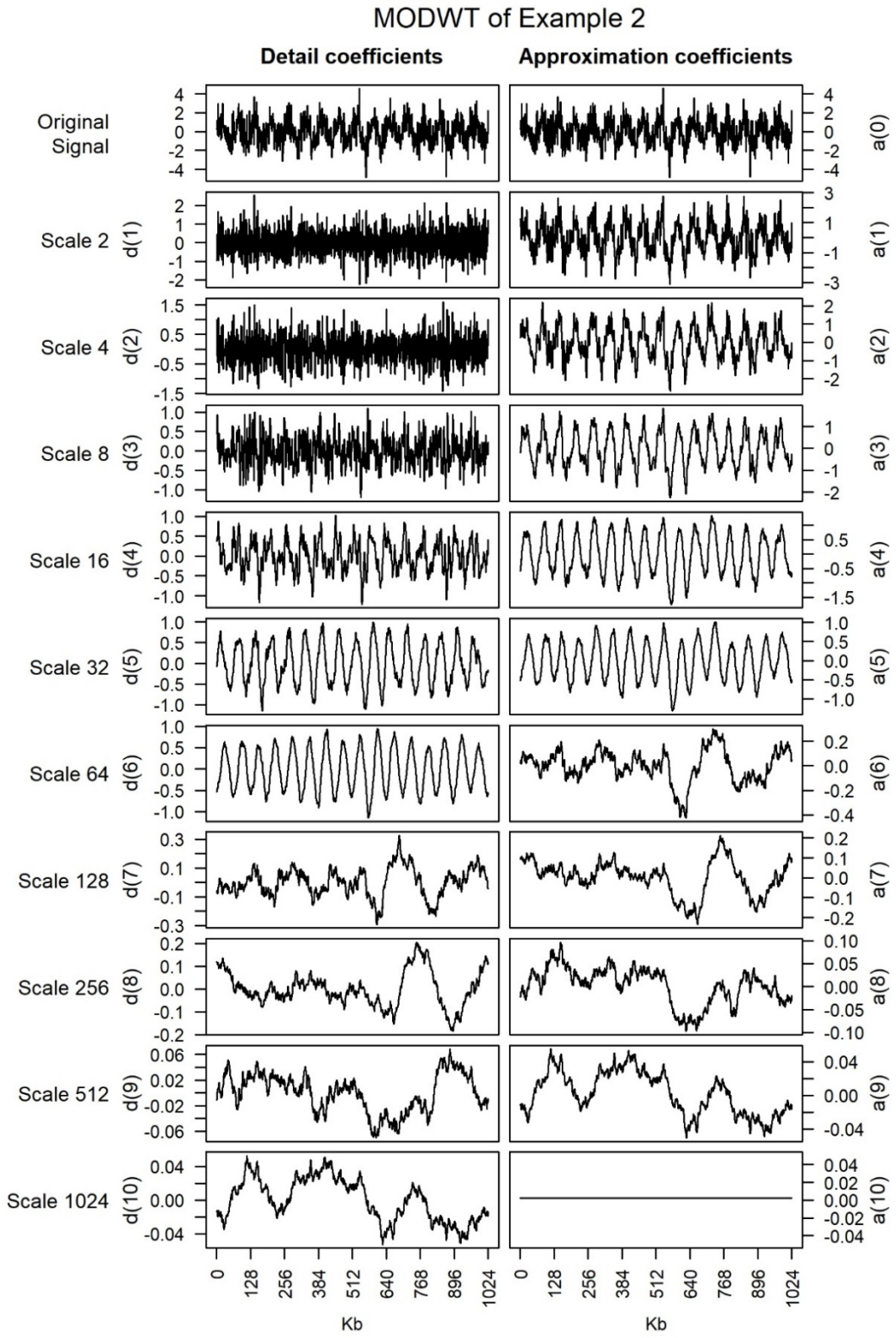


Figure 5-7 MODWT decomposition of example 2

This figure shows the maximal overlap discrete wavelet transform of the example 2 dataset. Unlike the DWT, this decomposition does not depend on the choice of start point for the series. This figure shows clear periodicity in scales 64 Kb and finer for the detail coefficients.

The power spectrum for the MODWT can still be calculated, as the energy is still conserved over all the scales in the MODWT – the following formula still holds for the MODWT:

$$\sum_{i=1}^{N} x_i^2 = \sum_{j=1}^{J} \sum_{i=1}^{N_j} \left(\tilde{d}_i^{(j)}\right)^2 + \sum_{i=1}^{N_J} \left(\tilde{a}_i^{(J)}\right)^2 \qquad (5.18)$$

And:

$$var(x) = \frac{\sum_{j=1}^{J} \sum_{i=1}^{N_j} \left(\tilde{d}_i^{(j)}\right)^2}{N} \qquad (5.19)$$

Note that this final formula only holds when $N = 2^J$, otherwise an adjustment would have to be made involving the $\tilde{a}^{(J)}$ terms.

Thus, the power spectrum can be determined as before, and Figure 5-8 shows the proportion of variance for each scale using the MODWT decomposition. This figure shows there is a large proportion of the variance explained at both the 32 Kb and 64 Kb scales in examples 2 and 3 as expected given the results in the previous power spectrums in Figure 5-4 and Figure 5-6.

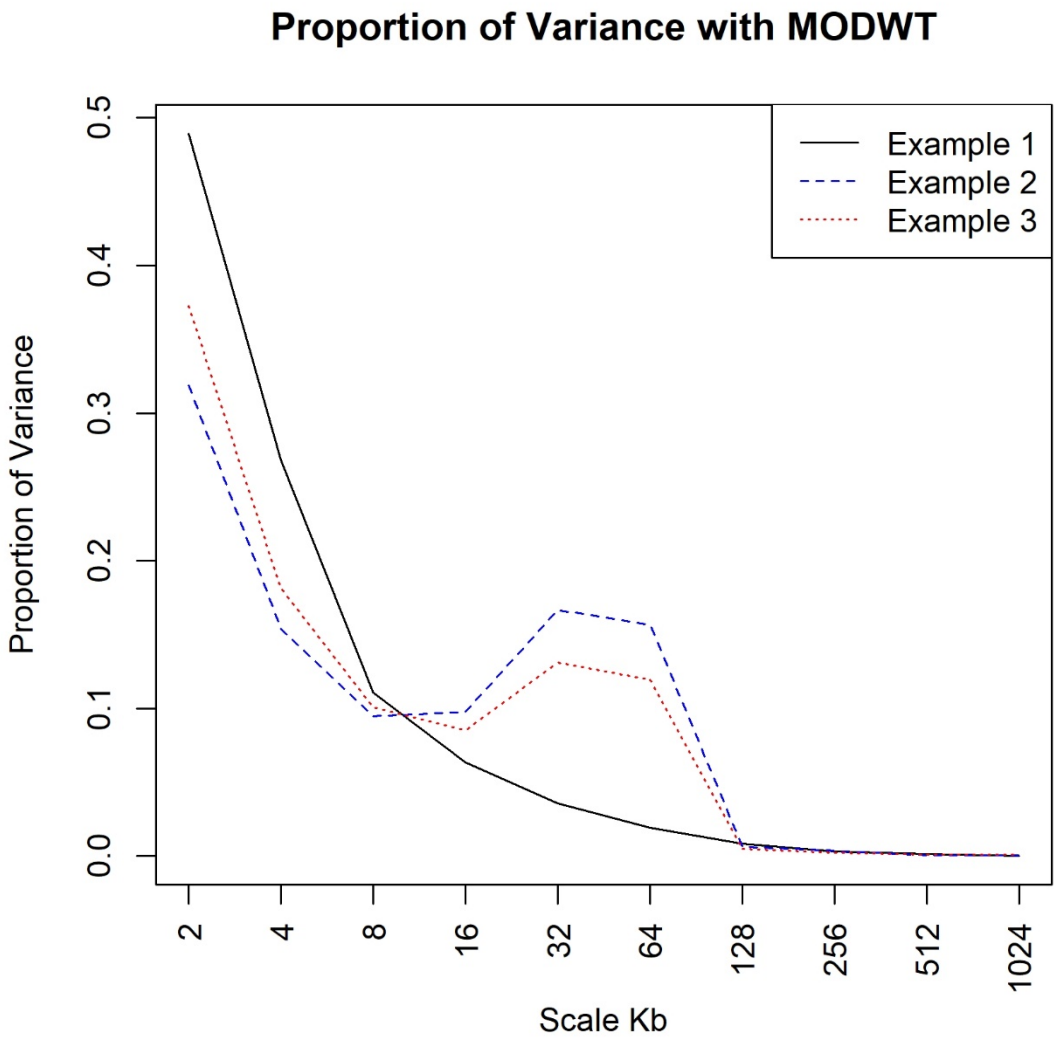


Figure 5-8 Power spectrum of the three examples using MODWT

This figure shows the proportion of variance contained in each scale for the three examples, this time calculated using the MODWT. This graph is smoother than the previously calculated power spectrums and shows an increase in energy for both scales 32 Kb and 64 Kb in examples 2 and 3.

Another useful piece of information that can be drawn from the DWT or MODWT is the correlation between detail coefficients. This is a way to compare different signals to see if there is any correlation in the changes i.e. if a change in one can predict the other and at what scale. This can be utilised to compare the same signal over multiple data sets (e.g. recombination rates in multiple populations) or for comparing different signals in the same population, such as GC content, recombination, and diversity among others in Spencer *et al.* [472]. The correlation for each scale is calculated by performing a Kendall's Tau test using the detail coefficients for the scale for both sets of data.

Figure 5-9 shows the correlations between the three examples, with the statistically significant points circled ($p<0.01$). As expected, the correlations between example 1 and the other two examples is low across all scales. There is significantly positive correlation between examples 2 and 3, the examples containing the sine wave, for scales 8 Kb up to 64 Kb. This is expected, as both datasets move in the same way (other than noise) up to windows to 64 Kb in length. The significant correlations at 128 Kb and 256 Kb are somewhat surprising, as they are not due to anything inherent about the examples and therefore must be due to random chance. It is precisely because of these spurious correlations that the analysis in the next section with real human data compares two datasets from each population, so like-for-like comparisons can validate each other.

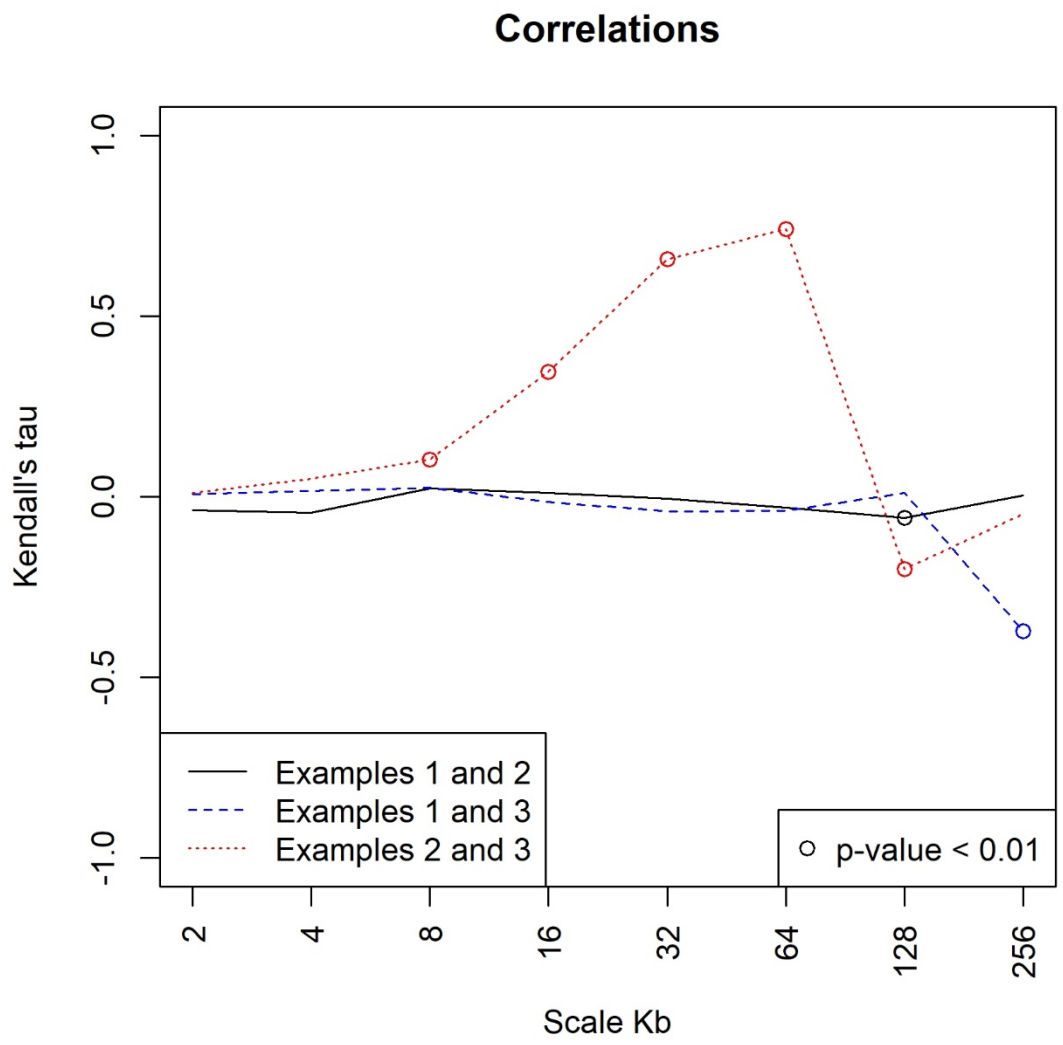


Figure 5-9 Correlations between the three examples

This figure shows the correlations between the detail coefficients of the three examples for each scale. This was calculated using the MODWT detail coefficients for each example data set. The correlations were calculated in R using Kendall's Tau. Correlations with p-values of less than 0.01 have been circled. While the correlations between pairs including example 1 are relatively flat, the correlations between examples 2 and 3 (in red dots) are significantly higher than zero for scales 8 Kb to 64 Kb.

5.2.5 Continuous wavelet transform

The continuous wavelet transform (CWT) is like the DWT except it is defined over a continuum rather than at specified intervals as in the DWT. In the previous section it was stated that wavelets need to satisfy the conditions given in equations (5.1) and (5.2). However, there is also another

condition they need to satisfy to be able to reconstruct the signal from the CWT coefficients. This is called the admissibility condition, and it is satisfied when:

$$0 < C_\psi < \infty$$

Where the scaling constant C_ψ is defined as:

$$C_\psi = \int_0^\infty \frac{|\Psi(f)|^2}{f} df \tag{5.20}$$

And Ψ(f) is the Fourier transform of the wavelet function:

$$\Psi(f) = \int_{-\infty}^{\infty} \psi(s) e^{-2i\pi fs} ds \tag{5.21}$$

There is a family of wavelets called the Morlet wavelets that can be used for CWT analysis. As in Chan *et al.* [443], here the Morlet wavelet with parameter $\omega_0 = 6$ is employed as it is widely used and appropriate for extracting features from signals [483]. The Morlet wavelet is defined as:

$$\psi(s) = \pi^{-1/4} e^{-i\omega_0 s} e^{-s^2/2} \tag{5.22}$$

Note, this is a simplified version of the Morlet wavelet that is accurate for large (≥5) ω_0 but will not integrate to exactly zero as required for the wavelet conditions [484]. See Percival and Walden [457] for a proof that the non-simplified Morlet wavelet satisfies the conditions. Figure 5-10 shows the Morlet wavelet with parameter $\omega_0 = 6$. As the Morlet wavelet is complex, the real and imaginary parts have been plotted separately.

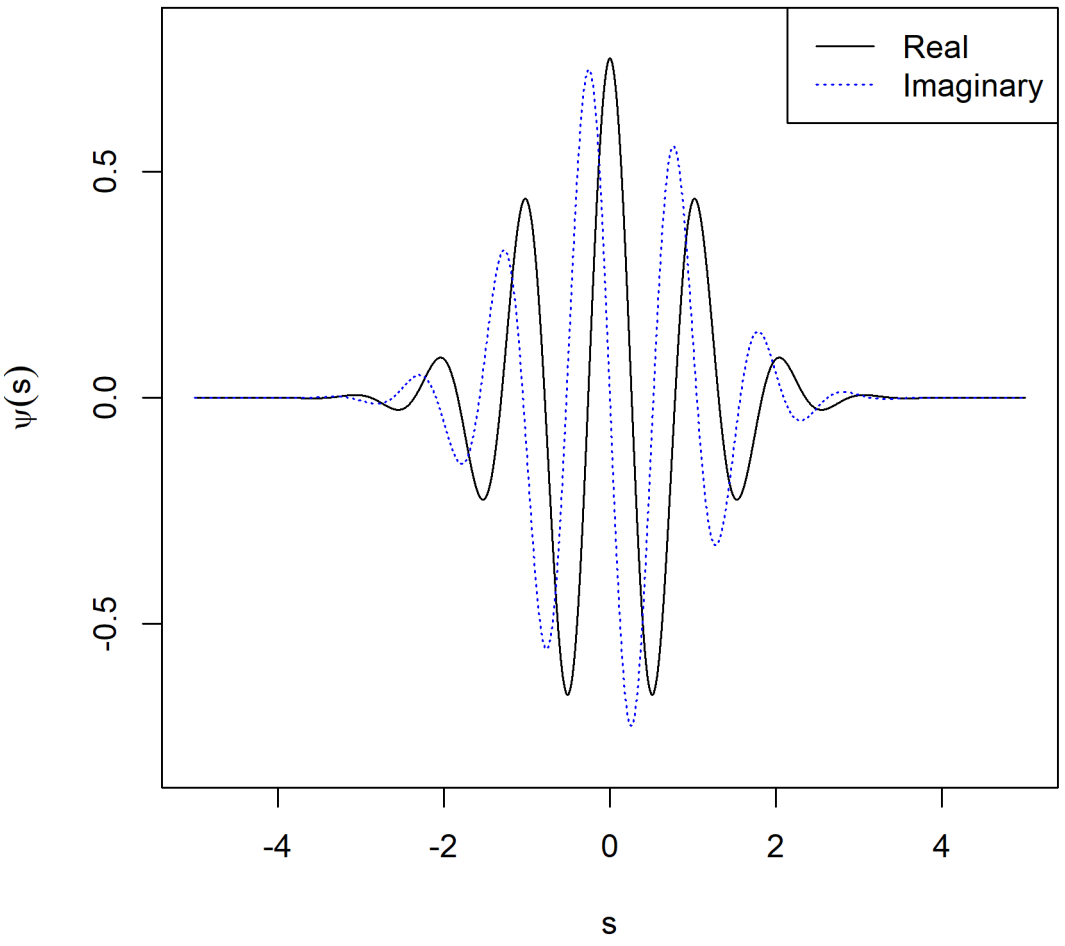


Figure 5-10 The Morlet wavelet

This figure shows a plot of the Morlet wavelet with parameter $\omega_0 = 6$. As the Morlet wavelet is complex, it has been split into its real and imaginary parts in solid black and blue dotted lines respectively.

The coefficients for the CWT are calculated using the following formula:

$$W(\lambda, t) = \int_{-\infty}^{\infty} \psi_{\lambda,t}(s) x(s) ds \qquad (5.23)$$

Where λ is the scale, t is the position and:

$$\psi_{\lambda,t}(s) = \frac{1}{\sqrt{\lambda}} \psi\left(\frac{s-t}{\lambda}\right) \qquad (5.24)$$

Just like the DWT the original signal can be recovered from the CWT coefficients. The formula to achieve this is:

$$x(t) = \frac{1}{C_\psi}\int_0^\infty \left[\int_{-\infty}^{\infty} W(\lambda, t)\frac{1}{\sqrt{\lambda}}\psi\left(\frac{t-s}{\lambda}\right)ds\right]\frac{d\lambda}{\lambda^2} \quad (5.25)$$

Where C_ψ was defined in equation (5.20).

Again, as in the DWT in equation (5.11) and the MODWT in equation (5.18), the energy of the signal can be decomposed. The formula is:

$$\int_{-\infty}^{\infty} x^2(t)dt = \frac{1}{C_\psi}\int_0^\infty \left[\int_{-\infty}^{\infty} W^2(\lambda, t)dt\right]\frac{d\lambda}{\lambda^2} \quad (5.26)$$

The wavelet power is defined as $|W^2(\lambda, t)|$. To calculate and plot this, the biwavelet R package was used [481]. This package uses the normalised, bias-corrected version of the wavelet power as in Liu *et al.* [485], to allow for accurate comparisons across positions and scale. An adjustment needs to be made for the difference between the wavelet scale and the Fourier period. Using the formula from Meyers *et al.* [486], the Fourier factor for Morlet wavelets is defined as:

$$\frac{4\pi}{\omega_0 + \sqrt{2 + \omega_0^2}} \quad (5.27)$$

For ω_0 = 6 as used here, this resolves to ≈1.033044 as also stated in Grinsted *et al.* [483].

The edge effects are dealt with by defining a cone of influence (COI), where values falling within the cone should be disregarded. The COI is calculated by dividing the distance from the edge by $\sqrt{2}$, and multiplying by the Fourier factor (calculated above). The $\sqrt{2}$ factor is determined by the *e*-folding time, where the power for values at the edge will have dropped by a factor of e^{-2} [487].

The significance of the wavelet power is calculated by comparison to an expected red-noise spectrum. A red-noise spectrum has increasing power towards lower periods, in contrast to a white-noise spectrum which has flat power regardless of period. This is calculated over a few steps. First an autoregressive model with lag 1, or AR(1), is fitted to find the to find the parameter α where the formula is:

$$X_t = \alpha X_{t-1} + \varepsilon_t \quad (5.28)$$

This is a simple regression of *X* with itself, a step behind. ε represents Gaussian white-noise. The theoretical Fourier power spectrum for the AR(1) process is defined as [488]:

$$P_k = \frac{1 - \alpha^2}{1 - 2\alpha\cos\big(2\pi f(k)\big) + \alpha^2} \quad (5.29)$$

Where *k* is the index for each scale and *f(k)* represents the frequency of *k,* which is one over the scale adjusted by the Fourier factor as defined above. To find significant values the chi-squared distribution is invoked. For the Morlet wavelet, two degrees of freedom (DOF) is appropriate due to the real and imaginary parts, and thus for a 95% significance level the quantile function for $\chi_2^2 \approx 5.99$. As in Torrence and Compo [487], the wavelet power over the variance (σ^2) is divided by $\frac{P_k \chi_2^2}{2}$ (halving removes the DOF factor for complex wavelets) to get the values for ascertaining significance. These final values will be greater than one where the wavelet power is significant at 95%.

Figure 5-11 shows the plot of the CWT for example signals 1 and 2. These plots show the wavelet power, which can be interpreted as the degree of change in that position at that scale in the original series. If the signals had been completely flat, the plots would have been a solid dark blue. The solid red region through the length of the CWT graph of example 2 shows the periodicity of the sine wave that was added to the noise across all positions. However, it should be noted that even example 1, which is purely noise, shows significant regions purely by chance.

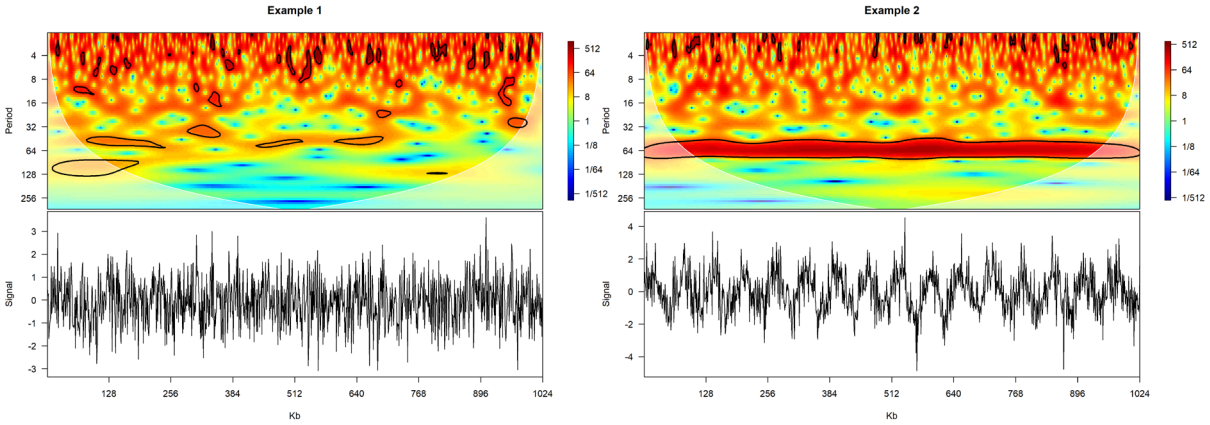


Figure 5-11 CWT of examples 1 and 2

These plots show the continuous wavelet transform for examples 1 and 2, with the original signals plotted below. The colours represent the wavelet power at that location and scale. The higher the power, the more energy the original signal had at that scale and location. Low power is coloured in blue (if the original signal had been completely flat, the plot would be a solid dark blue) and high power in red. Regions significant at the 95% level are surrounded by a black contour line. Regions within the cone of influence, i.e. affected by edge effects and thus should be discarded, are shown in the shaded white areas. The CWT for example 3 can be found in Appendix A.1.3.

5.2.6 Wavelet coherence

Wavelet coherence can be thought of as the correlation between the wavelet coefficients of two different signals, assessed over all positions and scales. Again, the R package biwavelet was used for the analysis and plots. The formula for wavelet coherence between two signals X and Y is as follows:

$$R^2(\lambda,t) = \frac{\left|S\left(\lambda^{-1}W^{XY}(\lambda,t)\right)\right|^2}{S(\lambda^{-1}|W^X(\lambda,t)|^2)\cdot S(\lambda^{-1}|W^Y(\lambda,t)|^2)} \quad (5.30)$$

where $W^{XY} = W^X W^{Y*}$ (where * is the complex conjugate), and *S* is a smoothing function. Smoothing is achieved across both position and scale by taking a weighted moving average, by way of convolution - that is by using a fast Fourier transform, applying a filter, and then employing an inverse fast Fourier transform. For smoothing across positions, the applicable filter for the Morlet wavelet transform is $e^{-\frac{t^2}{2\lambda^2}}$, see Torrence and Webster [489]. For smoothing across scales, the appropriate filter is a boxcar function of width 0.6 [483], where the width is the scale averaging factor for the Morlet wavelet with parameter $\omega_0 = 6$ as defined in Torrence and Compo [487]. Note that the $R^2(\lambda,t)$ equation in (5.30) is very similar to the standard squared correlation coefficient formula.

A Monte Carlo method is used to find the significance of the coherence. To do this, two new signals are simulated that have the same AR(1) α coefficients as the two original signals as calculated in equation (5.28). For both new signals the CWT is performed and then the coherence between them is calculated. This is repeated 1,000 times, as recommended in Grinsted *et al.* [483], and then for each position and scale the 95th percentile is found across simulations. Any coherence values from the original signals that are higher than this 95th percentile value are considered to be significant.

Figure 5-12 shows the wavelet coherence plots for each pair of example signals. The top two plots both involve example 1, which was just noise, whereas the bottom plot shows the coherence between the two signals with a sine wave. It is clear that the coherence between the examples with the wave is highest at and around the 64 Kb scale - aligning with the period of the wave. This high coherence persists across all positions, as expected. As in the CWT graphs, there are regions in all three plots that are highlighted as significant that exist purely by chance. Thus, it is important when interpreting these graphs to identify strong, large regions with a real-world physical explanation.

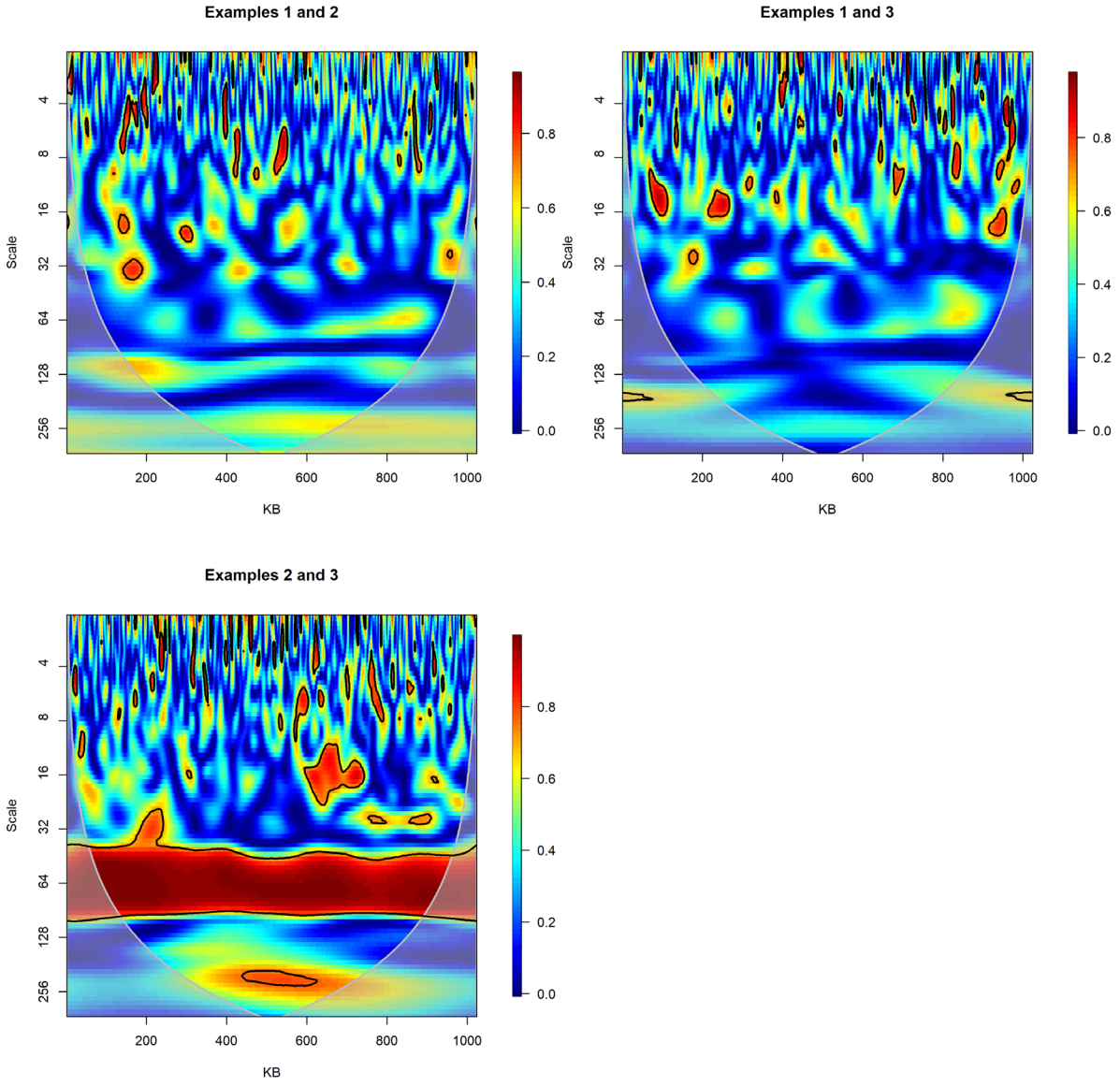


Figure 5-12 Wavelet coherence

These figures show the wavelet coherence between each pair of example signals. The coherence can be thought of as the squared correlation, values near 1, in red, imply signals have similar changes at that position and scale. The top two graphs compare the examples with a sine wave (examples 2 and 3) with a signal that is purely noise (example 1). The bottom graph, comparing both signals with a sine wave, clearly shows a high coherence at the 64 Kb, scale corresponding to the period of the sine wave. Regions enclosed in a black contour line are significant at the 5% level. The COI is shaded as in Figure 5-11.

This example has shown how to interpret wavelet analysis with three datasets containing noise and a repeating pattern. However, signals originating from the real world are likely to contain many more features such as large changes that either immediately or eventually return to

baseline, changes that persist, inconsistent cycles, and outliers. These can all be identified using the range of wavelet techniques described here. For example, for changes that last for different lengths of time, see Dong *et al.* [490] whose wavelet analysis identified two historical droughts that had different lengths and thus appeared on different scales. Cazelles *et al.* [491] explored inconsistent cycle lengths with simulated data and went on to look at real life signals in grouse populations where the periodicity indicated by the power spectrum was limited to only discrete regions of the signal. Wavelet analysis is a powerful method for analysing signals as underlying periodic trends, regions with long term changes, and regions with shorter high impact changes, can all be identified as part of the same analysis by considering all scales simultaneously.

5.3 Methods

The code for this chapter can be found at https://github.com/chorscroft/PhD-Thesis/tree/main/Chapter_5.

All wavelet analysis was undertaken using the R programming software v3.3.0 [492], using the IRIDIS high performance computing facility at the University of Southampton. The R packages used were intervals v0.15.1 [493], wmtsa v2.0-3 [475] and biwavelet v0.20.17 [481].

The output from LDhat for the four datasets in Chapter 4 were used: CEU, Wellderly, Baganda and Zulu. The datasets were trimmed to the region chr22:20000428-51218377 to avoid missing data in any of the datasets at either end.

The data were transformed from $\rho = 4N_e r$ per Kb into centimorgans per megabase (cM/Mb). This could have been achieved by using an estimate of N_e, for example of 10,000 as in Takahata [185] and Auton and McVean [450]; however, these numbers are inconsistent in the literature, and may not be applicable to both European and African datasets [187, 188]. Another way to do the conversion is to set the map length equal to another source and convert that way [438]. This is the method that was applied, using a map length of 60.67105 cM from the Bherer dataset, recalculated for just the section of the chromosome being analysed here. The ρ values were cumulatively summed, weighting for the distance between SNPs, and then proportioned based on the map length to get cumulative cM. This was then converted to cM/Mb based on the distances.

The data were then binned into 1 Kb bins using the code provided by Bherer *et al.* [406], removing any gaps bigger than 50 Kb. Finally, the datasets were merged and the longest section of chromosome 22 where all four datasets had complete data was found, and the centre of that region where the number of bins was equal to a power of two was selected to be analysed (chr22:29527000-45911000).

The analysis then followed the same format as in the introduction to wavelets section of this chapter above. The MODWT was used rather than the DWT to avoid the problems discussed in section 5.2.3.

5.4 Results

The four recombination datasets from the previous chapter (CEU, Wellderly, Baganda and Zulu) were decomposed using the Haar wavelet MODWT as described in section 5.2.4. The power spectrums for each decomposition were calculated and were plotted in Figure 5-13. As expected, the majority of the variance is contained in the smaller scales, due to the large, narrow peaks around recombination hotspots. As noted in Auton [494], this is not surprising and in order to explore the recombination rate by different scales, the log of the recombination rate may be more appropriate. The DWT of the $\log_{10}$ transformed data can be seen in Figure 5-14. The rough peak around the 16-64 Kb range is perhaps expected given there is, on average, around one recombination hotspot every 50 Kb in the human genome [136].

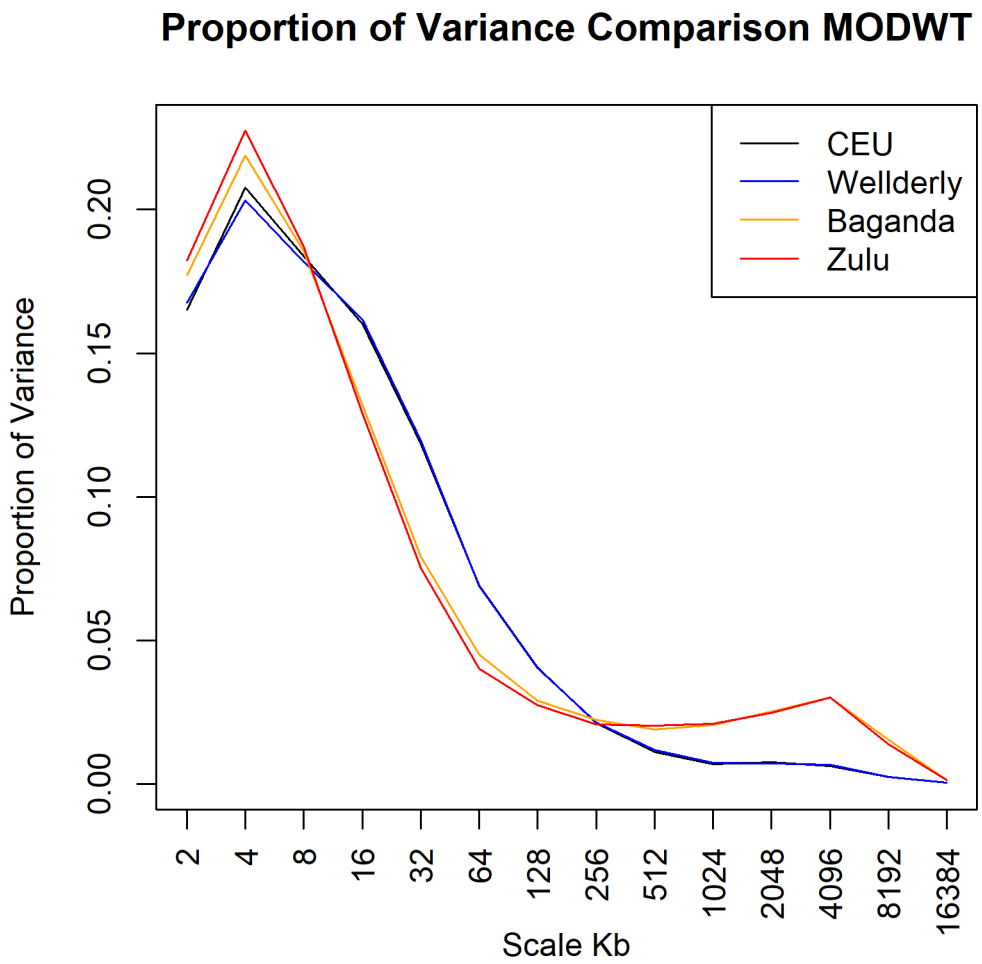


Figure 5-13 Power spectrums for recombination rates

This graph shows the proportion of variance from the original signal at each scale for each of the four datasets: CEU, Wellderly, Baganda and Zulu. The graphs have a very similar shape, all of them peaking at the finer scales.

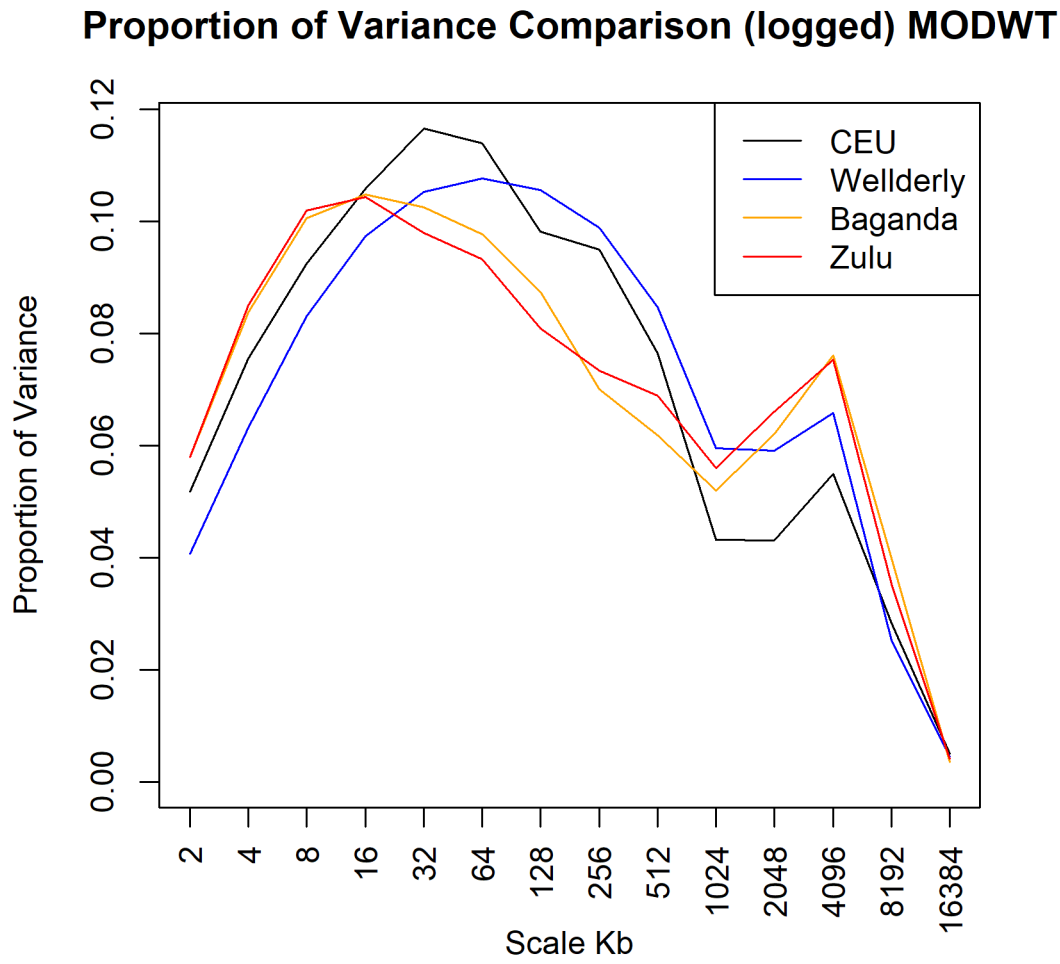


Figure 5-14 Power spectrums for log-transformed recombination rates

The graph shows the proportion of variance in the log-transformed recombination rates for each of the four datasets: CEU, Wellderly, Baganda and Zulu. The graph shows all show a similar pattern, peaking around the medium scales.

To check that the results are consistent across the chromosome and not just in this central region, the analysis was repeated using the 16,384 Kb regions at the extreme ends of the section of chromosome 22 with complete data across all datasets. The wavelet coefficients were different, as to be expected, however the power spectrums were similar enough to give confidence that the region being analysed is representative.

Correlations were calculated between the detail coefficients of each dataset, as described in section 5.2.4. Approximation coefficients are proportional to the averages and detail coefficients describe the changes in the signal, thus correlation between detail coefficients of two signals implies that as one signal changes at that scale, as does the other. As the data were not normally distributed, Kendall's Tau was the appropriate measure of correlation. Figure 5-15 shows the correlations between each pair of (log-transformed) datasets at each scale.

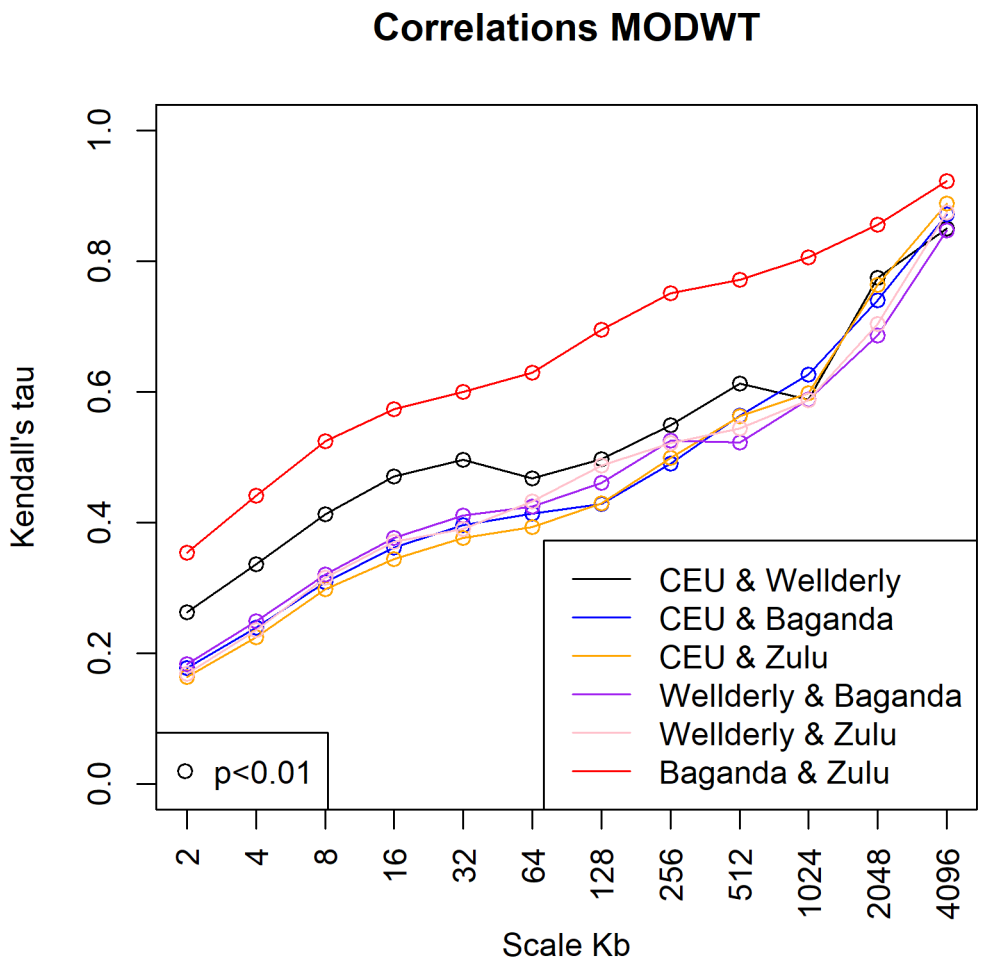


Figure 5-15 Correlations between detail coefficients

This graph shows the correlations between the detail coefficients of the DWT of the $\log_{10}$ transformed datasets. Circles indicate statistically significant correlations at the 0.01 level. All graphs show an upward trend as the scale gets wider. The African pair (Baganda and Zulu in red) are more correlated with each other than all the other pairings at every scale, whereas the European pair (CEU and Wellderly in black) are more correlated with each other than the other pairings at only the fine- to mid-scale.

The figure shows an upward trend for each pairing – that is, the wider the scale, the higher the correlation between the detail coefficients. This could be because recombination tends to cluster leaving cold spots in between, whereas the precise location of hotspots can be mobile over generations [428]. The African datasets are more correlated with each other on every scale than any other pairing, which was unexpected. Using the correlations between the European pair (CEU and Wellderly in black) as a baseline, it is clear that the European and African pairs differ mostly at

the fine scale, and as the scales increase the correlations become very similar to what would be expected for two samples from the same (European) population.

The datasets were then decomposed using the Morlet wavelet CWT with parameter $\omega_0 = 6$ as described in section 5.2.5. Figure 5-16 shows the CWT decomposition for each of the four datasets. The signals from the log-transformed data are also shown under each of the graphs. The colours on the graphs represent the magnitude of the detail coefficients – for example if the recombination rate was flat for the whole region, the graph would be entirely blue due to there being no changes across the region at any scale. The black lines indicate areas which are significant at the 5% level using a χ^2 test when compared to a red noise null power spectrum modelled using an autoregressive process with a lag of 1. The graphs show that the significant power differences are found in the narrowest scales and the widest scales. The narrow scales also contain most of the blue, low power areas, as can be seen more clearly in the zoomed-in view in Figure 5-17. These regions of static rates could explain why the proportion of variance in the power spectrum is lower for the narrower scales, even though there are significantly high regions too, which can be observed most clearly in the European zoomed-in graphs.

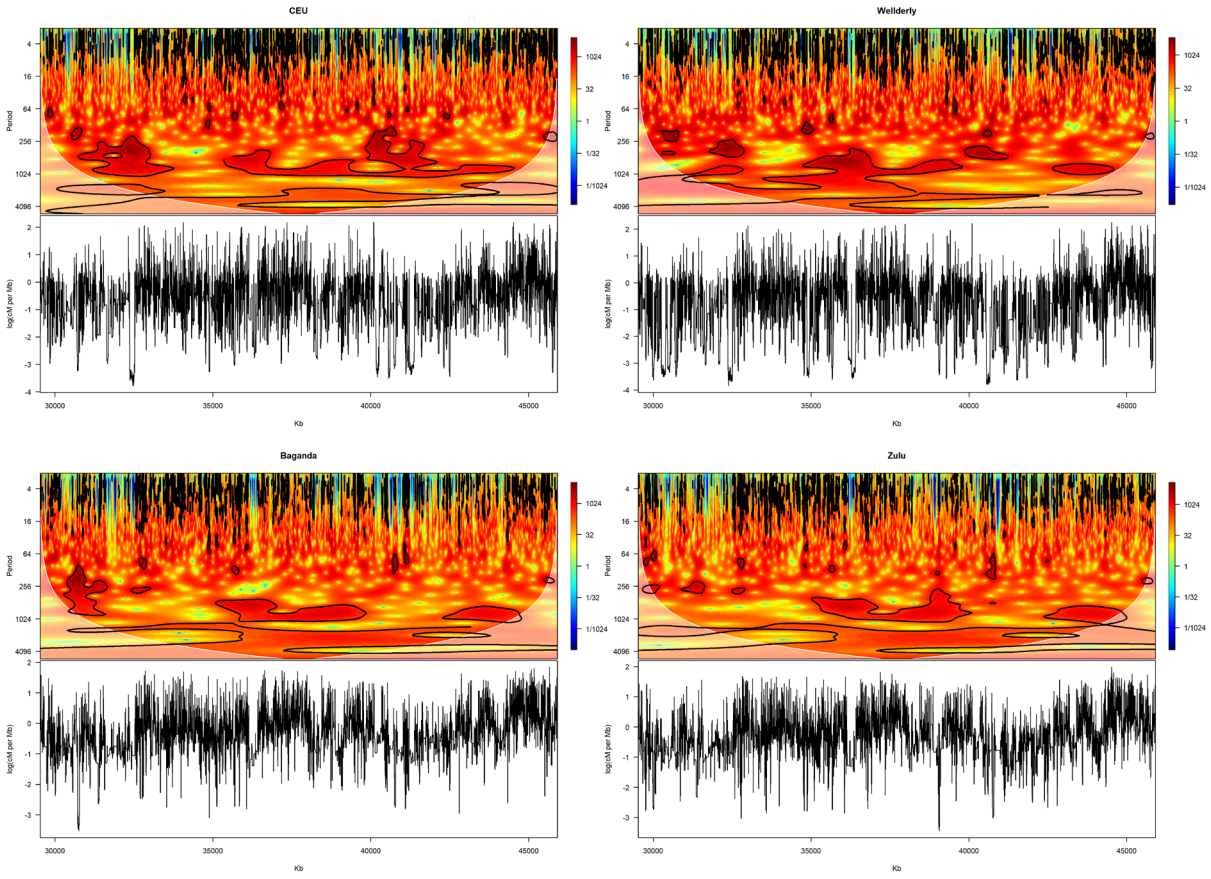


Figure 5-16 Continuous wavelet transforms

The CWT of the four log-transformed datasets showing the CWT and the original signal below. The colours indicate the magnitude of the detail coefficient – a flat, constant recombination rate would be dark blue. The black outlined areas are significant at the 5% level. The white shaded areas are under the cone of influence, where data may be affected by edge effects.

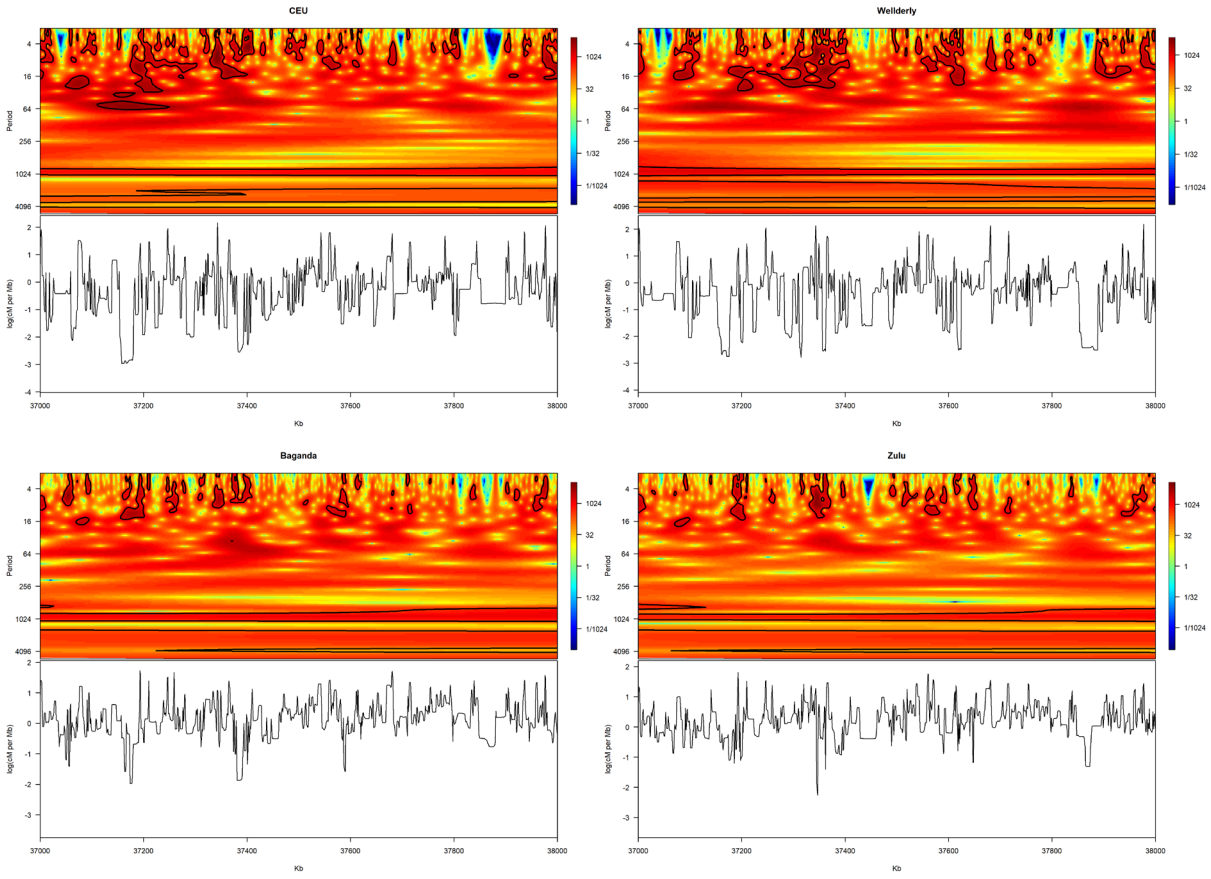


Figure 5-17 A close-up of the CWT for each dataset

These figures show the CWT for the four datasets as before, but zoomed-in to only show the 1 Mb region on chromosome 22 between 37,000,000-38,000,000 bp.

Wavelet coherence analysis assesses whether the correlations that were seen in Figure 5-15 are global across the whole region or whether there are differences in the correlation not only for multiple scales but also at different points along the chromosome. It works by calculating a smoothed correlation coefficient as defined in Torrence and Webster [489] using the CWT previously calculated. To test for significance, a Monte Carlo method was applied, using 1,000 randomisations as recommended in Grinsted *et al.* [483]. The results of the wavelet coherence analysis for each pair of datasets can be seen in Figure 5-18. The first striking thing is the amount of red on the Baganda and Zulu graph, indicating a strong correlation across most of the region and on medium to wide scales. For the rest of the scales, there are pockets of strong correlation and pockets containing little correlation with relatively little middle ground, implying the correlations are not consistent across the region. In contrast to this, the graph of the two European datasets (CEU and Wellderly) has many more blue regions especially in the medium scales (64 to 256) and a large area at the 1-2 Mb scale around the 41,000 Kb mark. The graphs comparing the European and African datasets all show a lack of correlation between the changes in the two signals at scales of around 64 Kb and less, with more correlation at the wider scales –

though noting a lot of the red areas are below the cone of influence. There is an interesting band of low correlation across most of the European and African comparison graphs around the 2 Mb scale which is not present when comparing the two African populations, with no obvious cause. Checks for replication on other chromosomes should be carried out before assigning any biological meaning to this phenomenon. The highly correlated red area on the CEU and Wellderly graph from the 256 to 1024 Kb scale around 32.5 Mb is not replicated in the graphs comparing these datasets to either of the African datasets. This is due to a roughly 150 Kb wide cold spot present in both European datasets but not in the African data, as shown in Figure 5-19. A CNV has been observed that almost exactly corresponds with this cold spot; however, it is unclear if this is the cause [495, 496].

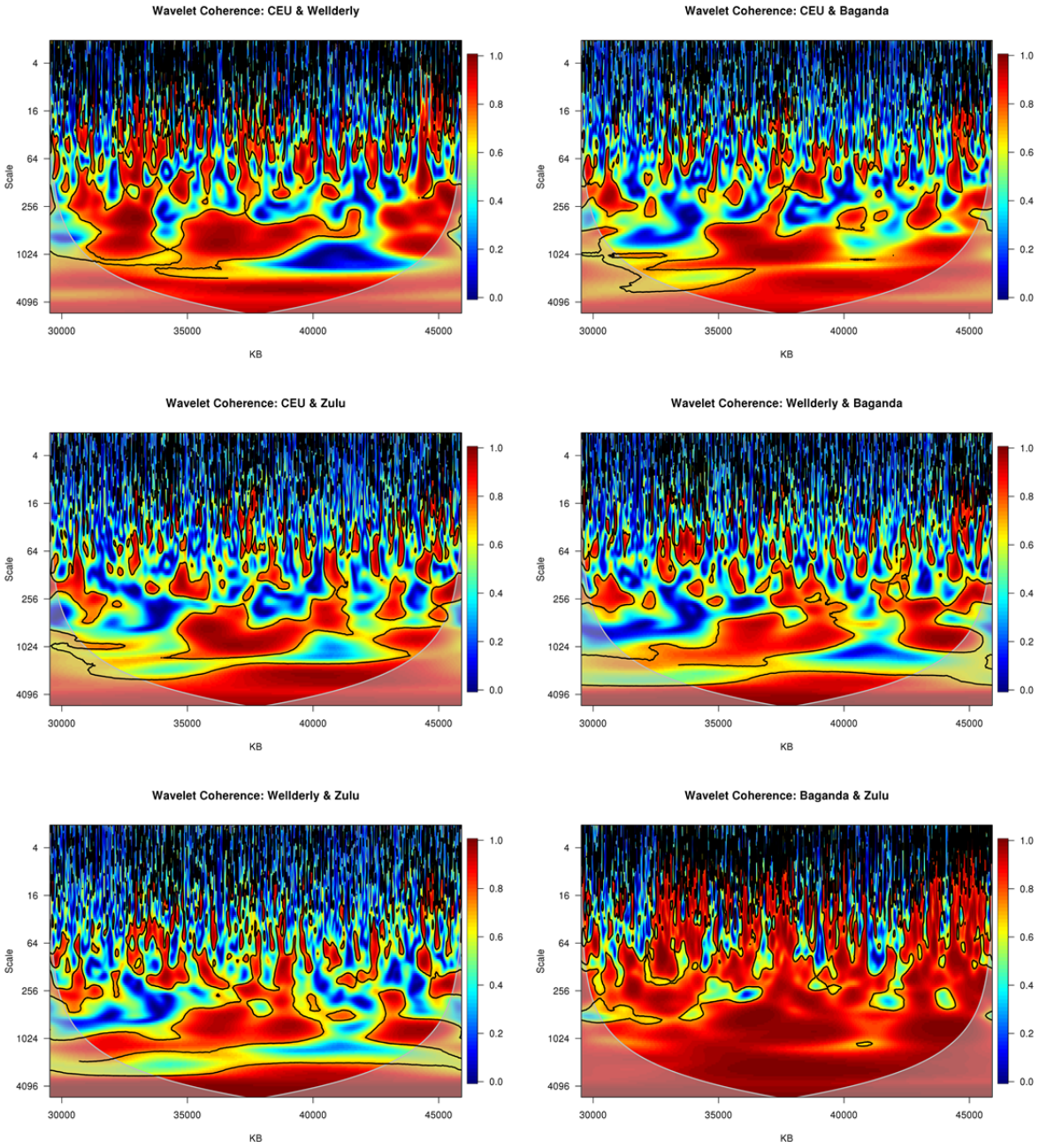


Figure 5-18 Wavelet coherence between each pair of datasets

Wavelet coherence analysis applied to each pair of recombination rate datasets. Red colours indicate higher correlation between the datasets at that location and scale. The cone of influence is shown in white. The black lines border regions of the graph at 5% significance.

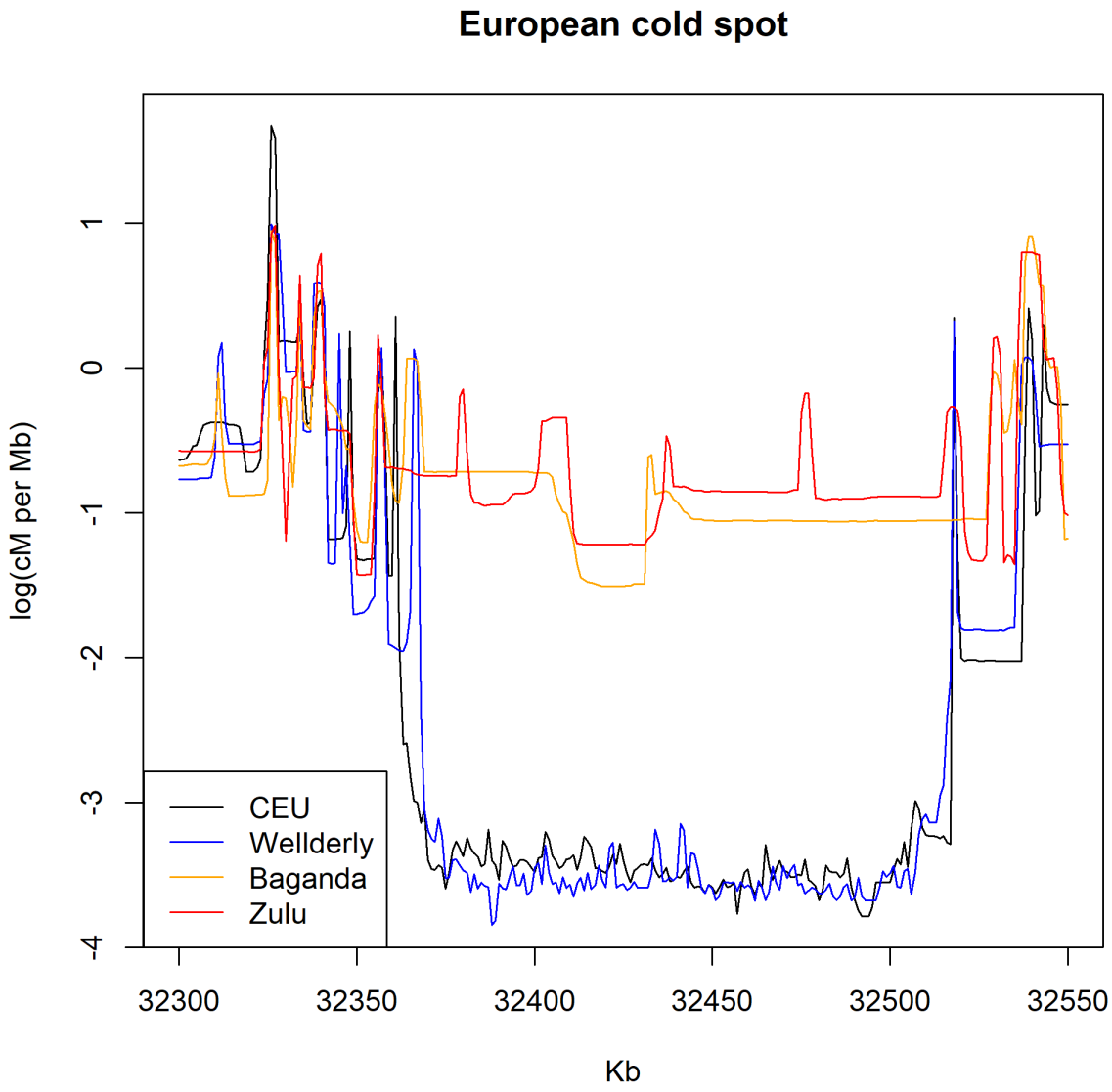


Figure 5-19 European cold spot

The data for the four datasets in $\log_{10}$ cM per Mb along chromosome 22. This shows a cold spot for recombination activity between 32.3 Mb and 32.55 Mb in the European data (black and blue lines), but not in the African data (red and orange lines).

5.5 Discussion

The wavelet analysis in this chapter has shown that it is possible to compare trends in recombination rates across a chromosome between different human populations from the fine scale through to the wide scale simultaneously. The change in recombination rates was shown to differ across populations, especially at the fine scale. While each dataset individually showed a

similar trend across scales, comparing across the chromosome showed rates are not uniform and are not consistent across populations.

The most variation in recombination rates was contained within the mid-scales for each of the four datasets, after logging. This is consistent with the estimate of one hotspot per 50 Kb previously published in Myers *et al.* [136]. The results are also similar to the power spectrums published in Bherer *et al.* [406] (whole genome) and Spencer *et al*. [472] (region of chromosome 20), although neither of these show the spike at the 4 Mb scale in this analysis, which implies this is probably due to a random quirk in this region of chromosome 22. The proportion of variance in the African datasets peaked slightly more towards the fine scale, perhaps reflecting that the African genome has more potential recombination locations than the European genome, and thus hotspots, while less intense, are more common [102, 418]. In all datasets there were regions of much change in recombination rate, and regions of very little change. This indicates the presence of recombination deserts - small regions with little recombination activity and few, low-intensity hotspots [136].

When comparing the datasets to each other, it was clear the changes at the wide scale were more correlated than those at the fine scale. This reflects the knowledge about broad recombination patterns, such as recombination being more frequent towards the telomeres and supressed around the centromere [136]. The finer the scale, the less correlated the changes in rate became for all comparisons. This is because recombination hotspots are highly mobile [428]. The comparison between two *D. melanogaster* populations in Chan *et al.* [443] shows a similar increase in correlation between the populations as the scale size increases. The biggest differences in correlations between European and African populations when compared to baseline were seen at the fine scale. Looking across the chromosome in the coherence analysis, it was clear there was little uniformity in correlations across the length of the region at most scales.

It was somewhat surprising that the African datasets were more correlated to each other than the European data sets were, over every scale. It is known that there is far more genetic diversity in Africa than there is in Europe [183]. The higher correlation could be because the Baganda and Zulu populations are more closely related than the two European datasets: they both speak languages from the Bantoid family, a subgroup of the Niger-Congo languages [103], and other research using LDMAPs has shown the two populations are much more closely related than to Ethiopians [497, 498]. However, the so-called “Bantu expansion”, where West African Bantu-speaking people spread west and south across the continent, happened around 5,600 years ago, integrating with the people who had been there for much longer [499]. Therefore, while their languages may have a fairly recent common ancestor, an estimate of the split between the Baganda and Zulu populations is around 29,000 years ago [500], so this is unlikely to be the cause.

Another theory could be that the European pairing has lower correlations due to selection. The OOA event and subsequent spreading across the Eurasian continent brought huge selective pressures as new environments, diets and climates were encountered [184]. Hotspots will move around due to selective pressure in locations where it would be beneficial for alleles along a chromosome to be separated [420, 501]. However, the African populations will have experienced their own set of selective pressures and so this may not explain the difference. Regardless of the reasons for the difference, it has been shown there is a difference between African and European recombination rates. Therefore, it is important when using recombination maps for analysis, for example when searching for regions under selection, to use maps made for the specific population being analysed. This is especially true when analysing genetic data on the fine scale.

Previous work has been undertaken to describe the differences between recombination rates of human populations, although none appear to have used wavelet methods. The work of Hinch *et al*. [418] on African American recombination maps reinforces the conclusions drawn here regarding similarity on the very wide scale, and differences on at the very fine scale. They conclude that variations in *PRDM9* are the driving force behind the differences in hotspot location. Looking further back into the evolution of *Homo sapiens*, Lesecque *et al*. [429] considered the recombination differences between modern humans and Denisovans, whose lineages are thought to have split around 750 kya [502], and discovered that they share few hotspots, even though the motif associated with hotspots in modern humans was already being actively targeted by the PRDM9 protein before the split. Even further back in evolutionary history, humans and chimpanzees (*Pan troglodytes*) share very few hotspots, with PRDM9 protein binding motifs known to be active in humans found to be inactive in chimpanzees, despite the genetic similarities between species [187, 427]. This work has highlighted the *PRDM9* gene as being an interesting source of study itself due to the selective pressure it has been under during the evolution of hotspots in humans and most other mammals, but yet is absent or inactive in other species such as dogs (*Canis familiaris*) and chickens (*Gallus gallus*) [419].

This study has shown that wavelet analysis is a useful tool for examining genetic data. The results are consistent with other studies and are biologically interpretable. Further work to examine the patterns on other chromosomes would be interesting, as well as using larger datasets of people from different populations. Bigger datasets with high coverage for more populations of people would be incredibly useful for understanding genetic variation and diversity, and is especially relevant in healthcare and to disease-risk, which can be very different in different populations [503]. The conclusion to be drawn from this work is that recombination rates are constantly evolving, due to the self-destructive nature of hotspots. It is therefore of vital importance when studying recombination and any process affected by recombination, for example detecting

selective events, that the population sample being studied are all indeed from the same population, and that recombination rate estimates for that specific population are used.

Chapter 6 Gene density and effective bottleneck time

6.1 Introduction

Researchers interested in identifying regions under selection in the genome can look to linkage disequilibrium (LD) patterns for clues, as changes in LD along a chromosome can be indicative of both positive and purifying selection in a population [263, 504]. Regions of reduced SNP diversity, high SNP correlation and thus higher LD than background levels all indicate that a selective event has occurred, or is occurring, in the region [505]. At the chromosomal level, LD maps can be created to ascertain LD across the whole chromosome, and then different chromosomes can be compared. Where some chromosomes have shorter LD maps than others, this implies more selection activity on these chromosomes [407].

When populations experience a bottleneck, LD increases as some haplotypes and variants are lost from the population. It then takes many generations for the LD caused by the bottleneck to decline, even after the population returns to its original size. An estimate of the time since the bottleneck occurred can be estimated by comparing the amount of LD in the genome with the frequency of recombination. If there is more LD than expected given the recombination activity in the genome, this is indicative of a bottleneck. If the population has experienced multiple successive bottlenecks, then the effect of the bottlenecks on LD will be compounded.

Effective bottleneck time (EBT) is a single value that describes the time in generations since a population had a bottleneck [442]. As populations can have many bottlenecks, this value will not necessarily align to a known individual bottleneck. It will instead represent the accumulation of bottlenecks over time, and hence is qualified with the word “effective”. This is analogous to the effective population size N_e, a common term used in population genetics, described in detail in section 1.10. The effective population size describes the size of the population in terms of its genetic diversity, and not necessarily its actual census size. Likewise, the effective bottleneck time describes the time in generations since a bottleneck in terms of the amount of LD in the genome. If there has only been one bottleneck, the EBT should align to that time; however, if there has been multiple bottlenecks the EBT would be comparatively shorter, as if there had been one recent bottleneck that would explain the mismatch between LD and recombination.

The EBT for a population is defined as the ratio of the LD map in LD units (LDU) and the length in morgans (M) [442]. Morgans are a unit of genetic distance, where 0.01 M = 1 cM. A distance of one morgan between two SNPs means that on average one crossover is expected to occur per meiosis between these SNPs. The intuition for the EBT comes from the Malécot equation and the

definition of linkage disequilibrium units (LDU) as defined in equations (4.2) and (4.3). LDU are calculated in terms of ε and d, as d is directly observable as the distance between SNPs and ε can be estimated, see section 4.2.4. However, this term in the original equation was θt, where θ is the frequency of recombination and t is the time in generations since LD began to breakdown from a previous bottleneck [147, 440, 442]. For a chromosome, the length in LDU is:

$$\sum_i \varepsilon_i d_i = \theta t \tag{6.1}$$

By using the genetic distance in morgans as an estimate for θ, and dividing both sides, an estimate for t can be ascertained. This means that the EBT for each chromosome can then be defined as:

$$EBT = \frac{LDU}{M} \tag{6.2}$$

LDU are smaller if there is linkage between SNPs. As bottlenecks cause an increase in linkage that this then broken down over time, the more recent and the more frequent a population has experienced a bottleneck, the smaller the LDU and thus the smaller the EBT.

While it would be tempting to assume that the EBT should be the same across all autosomes for a population, as they all have the same demographic history, this has been shown to not be the case [498]. Furthermore, this observation was much more pronounced in African populations than European populations. When EBT was first being investigated, only European data were considered, and so the chromosomal variation was not as apparent. This serves as a reminder of the importance to include diverse populations in human genomic studies [503]. The previous studies showed that the X-chromosome has a much smaller EBT than the autosomes in an ancestrally-European sample [407]. This is because the X-chromosome has increased levels of both purifying and positive selection, due to the hemizygosity of the chromosome in males [506]. While the autosomes are not hemizygous, the differing levels of selection experienced by individual chromosomes could potentially explain the differences in EBT.

It is now clear that both selection and bottlenecks can increase the amount of LD in the genome from expected levels. Selection affects LD on the local scale, whereas bottlenecks increase LD across the whole genome. By considering EBT on the chromosomal scale, the intersection of these factors can be considered. The EBT for a chromosome can be interpreted as the time in generations since LD began to decline to background levels, after the LD was increased by bottlenecks and selection events.

Gene density can be measured in number of genes per Mb. This can vary across the human genome, from an average of 2.8 genes per Mb on chromosome 13 up to 23.56 genes per Mb on

chromosome 19 [399, 507]. Gene density is a predictor for selection, as functional genic regions are likely to be the target of selection, although there are functional regions outside of genes, for example regulatory elements [504].

The aim of this study was to test the theory that EBT is negatively correlated with gene density due to purifying selection, by simulating genetic data. This was to be tested without any confounders that can be present in human genetic data, such as variable recombination rates, SNP densities, sample size and demography.

6.2 Methods

The code for this chapter can be found at https://github.com/chorscroft/PhD-Thesis/tree/main/Chapter_6.

The SLiM v3.3 simulation software [508] was used to generate populations of chromosomes with varying gene densities to test the hypothesis that EBT is negatively correlated with gene density. All simulated chromosomes had a length of 10 Mb. Each simulation was initialised with a population of 10,000 and was run for 5,000 generations. The overall mutation rate was set at 10^{-8} per site per generation [102] and the recombination rate was set at 10^{-8} per adjacent bases per generation.

16 different simulation models were created, each with a different gene density, ranging from 5 genes per Mb up to 20 genes per Mb. The genes were always 10,000 bp in length and were distributed uniformly across each chromosome. Mutations arising outside of these genic regions were always neutral and had no effect on fitness. Mutations within genes were either neutral or deleterious, with a 20% chance of neutrality. Deleterious mutations were all given a dominance coefficient of 0.5 and a fixed fitness effect of -0.03. The SLiM code can be found in Appendix A.2.

Samples of 100 individuals were taken from each simulation in VCF format. These files were then converted to .tped format and then cleaned by removing variants with minor allele frequency < 0.01 and Hardy-Weinberg equilibrium p-value < 0.001 using PLINK v1.90 [432].

To create the LD maps for these populations, the files were then passed to LDMAP to find the length of the maps in LDU [439]. This software is described in section 4.2.4. LDU describe the reduction in LD between SNPs to expected levels. SNPs that are highly correlated, i.e. in LD, will have a short LDU distance, because LD has not reduced to background levels yet. SNPs that are not correlated at all and are randomly associated will have a comparably large distance in LDU.

The length in morgans was equal to 0.1 M for each map. This is because the recombination rate was set to 10^{-8} per base, and with 10 Mb chromosomes: 10,000,000 bases * 10^{-8} = 0.1 expected

events per generation along each chromosome. This is equivalent to the rough estimate of 1 cM per Mb that is sometimes used for humans [509].

Further analysis and graphs were all created in R v3.6.0 [480]. Squared correlations were calculated using the linear model function in R and returning the r^2 value. A line of best fit was fitted similarly.

6.3 Results

Firstly, the LDU maps were created in LDMAP: one for each of the population samples with a gene density of 5 genes per Mb through to 20 genes per Mb. Figure 6-1 shows the LDU maps plotted together on one graph. As the gene density increases the total length of the map decreases. This is because the purifying selection was only present in the genic regions of chromosomes, and thus affected the populations with a denser gene density more than gene-sparse populations. The purifying selection increased the LD between SNPs, reducing the LDU between SNPs, and thus resulted in an overall shorter map. Gene-sparse populations experienced less purifying selection, so SNPs were less associated, meaning longer LDU distances and longer maps.

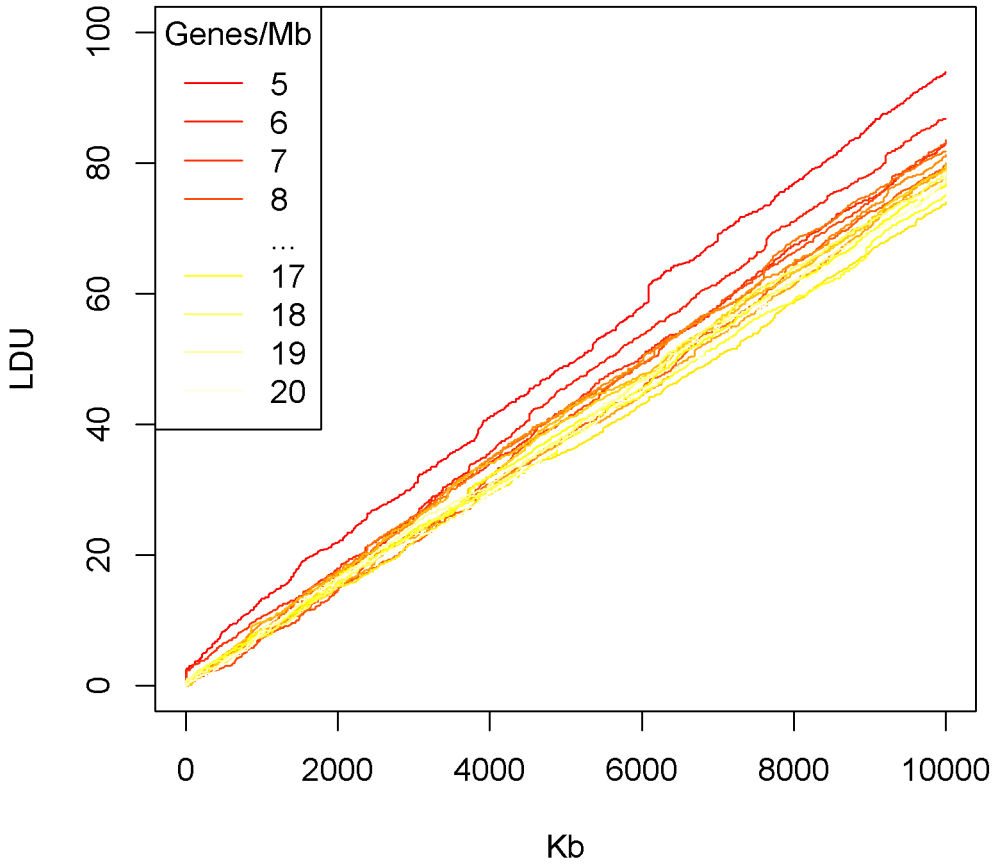


Figure 6-1 LDU maps of each of the simulations

This graph shows the LDMAP output for each of the samples taken from the simulation outputs. Each simulation was run with the same parameters, apart from the gene density, which ranged from 5 genes per Mb (red) through to 20 genes per Mb (yellow). LDU map length is longer when the gene density is sparser.

Effective bottleneck time was calculated for each of the models by taking the overall LDU map length and dividing by the distance in morgans, which was 0.1 M for each map. Figure 6-2 shows a plot of gene density against EBT. The figure shows a strong negative correlation (r^2 = 64%, p-value = 0.0002) between gene density and EBT. The correlation is imperfect due to the noise generated by the neutral mutations and the small sample size of individuals taken for each simulation.

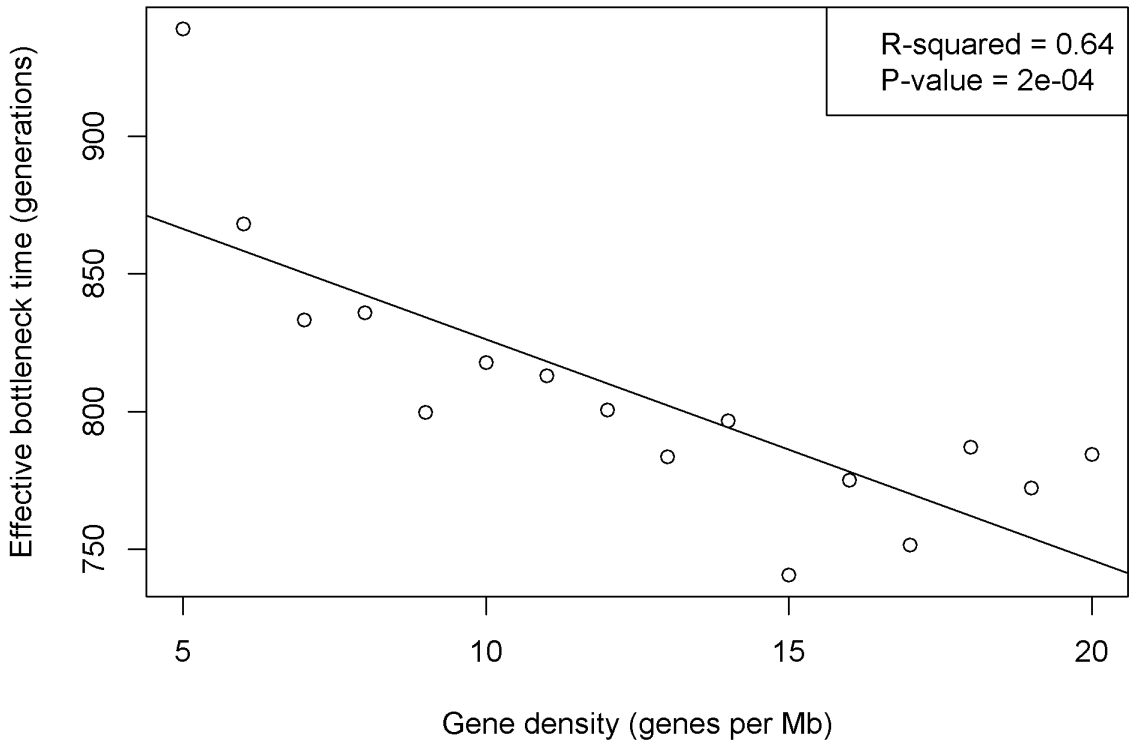


Figure 6-2 Gene density by effective bottleneck time

This figure shows the EBT for each model plotted against the gene density. The figure shows a strong negative correlation between gene density and EBT. The r^2 value and line of best fit included on the graph were calculated in R.

6.4 Discussion

The results showed evidence that EBT is negatively correlated with gene density, as hypothesised. LDU map lengths varied depending on the density of genes, even though recombination rates were held steady. For populations with chromosomes containing a higher density of genes, the maps were shorter, showing the effect of purifying selection on LD. Effective bottleneck times, measured in generations, were fewer for populations with gene-dense chromosomes as opposed to gene-sparse, which again illustrates the effect of selection.

This study was a successful proof of concept. The relationship between gene density and EBT has been established and shown. However, even with all the other variables held constant, the correlation was not perfect. This means correlations found in real-life data should be more compelling as there will be extra confounders, for example variable recombination rates. This study also strengthens arguments that any relationships found between EBT and gene density are

real and are not spurious due to different sample sizes or SNP densities, as mismatches in these can cause biases when using data from multiple sources [137].

Further work is now required to assess the EBT of different human chromosomes given the genetic density disparity between them, and to consider different human populations with different demographic histories and selective pressures, for example European and African populations. LDU maps have already been created for some African populations [498]. These maps are longer on every chromosome than maps made for Europeans, due to the shorter European history. This is in contrast to genetic maps built for African populations, where wide scale recombination patterns are almost identical to European maps, with most differences found at the fine scale [418]. Thus, EBT would be expected to be larger for African populations than European populations, consistent with what is known about human history [182].

This piece of work has shown that EBT is negatively correlated with gene density, as predicted, and has been a proof of concept so that analysis on real-life data with confounders can proceed with the knowledge that the basic assumption is sound.

Section 3 Identifying selection

In this section I investigate the Z_α statistic further, especially the extended family of Z_α statistics that allow for an adjustment based on the expected squared correlation between SNPs. In the first chapter I discuss the R package zalpha that I coded, documented, tested, and successfully submitted to CRAN (the Comprehensive R Archive Network). Submission to CRAN is a peer-reviewed process and means the package is easily discoverable and is straightforward to install. I would like to thank the reviewers for their valuable feedback on the zalpha package and on the paper announcing the zalpha package [510]. I would also like to thank the volunteers at CRAN for their time reviewing the software.

The final analysis chapter brings together all the knowledge and work from the previous chapters and applies it to the domestic dog. I clean the data, create the LD map and analyse it with wavelets, and then apply the Z_α suite of statistics to the data using the R package I created to find candidate regions for selective sweeps.

Chapter 7 Development of the zalpha package

7.1 Introduction

The Z_α statistic was introduced in Chapter 2, and in Chapter 3 was applied to both simulated and real-life data, where it performed well and warranted further investigation and use. Z_α comes from a family of statistics that were developed with the aim of exploiting LD fluctuations to identify regions under selection. Furthermore, some of the statistics also contain an adjustment for expected squared correlations in the genome. The aim of this is to increase the accuracy of Z_α by correcting for misleading relationships between alleles due to variable recombination rates across the genome. Z_α also has a sister statistic, Z_β, that can be used in conjunction with Z_α to determine if a sweep is in progress or near fixation. While the statistics had been published previously by Jacobs *et al.* [257], there was no publicly available software for applying them in a consistent and reproducible way.

This chapter concerns a new piece of software, the zalpha R package, that contains functions to facilitate the use of the Z_α statistics introduced in the paper by Jacobs *et al.* [257]. The R language is open source and is widely used for statistical and data analysis. R packages are containers for related functions for achieving a specific goal or performing a piece of analysis. The Comprehensive R Archive Network (CRAN) is an archive of R packages that have been submitted and are peer reviewed before they are accepted onto the network. A rigorous checking procedure is in place to make sure packages are tested and work correctly on different operating systems. Acceptance into CRAN means packages are easy to download and use, and package owners are automatically informed when any changes to R core may affect the functionality of the package. Having a package hosted on CRAN therefore means code is public, reproducible, supported, and legitimised through peer-review. The aim was to have the zalpha package accepted onto CRAN.

Simulated models were created to test that the zalpha package worked as expected. The aim of this was to show that the Z_α family of statistics can distinguish a region of the genome under selection from another region that is not under selection. A further step was to take outputs from the selected regions at different time points, to test whether the different statistics could differentiate between sweeps in progress and sweeps near fixation. Models with both uniform and variable recombination rate were considered, to show the benefit of adjusting for the latter.

The aim of this project was to create a package of statistics that would allow a researcher to easily apply the Z_α statistics to their data. The zalpha package version 0.1.0 was accepted onto CRAN on March 16th 2020, with the updated version 0.2.0 accepted on July 26th 2020. The development

version of the package is hosted on GitHub at https://github.com/chorscroft/zalpha. The next section of this chapter describes the Z_α statistics, and section 7.3 goes into the details of the zalpha R package itself. Sections 7.4 onwards contain details on the methods and the results of the simulation study.

7.2 The statistics

7.2.1 The Z_α statistic

The Z_α statistics are based on the idea that SNPs around the site of a selective sweep will be more correlated than SNPs in other parts of the genome where selection has not taken place. This is due to the hitchhiking effect; see section 1.9 for a fuller explanation. Figure 7-1A shows a population of chromosomes, where some alleles are perfectly correlated, and some are poorly correlated. The perfectly correlated alleles can be used to predict each other: if a chromosome contains a specific allele in the first position, the other positions can be accurately determined. Where a sweep has taken place, correlations between alleles are expected to be higher than in other regions of the genome. Z_α is calculated for each SNP and is defined as follows:

$$Z_\alpha = \frac{\binom{|L|}{2}^{-1} \sum_{i,j \in L} r_{i,j}^2 + \binom{|R|}{2}^{-1} \sum_{i,j \in R} r_{i,j}^2}{2} \qquad (7.1)$$

Where L and R are the set of SNPs to the left and right of the target SNP within a given window and r^2 is the squared correlation between a pair of SNPs. The notation $|L|$ and $|R|$ indicates the number of SNPs in each window, and the $\binom{n}{k}$ notion is shorthand for the binomial coefficient defined as:

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} \qquad (7.2)$$

Figure 7-1B shows a representation of the Z_α statistic calculated for a target SNP. The squared correlations are only calculated between pairs of SNPs on each side of the target SNP and are represented in the diagram as black circles. These symbolise the r^2 values that will be calculated and averaged in the Z_α statistic.

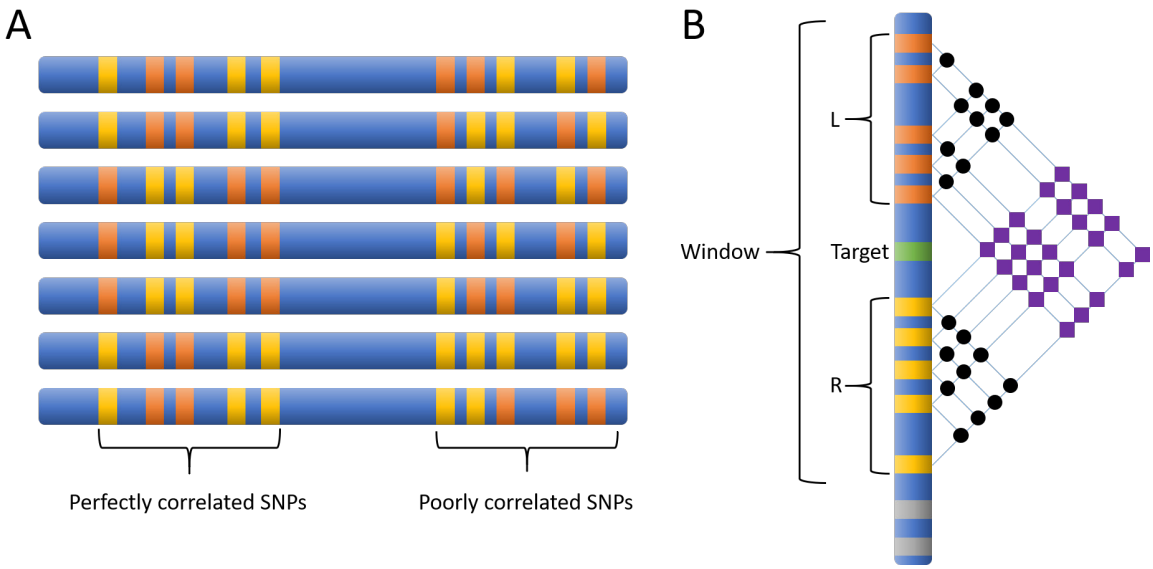


Figure 7-1 SNP correlation and Z_α visualisation

Diagram (A) shows a population of chromosomes where all the SNPs on the left are perfectly correlated and the SNPs on the right are poorly correlated. Given the information in the diagram, the colour of second SNP through to the fifth SNP can be perfectly predicted given the colour of the first SNP. Diagram (B) shows a representation of the Z_α statistic. For the target SNP in green, the window is defined and the SNPs falling within the window to the left and right are in sets L and R respectively. The black circles represent the squared correlations (r^2) between pairs of SNPs that will be used in the Z_α calculation. The purple squares represent the squared correlations considered by the Z_β statistic.

7.2.2 The Z_β statistic

The Z_β statistic is very similar to the Z_α statistic, as is defined as:

$$Z_\beta = \frac{\sum_{i \in L, j \in R} r_{i,j}^2}{|L||R|} \quad (7.3)$$

The difference is that the pairs of SNPs compared are always from different sides of the target SNP; see Figure 7-1B. The expected behaviour of the Z_β statistic is that when a sweep is in progress, it should increase just like Z_α. However, when the sweep nears fixation, the relationships between SNPs either side of a sweep break down, and Z_β will decrease. This is shown in Figure 2-1, which showed the lack of correlation across a beneficial mutation after fixation. Z_β can be used in conjunction with Z_α to infer whether a sweep is in progress or if it is near fixation. This is summarised in Table 7-1.

Table 7-1 Expected values of Z_α and Z_β given the stage of a sweep

Sweep progress	Expected Z_α values	Expected Z_β values
No Sweep	Low	Low
Mid-sweep	High	High
Nearing fixation	High	Low

The Z_α and Z_β statistics use the squared correlations between SNPs to find evidence of sweeps, where highly correlated regions imply a sweep may have occurred. However, recombination rates are variable across the genome, and this can cause regions to be highly correlated in the absence of any selective pressure. It would therefore be more useful to use Z_α and Z_β to find regions that are more highly correlated *than expected*, given the level of recombination in the region.

7.2.3 Z_α derived statistics

Z_α is the base for a family of related statistics that use the expected squared correlations between SNPs. These are defined and discussed below. By adjusting for expected squared correlations, Z_α can differentiate between regions that are highly correlated due to selection and those that are highly correlated because of a low level of recombination. To ascertain these expected squared correlations, a so-called LD profile is used. The LD profile is a look-up table that gives the expected squared correlation between SNPs given the genetic distance between them.

The idea behind this is that the correlation between SNPs is directly related to the genetic distance between SNPs. To use one of the statistics that requires an LD profile, the user must also know the genetic distance between the SNPs in order to look-up the expected values. As genetic distances are continuous, the LD profile consists of bins, or intervals, within which the expected r^2 values are defined. For example, bins could be of size 0.0001 cM, and thus the first bin in the LD profile would contain information for pairs of SNPs where the genetic distance between them is between 0 and 0.0001 cM. There is an example LD profile included in the zalpha R package, see section 7.3.3.2.

From a calculational point of view, the LD profile can be derived in multiple ways. Ideally, it would be calculated from a simulation of the population where every parameter is the same as the target population, but without selection. Failing this, using an independent data source of individuals from the same population would also be preferable. Finally, another acceptable way to calculate the LD profile would be to use the sample being analysed, as long as the bins for the genetics distances are large enough to have many observations so as to not be biased by outliers. The R package contains a function to aid in the creation of an LD profile. As well as the expected r^2

values, the standard deviation and the estimated Beta distribution parameters can also be calculated for each bin.

The use of LD profiles to adjust the statistics for expected squared correlations is the unique feature that sets the Z_α statistics apart from other statistics for identifying evidence of selection. The statistics will now be defined. The implementation of these statistics within the R package is discussed in the next section.

The formula for the first of these adjusted statistics is:

$$Z_\alpha^{E[r^2]} = \frac{\binom{|L|}{2}^{-1} \sum_{i,j \in L} E[r_{i,j}^2] + \binom{|R|}{2}^{-1} \sum_{i,j \in R} E[r_{i,j}^2]}{2} \quad (7.4)$$

This formula contains the term $E[r_{i,j}^2]$. This is the expected squared correlation between SNPs *i* and *j* given the genetic distance between them. This information is ascertained from the LD profile. This statistic is the same as the Z_α statistic except for the use of expected squared correlations rather than observed squared correlations, and so can be thought of as the expected value of Z_α. While the other statistics defined below are stand-alone, $Z_\alpha^{E[r^2]}$ purely describes the expected value for Z_α given the expected r^2. Thus, for locating sweeps it should be used in combination with Z_α, for example by calculating $\frac{Z_\alpha}{Z_\alpha^{E[r^2]}}$, which will be greater than one when Z_α is large even after taking into account the expected squared correlations.

$$Z_\alpha^{\frac{r^2}{E[r^2]}} = \frac{\binom{|L|}{2}^{-1} \sum_{i,j \in L} \frac{r_{i,j}^2}{E[r_{i,j}^2]} + \binom{|R|}{2}^{-1} \sum_{i,j \in R} \frac{r_{i,j}^2}{E[r_{i,j}^2]}}{2} \quad (7.5)$$

This statistic includes the calculated observed squared correlation r^2 inside the statistic as well as the expected squared correlation and uses the ratio of the two as the measure of association between SNPs. Unlike the previous statistic that needs to be combined with Z_α, this statistic is stand-alone as it contains both the observed and expected r^2 values.

$$Z_\alpha^{\log\left(\frac{r^2}{E[r^2]}\right)} = \frac{\binom{|L|}{2}^{-1} \sum_{i,j \in L} log\left(\frac{r_{i,j}^2}{E[r_{i,j}^2]}\right) + \binom{|R|}{2}^{-1} \sum_{i,j \in R} log\left(\frac{r_{i,j}^2}{E[r_{i,j}^2]}\right)}{2} \quad (7.6)$$

This statistic uses the log (base 10) of the ratio of the squared correlation and expected squared correlation between SNPs. This may be useful as the ratio of the squared correlation and expected squared correlation can be extreme, and thus the log-transformed value reduces the effect of outliers.

$$Z_{\alpha}^{ZScore} = \frac{\binom{|L|}{2}^{-1}\sum_{i,j\in L}\frac{r_{i,j}^2 - E[r_{i,j}^2]}{\sigma[r_{i,j}^2]} + \binom{|R|}{2}^{-1}\sum_{i,j\in R}\frac{r_{i,j}^2 - E[r_{i,j}^2]}{\sigma[r_{i,j}^2]}}{2} \quad (7.7)$$

This statistic uses the expected squared correlation and the standard deviation of the expected squared correlation. This assumes that squared correlations are normally distributed and calculates a standardised score. However, it is unlikely the r^2 values will be normally distributed, so the next statistic may be more appropriate.

$$Z_{\alpha}^{BetaCDF} = \frac{\binom{|L|}{2}^{-1}\sum_{i,j\in L}F(r_{i,j}^2; a, b) + \binom{|R|}{2}^{-1}\sum_{i,j\in R}F(r_{i,j}^2; a, b)}{2} \quad (7.8)$$

Where $F(r_{i,j}^2; a, b)$ is the cumulative distribution function of the beta distribution with parameters *a* and *b* as defined by the LD profile. This method essentially estimates p-values for the squared correlations between SNPs and averages these. This method may be more realistic than assuming a normal distribution as r^2 values are likely to resemble different distributions based on the genetic distance. For instance, where r^2 is expected to be high due to a small genetic distance the distribution may be negatively skewed, whereas where there is a large genetic distance the expected r^2 distribution may be positively skewed. The Beta distribution is a good fit as it can take many different shapes due to its two parameters, and it is only valid for values between zero and one, the same as r^2.

For each of the Z_{α} statistics involving adjustments for expected r^2, there is an equivalent Z_{β} statistic. The formulae for these are as follows:

$$Z_{\beta}^{E[r^2]} = \frac{\sum_{i\in L,j\in R}E[r_{i,j}^2]}{|L||R|} \quad (7.9)$$

$$Z_{\beta}^{\frac{r^2}{E[r^2]}} = \frac{\sum_{i\in L,j\in R}\frac{r_{i,j}^2}{E[r_{i,j}^2]}}{|L||R|} \quad (7.10)$$

$$Z_{\beta}^{\log\left(\frac{r^2}{E[r^2]}\right)} = \frac{\sum_{i\in L,j\in R}log\left(\frac{r_{i,j}^2}{E[r_{i,j}^2]}\right)}{|L||R|} \quad (7.11)$$

$$Z_{\beta}^{ZScore} = \frac{\sum_{i\in L,j\in R}\frac{r_{i,j}^2 - E[r_{i,j}^2]}{\sigma[r_{i,j}^2]}}{|L||R|} \quad (7.12)$$

$$Z_{\beta}^{BetaCDF} = \frac{\sum_{i\in L,j\in R}F(r_{i,j}^2; a, b)}{|L||R|} \quad (7.13)$$

Suggested ways of using the Z_β formulae to infer the stage of the sweep are $Z_\alpha - Z_\beta$ or $\frac{Z_\alpha}{Z_\beta}$, using the equivalent Z_α and Z_β formulae if adjusting for expected squared correlations. First, the Z_α statistic should be used to find regions with outlying Z_α values, and then the Z_β statistics can be used to ascertain the stage of the sweep, see Table 7-1.

Finally, the size of the sets L and R themselves can be interesting as regions of low SNP density can be indicative of a sweep. This is due to the hitchhiking effect increasing the frequency of one or a small number of haplotypes. This can result in some polymorphisms being removed from the population entirely as they either hitchhike along to fixation, or because they reside on less-favourable haplotypes that are removed from the population, see Figure 2-2 for an example of this [511]. There are two statistics included in the package for assessing this, collectively called the diversity statistics:

$$LplusR = \binom{|L|}{2} + \binom{|R|}{2} \tag{7.14}$$

$$LR = |L||R| \tag{7.15}$$

The diversity statistics are expected to be reduced around regions undergoing selection than in other regions of the genome. However, there are many other reasons that a region may be SNP-poor, for example regions that are hard to sequence due to repetitive elements, and thus should not be solely used as evidence.

7.3 The zalpha R package

The zalpha R package contains multiple elements, including functions, documentation, example datasets, and test files. These will all be explained in detail over the next few sections. The zalpha package itself can be found at https://github.com/chorscroft/zalpha and at https://cran.r-project.org/web/packages/zalpha/index.html.

7.3.1 Functions

The zalpha package contains sixteen main functions and four small helper functions. The sixteen main functions consist of: one function for creating an LD profile, two diversity statistics, six Z_α statistics, six Z_β statistics, and one function called Zalpha_all that will run all the statistical functions simultaneously.

7.3.1.1 Inputs

The inputs for the different statistical functions are described here. Table 7-2 shows which inputs are required for each of the statistics. Note that each chromosome of the genome should be run separately.

- pos: This a vector containing the physical positions of each of the SNPs, in bp or Kb
- ws: The window size that Z_α should be calculated over for each SNP. This should be given in the same units as were used in pos.
- x: A matrix containing a value for each SNP for each chromosome (column) in each position (row). The number of rows should be the same as the number of elements in pos, and the number of columns will be equal to 2*n* for *n* diploid individuals. The value can be anything, for example 0 and 1 or ACGTs, so long as each SNP is biallelic. Missing values should be coded as NA.
- dist: A vector containing the cumulative genetic distances for each SNP, for example in cM. This vector should be the same length as pos.
- minRandL: This is the minimum number of SNPs that must be present within the window on both the left and right sides of the target SNP. This is to stop small numbers of SNPs causing extreme squared correlations and biasing the result. By default, this is set to 4.
- minRL: This is the minimum product of the number of SNPs to the left and right of the target SNP within the window. As above, this is to avoid regions with a dearth of SNPs from skewing the results. By default, this is set to 25.
- X: The user can set the software to only return values within a given region. If a vector of size two is given for this parameter, only results within those limits will be returned. These values must be within the limits of pos.

The following inputs are all part of an LD profile. These should all be provided as individual vectors of the same length. A more detailed overview of how these can be calculated can be found in section 7.3.1.3.

- LDprofile_bins: The lower bounds for each of the bins in the LD profile. Each bin should be the same size.
- LDprofile_rsq: The expected r^2 value for the corresponding bin.
- LDprofile_sd: The standard deviation for the r^2 values in the corresponding bin.
- LDprofile_Beta_a: The first estimated parameter for the Beta distribution fitted for the corresponding bin.

- LDprofile_Beta_b: The second estimated parameter for the Beta distribution fitted for the corresponding bin.

Table 7-2 Inputs for each statistical function in the zalpha package

These are the inputs for each function. An X indicates an input is required, and an O means it is optional. If there is a default, it is given in brackets.

Function	Eqn	pos	ws	x	dist	LDprofile_ bins	LDprofile_ rsq	LDprofile_ sd	LDprofile_ Beta_a	LDprofile_ Beta_b	minRandL	minRL	X
Zalpha **Zbeta**	(7.1) (7.3)	X	X	X							O(4)	O(25)	O
Zalpha_expected **Zbeta_expected**	(7.4) (7.9)	X	X		X	X	X				O(4)	O(25)	O
Zalpha_rsq_over_expected **Zbeta_rsq_over_expected** **Zalpha_log_rsq_over_expected** **Zbeta_log_rsq_over_expected**	(7.5) (7.10) (7.6) (7.11)	X	X	X	X	X	X				O(4)	O(25)	O
Zalpha_Zscore **Zbeta_Zscore**	(7.7) (7.12)	X	X	X	X	X	X	X			O(4)	O(25)	O
Zalpha_BetaCDF **Zbeta_BetaCDF**	(7.8) (7.13)	X	X	X	X	X			X	X	O(4)	O(25)	O
L_plus_R **LR**	(7.14 (7.15)	X	X										O
Zalpha_all		X	X	O	O	O	O	O	O	O	O(4)	O(25)	O

7.3.1.2 Outputs

The output of each of the statistical functions is a list where the elements are:

- position: A vector containing the physical location of each of the SNPs in the output
- function: The name of this element is the name of the statistical function. This is a vector containing a value for each SNP in the position vector. If a value could not be calculated, the value will be NA. For the Zalpha_all function, there will be an element in the output list for each of the statistical functions that was calculated.

These outputs can then be saved, combined, plotted, and used for whatever analysis the user requires.

7.3.1.3 create_LDprofile

For statistics requiring an LD profile, the user can supply their own or they can use this function to generate one. The inputs required are as follows:

- dist: A vector containing the cumulative genetic distances for each SNP, for example in cM.
- x: A matrix containing a value for each SNP for each chromosome (column) in each position (row). The number of rows should be the same as the number of elements in dist, and the number of columns will be equal to 2*n* for *n* diploid individuals. The value can be anything, for example 0 and 1 or ACGTs, so long as each SNP is biallelic. Missing values should be coded as NA.
- bin_size: The size of each bin. This should be big enough that many pairs of SNPs will fall into each bin. This should be given in the same units as dist.
- max_dist: The maximum genetic distance to be considered. If this is not given it will default to the size of the largest distance in the distance vector. Recall that only relationships between SNPs within a given physical window size will be considered, so the max_dist should be large enough to cover a realistic genetic distance that could exist within a given window size to avoid redundancy.
- beta_params: If this is set to TRUE, a Beta distribution will be fitted to each of the bins. Default is set to false.

The first two inputs can optionally be supplied as two lists of the same length, where each element in the dist list is a vector that corresponds to the same element in the x list, which would be a matrix. This is designed so that the information for all chromosomes in the genome can be

used to create an LD profile. Each chromosome would have its own dist vector and x matrix, and thus each element of the lists will correspond to each chromosome.

The function works by finding the distances between each pair of SNPs in the dist vector. For each pair, it assigns them to a bin, for example if the distance was 0.0053 cM, and the bin_size was 0.001, this pair would be assigned to bin 0.005. The function calculates all the squared correlations (r^2) between each pair of SNPs using the x matrix. For each bin, using just the SNP-pairs assigned to it, the average r^2 value is calculated, as is the standard deviation of the r^2 values.

If the parameter beta_params was set to TRUE, a Beta distribution will be fitted to each of the bins. This is achieved using the function fitdist within the R package fitdistrplus [512]. This R package is only required if the beta_params parameter is set to TRUE. If the package is not present, the user will be informed, and the function will cease. If any of the r^2 values in the bin are exactly zero or one, the function will not be able to fit a Beta distribution as the distribution is only defined over the exclusive range (0,1). In this case, an adjustment is made to the data as described in Smithson and Verkuilen [513].

The output from this function is a data frame, containing six columns:

- bin: The lower bound for each bin.
- rsq: The mean r^2 value for the bin.
- sd: The standard deviation of the r^2 values for the bin.
- Beta_a: The first estimated parameter of the Beta distribution fitted for the bin. This will be NA if the beta_params parameter was set to FALSE.
- Beta_b: The second estimated parameter of the Beta distribution fitted for the bin. This will be NA if the beta_params parameter was set to FALSE.
- n: The number of pairs of SNPs that fell within this bin. This can be used to ascertain that the bin_size supplied was an appropriate size.

These columns can be used as inputs for the other relevant statistical functions by extracting them as vectors.

7.3.1.4 Helper functions

There are four helper functions contained inside the zalpha package that cannot be directly called by the user but instead are used internally to aid in calculations.

- lower_triangle
  This function returns the lower left-hand triangle of a matrix, not including the diagonal, as a vector. This function utilises the lower.tri base R function.

- equal_vector

  The use of LD profile bins with non-integer cutoffs, e.g. 0.0001 cM, means it is important to make sure that these numbers are compared correctly and not affected by floating-point errors. This function compares a vector to a number and returns true or false for each value in the vector that is equal to the number. The values are considered equal if the absolute difference is within a very small margin, given by the precision of the user's own machine.

- assign_bins

  This function assigns the genetic distance between two SNPs to a bin in the LD profile. The in-built base R floor and ceiling functions for rounding are ill-equipped for comparing numbers that may be affected by rounding errors. This function accounts for this by utilising the all.equal and isTRUE functions from base R.

- est_Beta_params

  The create_LDprofile function has the options to estimate Beta parameters for each bin. The function fitdist from the R package fitdistrplus [512] is called to do this. The fitdist function has the option of including starting parameters for the fitting algorithm. This function generates these starting parameters, using the mean and variance of the r^2 values in the bin. The Beta distribution has a defined mean and variance as:

$$E(X) = \bar{x} = \frac{a}{a+b} \tag{7.16}$$

$$Var(X) = \bar{S}^2 = \frac{ab}{(a+b+1)(a+b)^2} \tag{7.17}$$

  Rearranging these gives these estimates for parameters *a* and *b*, and these are the values returned by this helper function:

$$a = \bar{x}^2\left(\frac{1-\bar{x}}{\bar{S}^2} - \frac{1}{\bar{x}}\right) \tag{7.18}$$

$$b = a\left(\frac{1}{\bar{x}} - 1\right) \tag{7.19}$$

7.3.2 Documentation

The zalpha package is fully documented. Each function has a help file associated with it written in R documentation (.Rd) format, which is a markup language very similar to LaTeX. The roxygen2 package was used to automatically generate these help files [514]. Each help file contains:

- A description of the statistic that is being implemented by the function
- A list of the parameters required, whether they are optional, if they have a default (and what the default is if they have one), and what format they need to be in
- A description of the result of the function, including what it means and what format it is in

- A working example of the function using the included datasets
- A reference to the Jacobs *et al.* [257] paper
- Hyperlinks to other related functions, if relevant

These help files are readily accessible by R users using the standard help commands. Because the help files were written using the .Rd format, a manual can be easily generated as a PDF file using the devtools function build_manual [515].

A vignette is included in the package. This contains a worked example of how to use the package using the example datasets and fully functioning code. The vignette is written in R markdown and contains code segments between paragraphs of text.

The vignette and the manual can be found at https://github.com/chorscroft/zalpha and at https://cran.r-project.org/web/packages/zalpha/index.html.

7.3.3 Datasets

The package contains two example datasets “snps” and “LDprofile” that can be accessed by users and are utilised in the vignette and the examples section of the help files for each function.

7.3.3.1 snps

The snps dataset contains 20 SNPs on 10 simulated chromosomes. It is formatted with zeros and ones as the SNP values, which could represent reference and alternate alleles, or ancestral and derived alleles, for example. The dataset was created such that the SNPs nearer the centre are more correlated to each other than SNPs further away, so that the Z_α statistic is higher around the middle of the region. This dataset models the expected behavior of a selective sweep.

7.3.3.2 LDprofile

The LDprofile dataset is a small example of an LD profile. It contains 50 bins of size 0.0001 cM ranging from 0 to 0.0049 cM, with the following stats associated with each bin: expected r^2, standard deviation, and the parameters for a fitted Beta distribution.

7.3.4 Testing

The package has been thoroughly tested to make sure it works as expected. Each function contains checks that the inputs are in the correct format and will warn the user with a meaningful error message when inputs are incorrect. Each function has a test file associated with it to allow for consistent and thorough testing using the test_that function from the testthat R package [516]. All test files and documentation are publicly available on the package’s GitHub.

The development version of the package on GitHub has also been linked to the continuous integration service Travis CI [517]. This triggers tests and checks on the package every time new code is pushed to the GitHub repository. The status of the package (i.e. whether it has passed or failed the Travis CI checks) is publicly available on the package's GitHub page.

7.4 Simulation methods

The code for this section can be found at https://github.com/chorscroft/PhD-Thesis/tree/main/Chapter_7.

7.4.1 Simulations with uniform recombination rate

All simulations were performed using SLiM version 3.3 [508]. Two models were created: one neutral model without selection, and one with a selected sweep. 100 simulations were generated for each model. Table 7-3 shows the parameters for the two models.

Table 7-3 Simulation parameters for zalpha package test

Parameters	Neutral simulation	Simulation with selection
Size of chromosome	1 Mb	1 Mb
Population size	10,000	10,000
Recombination rate	1e-8	1e-8
Recombination distribution	Uniform	Uniform
Mutation rate	1e-7	1e-7
Mutation distribution	Uniform	Uniform
Mutation fitness effect	0	0
Beneficial mutation introduced at:	N/A	Generation: 1,000 Location: 500,000 bp Dominance coefficient: 0.5 Fitness effect: 0.05 (fixed)
Output	At generation 1,500: ➔ 1,000 randomly selected chromosomes. "Neutral" dataset	When the beneficial mutation frequency was 50%: ➔ 1,000 randomly selected chromosomes. "Mid-way" dataset When the beneficial mutation frequency was 90%: ➔ 1,000 randomly selected chromosomes. "End" dataset

Simulations where the beneficial mutation was lost to genetic drift were discarded. All SNPs with a minor allele frequency of less than 5% were removed from the final datasets.

The final three datasets, each consisting of 100 simulations of 1,000 chromosomes, were:

- Neutral: A model with no selection
- Mid-way: A model with a sweep in progress
- End: A model with a sweep nearing fixation

To generate an LD profile, a 5 Mb chromosome was simulated using the parameters for the neutral model. All relevant statistics were then calculated for distances from 0 to 2 cM in discrete bins of size 0.0001 cM.

The zalpha R package v0.1.0 was used to calculate the Z_α family of statistics. A window size of 200,000 bp was used. Once these statistics were calculated they were either used as is or combined with others based on analysis undertaken by Jacobs *et al*. [257]. This resulted in 35 final statistics for further analysis.

7.4.2 Aggregate graphs

Aggregate graphs were produced using R version 3.4.2 [389]. For each model, the statistics were binned into 10,000 bp intervals and the mean value of each statistic with 95% confidence intervals was calculated across all the simulation runs. These were then plotted in R, with each point plotted in the centre of each bin.

7.4.3 ROC curves

To create ROC curves, the 35 statistics were calculated for the centre 600 Kb region of each scenario, to avoid any edge effects at each end of the simulated chromosome. The maximum value of each statistic in the region was then ascertained, apart from $\binom{|L|}{2} + \binom{|R|}{2}$ and |L||R| where the minimum was used. These values were passed to the R package pROC v1.15 [395] to generate the ROC curves. Three comparisons were undertaken: neutral vs end, neutral vs mid-way and mid-way vs end. This was to show which statistics are effective at detecting regions under selection, and which are effective at differentiating between different stages of sweeps. As well as the graphs, the area under the curve (AUC) and partial area under the curve (pAUC) were returned for each ROC curve. The pAUC was set for specificity between 0.95 and 1 (corresponding to a false positive rate of 5% or lower). The package also calculated confidence intervals of 95% for both statistics using the recommended bootstrap method, with sensitivity steps set to 0.01. The AUC is a valid statistic to use for comparison as each model contains the same number of observations.

7.4.4 Simulations with variable recombination rate

To simulate a population with realistic recombination variance, a 1 Mb section of the 2008 HapMap recombination map was chosen. Specifically, the region between 10 Mb and 11 Mb on chromosome 2 of the combined Phase 2 HapMap Release 22 (NCBI 36) [101].

The cM per Mb values given in the HapMap release were converted to the recombination rate per base pair per generation by dividing by 100,000,000. The recombination probabilities and the bp change-points were passed to the simulation software. Other than the recombination rate, the simulations were run exactly as described before using the parameters in Table 7-3.

To generate an LD profile, 100 1 Mb chromosomes were simulated using the parameters for the neutral model with variable recombination rate. All pairs of SNPs up to 2 cM apart were collated and relevant statistics calculated for distances in discrete bins of size 0.0001 cM. The Z_α statistics were calculated as before.

7.5 Results

7.5.1 Uniform recombination rate

Simulation models were built to assess the effectiveness of the Z_α family of statistics. This resulted in three sets of 100 simulated populations: one set without any selection pressure, one with a selective sweep mid-way through, and one with a selective sweep nearing fixation. For each simulation a sample population of 1,000 chromosomes was selected for analysis. The Z_α statistics were calculated for each of the samples.

ROC curves were created to assess whether each statistic can could distinguish between neutral simulations, simulations mid-way through a sweep, and simulations nearing fixation. For each ROC curve, the AUC and pAUC was calculated. As each scenario contains the same number of simulations, it is valid to use these statistics for comparison. A table summarising the AUCs, with confidence intervals, can be found in Appendix A.3.1.

Figure 7-2 shows an aggregated graph of the Z_α statistic. This figure was created by binning the SNPs along the region into 10 Kb bins and averaging the results across the 100 simulated populations. This figure shows a clear differentiation between the neutral model and the models with selection. The figure also shows that the further along the selection is in time, the higher the Z_α and the wider the effect is. This is as expected as the haplotype containing the selected mutation spreads through more of the population, increasing the correlation between SNPs in the nearby regions. The AUC for the Z_α statistic was 1 for both neutral/mid-way and neutral/end comparisons but dropped to 0.7 for mid-way/end. To distinguish between sweeps mid-way and those reaching the end, the Z_β statistic can be employed.

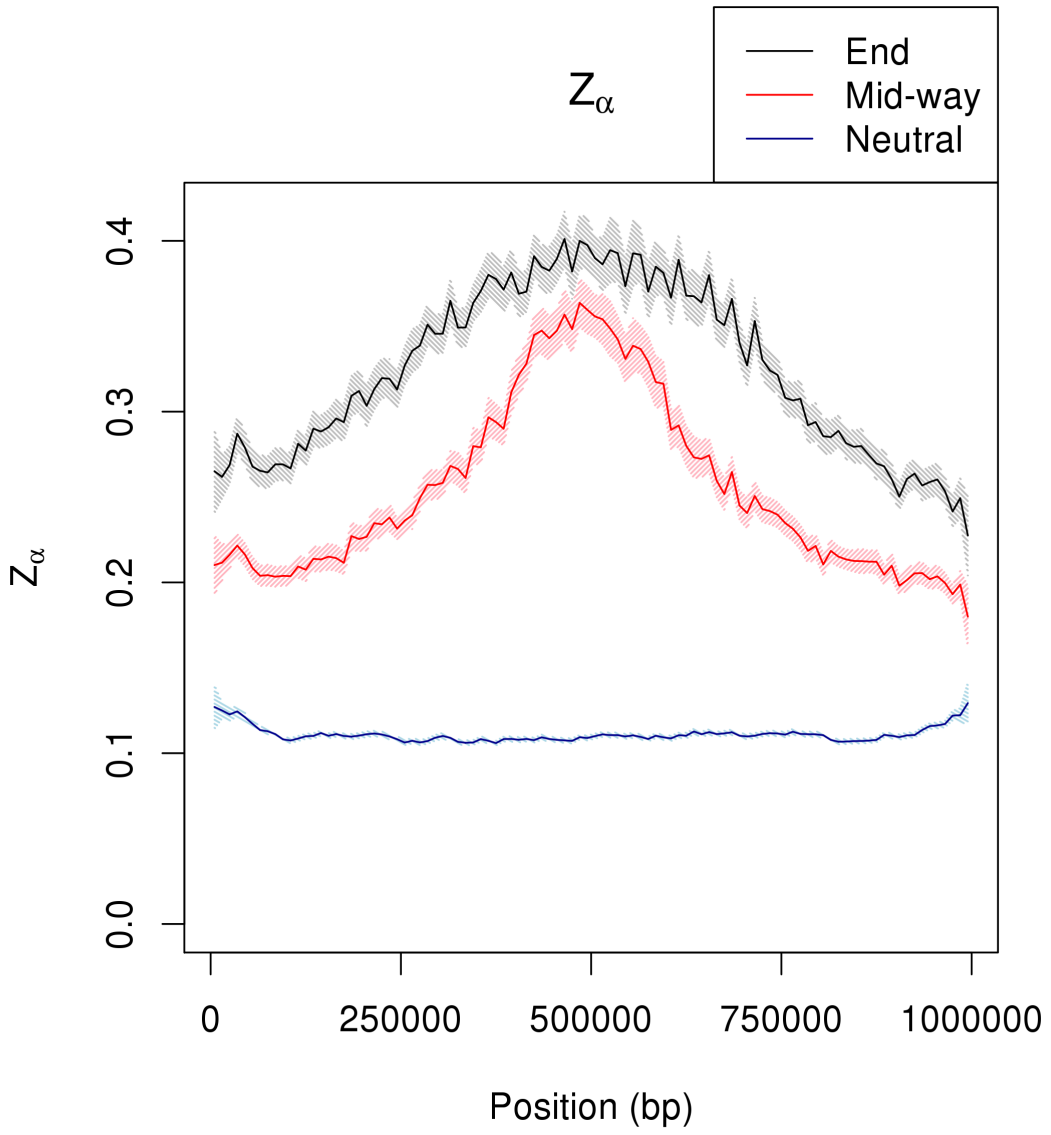


Figure 7-2 Z_α with simulated selection and a uniform recombination rate

The plot shows the aggregate Z_α values for each of the simulations in bins of size 10 Kb. For the two models including selection (mid-way in red and end in black) there was a selective event in the centre of the region.

Figure 7-3 shows a plot of Z_β, created in the same way as the Z_α plot. This plot shows different behaviours than the Z_α graph. For the mid-way plot, the Z_β looks similar to the Z_α plot - that is an increase around the site of selection. However, the end plot decreases around the selected site. As explained in section 7.2, as sweeps near fixation the correlation on either side of the sweep should remain high, but the correlations across the centre of the sweep will be reduced, due to recombination affecting each side independently. Thus, this graph shows the expected behaviour of the Z_β statistic. On its own, the Z_β statistic also had an AUC of 1 for both neutral/mid-way and neutral/end comparisons, and the mid-way/end comparison was 0.54. This is worse than the AUC for Z_α; however, by combining the two statistics an improvement can be made.

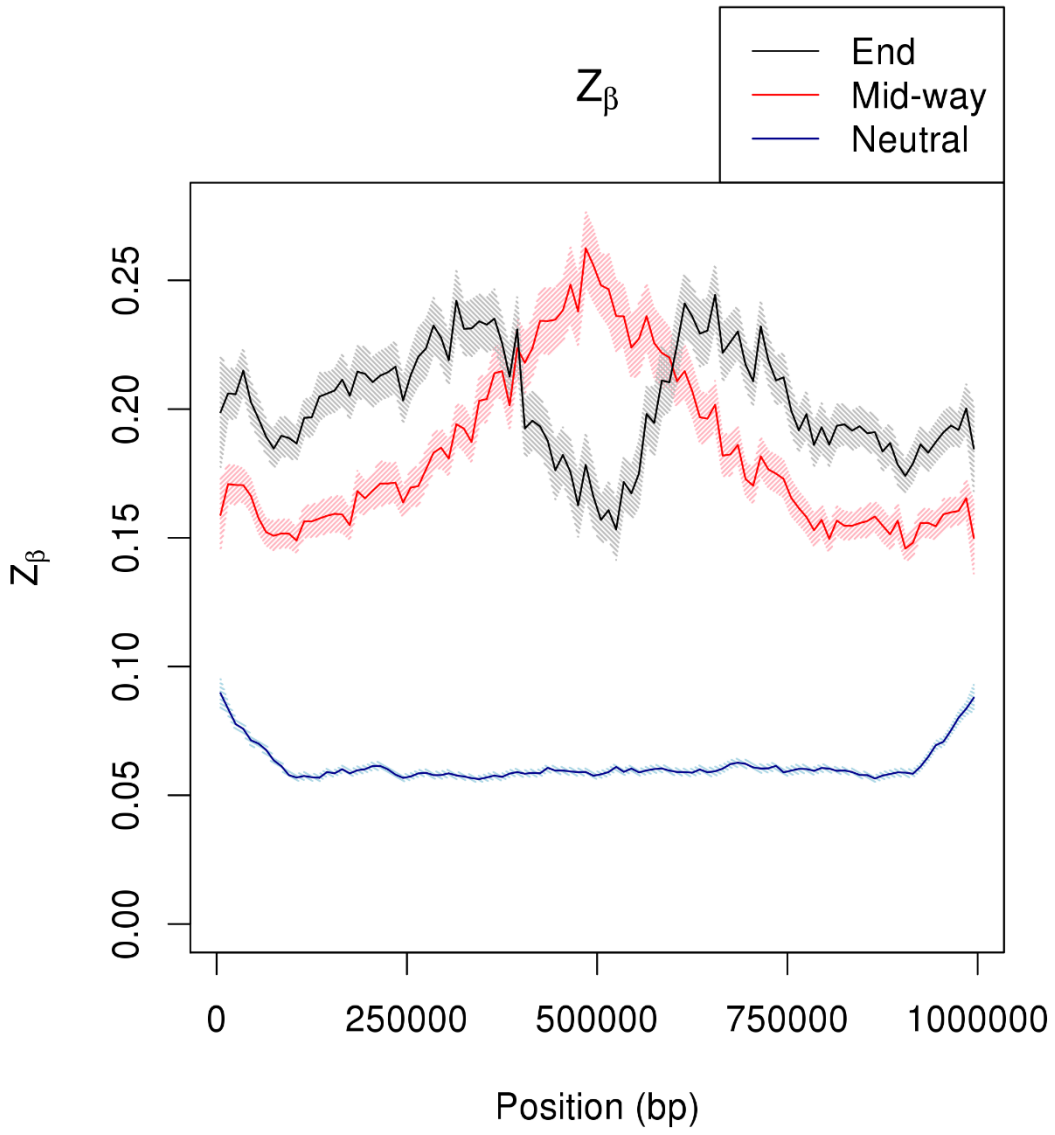


Figure 7-3 Z_β with simulated selection and a uniform recombination rate

The plot shows the aggregate Z_β values for each of the simulations in bins of size 10 Kb. For the two models including selection (mid-way in red and end in black) there was a selective event in the centre of the region.

Figure 7-4 shows the statistic Z_α/Z_β for the region. This combined statistic is a more effective way of distinguishing sweeps that are in progress to those that are nearing fixation, with an AUC of 0.95. While it is not as good at distinguishing sweeps from neutral (AUC = 0.77 for neutral/end and 0.86 for neutral/mid-way), this is unnecessary as the Z_α statistic provides this information. Therefore, this statistic should be used only where the Z_α statistic already indicates the presence of a sweep.

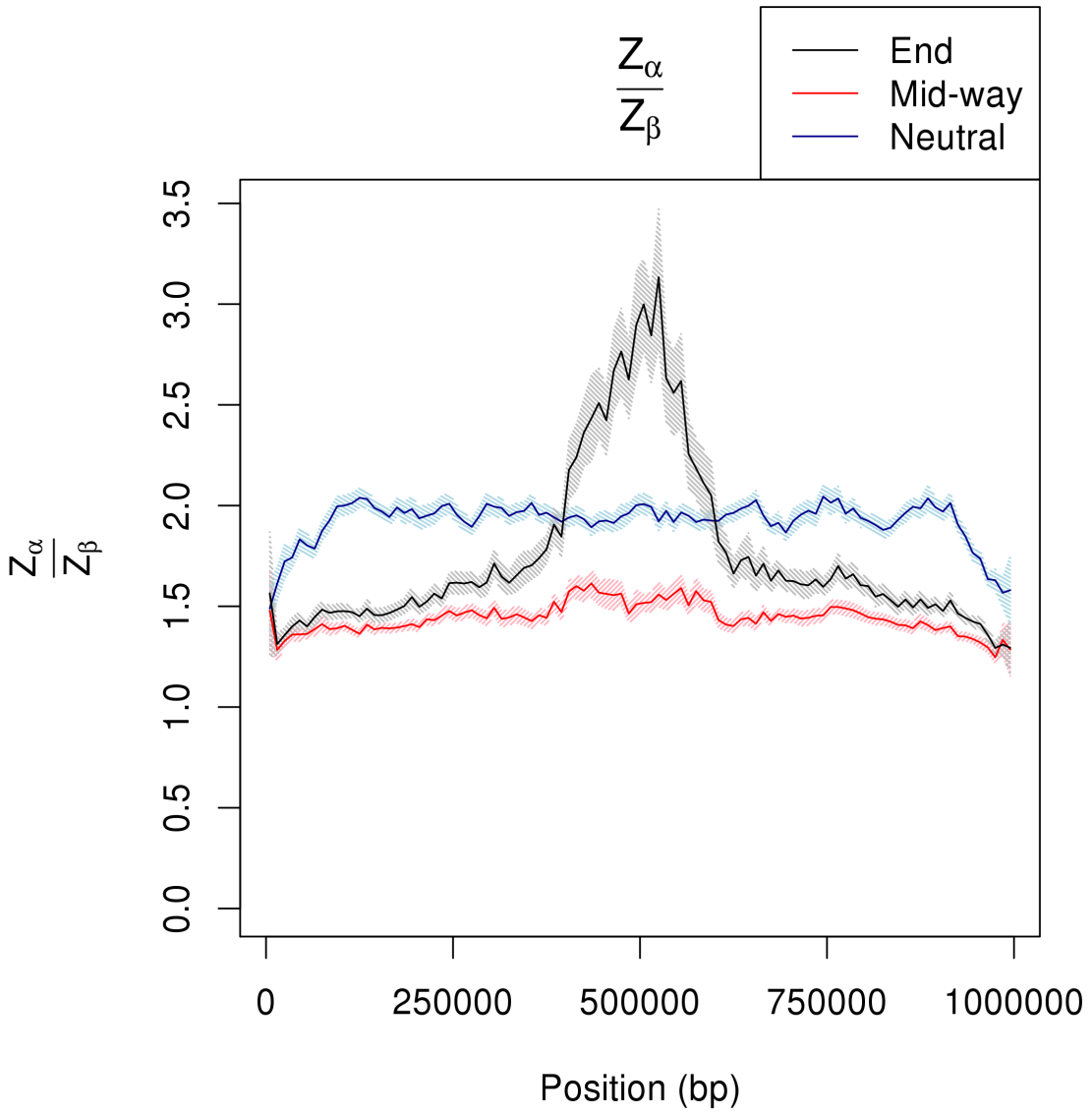


Figure 7-4 Z_α/Z_β with simulated selection and a uniform recombination rate

The plot shows the aggregate Z_α/Z_β values for each of the simulations in bins of size 10 Kb. For the two models including selection (mid-way in red and end in black) there was a selective event in the centre of the region.

While the statistics that adjust for variable recombination rates were included in this analysis, as the recombination rate was uniform for this scenario, they show much the same results as the base Z_α and Z_β statistics. The aggregate graphs for all the statistics can be found in Appendix A.3.2.

7.5.2 Variable recombination rate

To now assess the statistics that adjust for recombination rates, a variable recombination rate was included in the simulation model. A section of the HapMap recombination map was used to simulate a realistic recombination rate for a 1 Mb section of chromosome. The region chosen has been plotted in Figure 7-5 after conversion to recombination rate per base pair per generation.

This figure shows a few hotspots around the centre of the region, and a relatively cold spot around 700 Kb to 800 Kb. This region therefore is an ideal recombination map for the simulation study as it should interfere with the Z_α statistic.

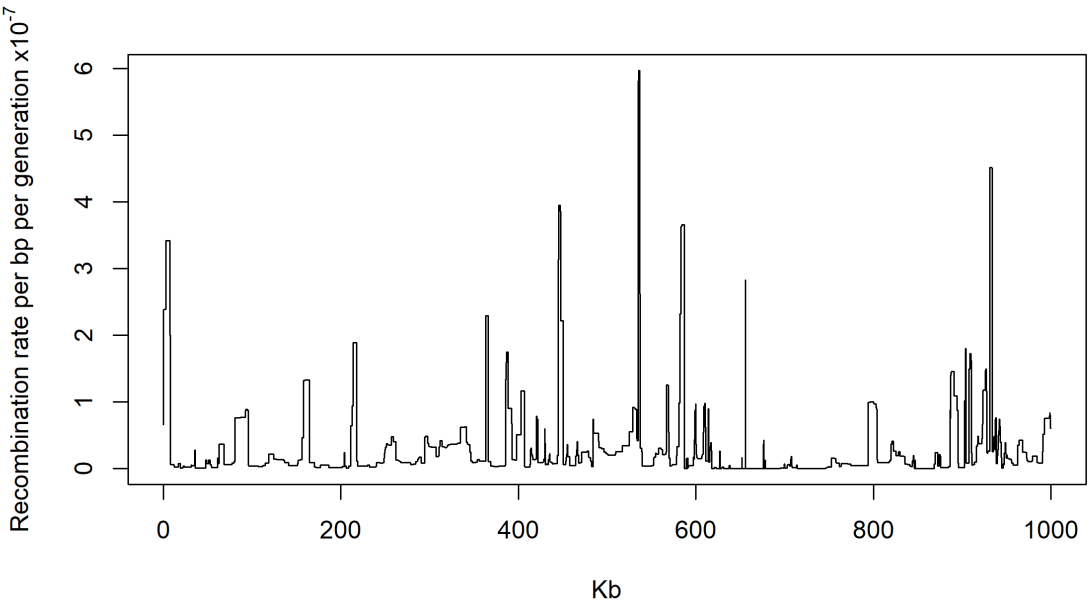


Figure 7-5 Plot of the variable recombination rate for simulations

This plot shows the recombination rate per base pair per generation for the variable recombination rate simulations. It is based on the 10-11 Mb region of chromosome 2 from the combined HapMap phase 2 release.

Figure 7-6 shows the Z_α statistic for the three scenarios with a variable recombination rate included. The maximum value for the statistics is now further along the chromosome than before, with the peak corresponding to the cold spot in the recombination map. The ROC analysis gave an AUC of 0.99 for the neutral/end comparison, 0.97 for neutral/mid-way, and 0.63 for mid-way/end.

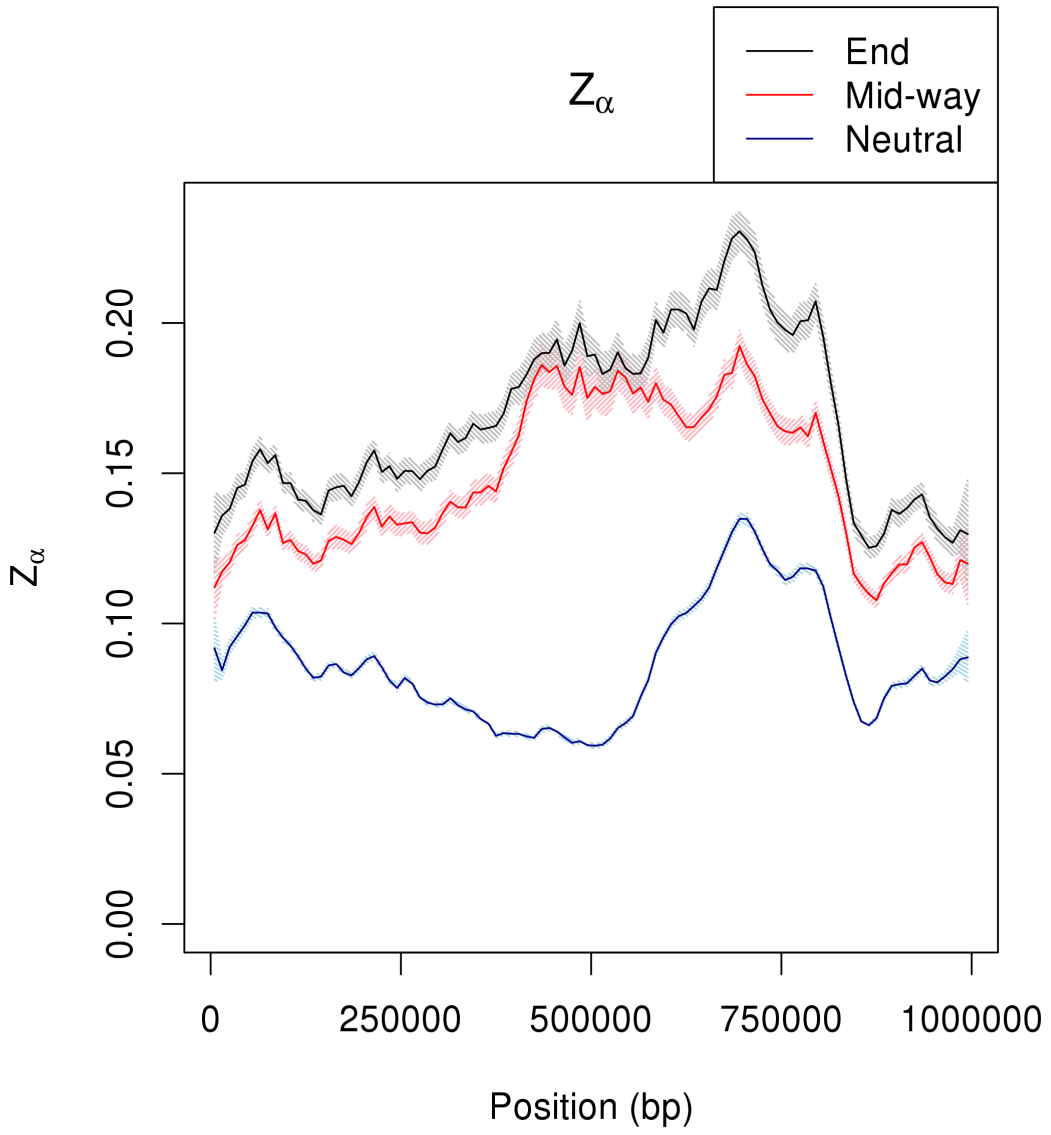


Figure 7-6 Z_α with simulated selection and variable recombination rate

The plot shows the aggregate Z_α values for each of the simulations in bins of size 10 Kb. For the two models including selection (mid-way in red and end in black) there was a selective event in the centre of the region.

Figure 7-7 shows the $Z_\alpha^{r^2/E[r^2]}$ statistic, a measure calculated by adjusting each r^2 value by the expected r^2 value given the genetic distances between SNPs. This graph clearly shows the peak around the centre of the region where the beneficial mutation is. This shows that adjusting for expected squared correlations can more accurately pinpoint the location of a selective event. The ROC analysis was almost perfect for neutral/end (AUC = 0.9998) and neutral/mid-way (AUC = 0.996), and like Z_α, is not good at distinguishing the mid-way/end comparison with an AUC of 0.61. As the AUC values for the neutral/selected and neutral/mid-way comparisons are so close to 1 for both Z_α and $Z_\alpha^{r^2/E[r^2]}$ it is hard to conclusively state that one is better than the other. Looking at

the partial area under the curve (pAUC) is another way of comparing the difference between Z_α and $Z_\alpha^{r^2/E[r^2]}$. The pAUC describes the region of the ROC curve where the false positive rate is 5% or lower. When performing selection scans, it is desirable to minimise false positives. For Z_α, the pAUC for neutral/end is 0.86, increasing to 0.996 for $Z_\alpha^{r^2/E[r^2]}$. Similarly, for the neutral/mid-way comparison, the adjusted statistic has an improved pAUC of 0.99 compared to 0.71 for Z_α. Thus, there is a clear improvement when using the adjusted version of the Z_α statistic.

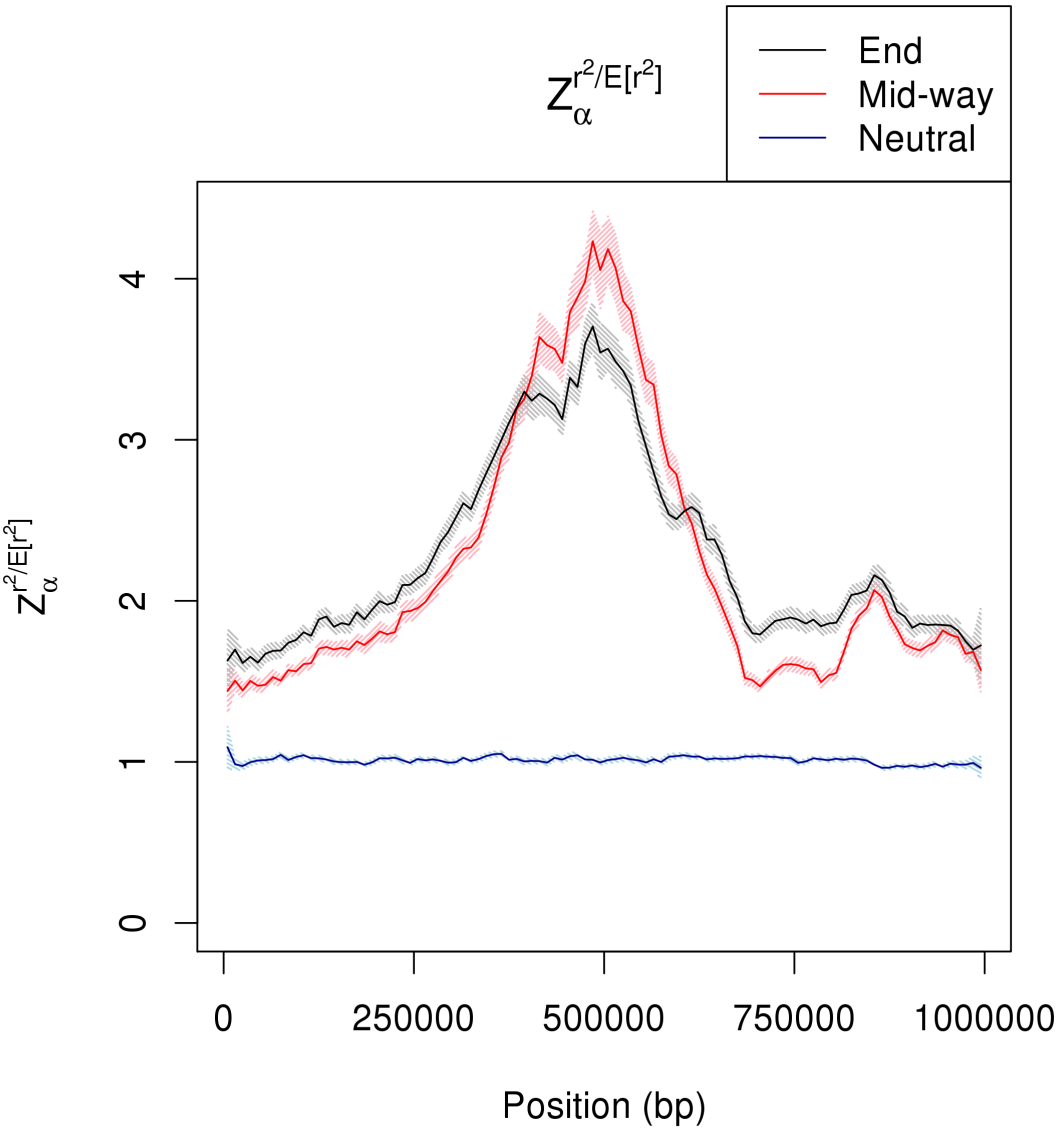


Figure 7-7 Z_α adjusted for expected r^2 with simulated selection and variable recombination rate

The plot shows the aggregate Z_α values for each of the simulations in bins of size 10 Kb. For the two models including selection (mid-way in red and end in black) there was a selective event in the centre of the region.

There are many ways to test for sweep progression as each Z_α variant has a Z_β counterpart, and there are numerous ways of combining them. An example is shown in Figure 7-8, which shows $\frac{Z_\alpha^{BetaCDF}}{Z_\beta^{BetaCDF}}$. The behaviour of the statistic when a sweep is mid-way and when it is nearing fixation are patently different. The AUC for this statistic when comparing these scenarios is 0.92, a marked improvement from solely using a Z_α variant.

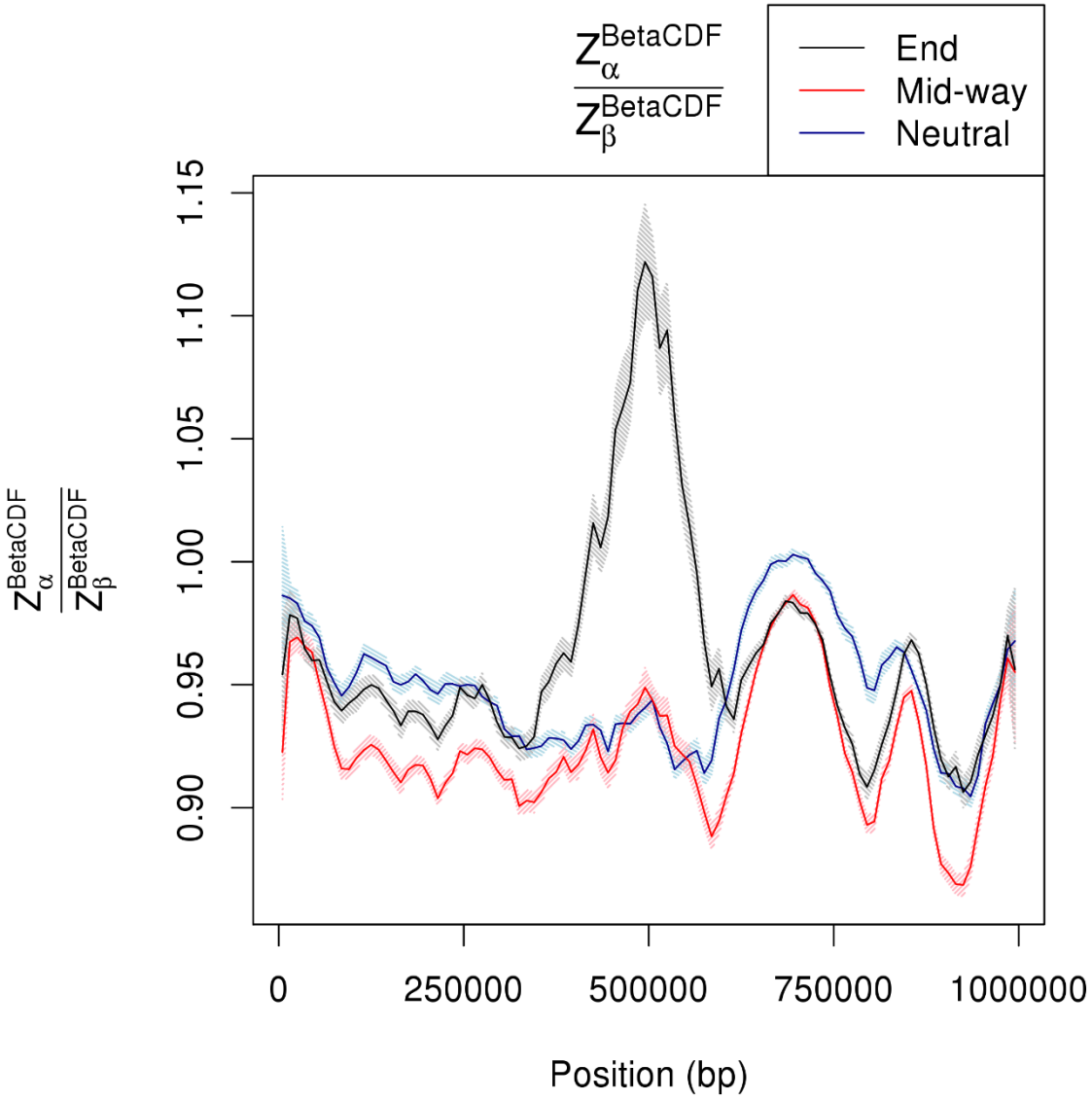


Figure 7-8 Z_α/Z_β adjusted for expected r^2 with simulated selection and variable recombination rate

The plot shows the aggregate $\frac{Z_\alpha^{BetaCDF}}{Z_\beta^{BetaCDF}}$ values for each of the simulations in bins of size 10 Kb. For the two models including selection (mid-way in red and end in black) there was a selective event in the centre of the region.

The graphs for all the statistics under the variable recombination rate scenario can be found in Appendix A.3.4, and a table summarising the AUC values in Appendix A.3.3.

7.6 Discussion

The zalpha R package is a new piece of software designed to allow researchers to easily examine their genomic data for evidence of selective sweeps. The innovative aspect of the Z_α statistics is the ability to adjust for expected squared correlations between SNPs given the genetic distances between them. Researchers can now apply these statistics in a reliable and reproducible way, using this free, open source package.

The results presented in this chapter showed that the statistics perform as expected given the motivation and science behind them. The addition of a variable recombination rate showed a marked improvement in the ability to distinguish between neutral simulations and those containing selection once correlations were adjusted for. The further step of assessing the stage of the sweep was also shown to be effective by using Z_α in combination with Z_β.

The simulation scenarios presented here were simple and basic to clearly show the effects of changing from a neutral scenario to one with selection, and to show the difference when a variable recombination rate was introduced. As a result, some of the statistics returned an AUC of exactly 1 under some of the conditions. This is unrealistic and unlikely to be the case once statistics are applied to real-world data. Genetic drift, demography, and mutation rates among other factors are all likely to have a confounding effect on the proficiency of the statistics.

One of the biggest challenges when writing this program was the issue of floating-point errors [378, 518]. Genomics research necessarily involves very small numbers and so it is imperative that any software can handle small numbers accurately. This problem was solved by creating a helper function to process comparisons between decimals. This was not a trivial problem to solve, as early attempts made the run-time for the code much longer than necessary with large datasets, but eventually a solution was found that balanced the need to compare non-integer values and quick running speeds.

There are a handful of R packages already in existence for performing selection scans on single populations, none of which use the Z_α statistics described in this chapter. The tsel R package [390] described and applied in Chapter 3, uses a method based on the estimated time to most recent common ancestor to find selection. Another package, PopGenome [519], has the benefit of being flexible as it can accept many different standard genomic data formats as inputs. As well as selective scans, for which it uses the CL and CLR methods by Nielsen *et al.* [280], it can also

perform other relevant analyses such as neutrality tests. For LD estimates, it uses Kelly's Z_{nS} [258] and Rozas' ZA/ZZ [520], which are similar statistics to the basic Z_{α} statistic. However, this package does not adjust for expected squared correlations using an LD profile or any other method. The R package rehh [521] implements the extended haplotype homozygosity (EHH) method and other related methods for single population scans. While these are popular methods, and are easily interpretable, they are not as effective for soft sweeps, potentially the primary type of sweep in humans [209, 370]. To take into account variations in the local recombination rate, the package includes the cross-population EHH (XP-EHH) method [271]; however, this necessitates a comparison with a second population. Each of these packages have distinct virtues and each assess evidence for selection in a different way. Results could be validated and improved upon by applying different methods through using a combination of packages. The ease of this is enhanced by them all being coded in the same language and environment, making data manipulation and comparison straightforward.

While Jacobs *et al.* [257] tested their statistics on simulations mimicking real-life data, they only applied them to two human chromosomes as a proof of concept. Now that this package has been created and tested on simulated scenarios to confirm that they work, the next step is to apply them to full, real-world datasets to establish if they can identify candidate regions of the genome in real populations.

Chapter 8 Application to the domestic dog genome

8.1 Introduction

The domestic dog (*Canis lupus familiaris*) is known as man's best friend, and as such people are curious about the history of their companions. Just as there are private companies offering genetic ancestry tests for humans, there are multiple companies offering similar services for their pets [522-525]. As well as revealing the breed mixes of their dogs, these companies can also test for hereditary diseases or risks. This allows owners to be alert for potential symptoms and care for their pets. Quite apart from helping the animals themselves, dogs have also been shown to be a good model for some rare human diseases as they are large mammals who, crucially, live in the same environment as their human counterparts with the same conditions [526]. Dogs are also a useful model for the study of evolution as they fall within a different clade, Laurasiatheria, than other well-studied mammals such as humans, chimpanzees and mice who are Euarchontoglires, and so can function as an outgroup [527].

Dogs have an unusual evolutionary path, in that they have gone through a double selection process. The first was the initial split from the grey wolf [528]. South East Asia around 33 kya has been posited as an origin for the domestication of dogs [529], and there is evidence that Chinese and other Asian dogs are the most closely related to the ancient ancestor [530, 531]; however, there is still debate as to the exact time and place of dog domestication [532-534]. The second selection process was during the Victorian times when the majority of breeds were formed, meaning the huge phenotypic differences between modern dog breeds today were formed in less than 200 years, roughly 100 generations [535]. Genetically, this has been problematic for dogs as the population bottlenecks during breed formation has caused negative mutations to accumulate in the regions around selected traits [536].

The history of dogs in the Americas is an interesting one and is briefly mentioned here as the data to be examined in this chapter consists of North American dog genomes. Dogs first arrived on the continent over 9 kya, almost certainly pre-domesticated as opposed to being domesticated from North American wolves [537, 538]. Then, around 1,000 years ago the Thule people arose from Siberia into Alaska and proceeded to expand across the North American continent, bringing sledding dogs with them [539, 540]. The arrival of Europeans to the Americas in the 15th century saw the commencement of an almost complete replacement of native American dogs with European dogs, by way of deliberate persecution, preference for European dogs, and the susceptibility of the native dogs to introduced diseases [541, 542]. The final introduction of dogs

to the Americas was during the Alaskan gold rush in the late 19th century [543, 544]. Genetic material from native American dogs is very rare to find in modern day dogs in the Americas [541, 544].

After the initial domestication of dogs from grey wolves, there is currently no evidence from within the dog genome that introgression from wolves continued. Conversely, the wolf genome contains evidence of gene flow from dogs back to wolves in most wolf populations [545]. This has resulted in wolves gaining a variant for black coats: a trait that has been selected for in wolves living in forest environments [546]. There is also evidence of hybridisation with jackals from stray dogs mixing with wild jackal populations, although it is yet to be seen if the benefits of hybridisation in this case, such as adaptive introgression of variants helpful for living in close proximity with humans, are greater than the negative effects, for example introduction of disease [547].

The *Canis* genus are potentially unique among mammals as they possess a non-functioning version of the *PRDM9* gene [423]. It is possible that the entire *Canidae* family have this non-functioning allele, potentially dating back 49 million years to the most recent common ancestor of dogs and pandas [548, 549]. In Chapter 4 it was discussed that *PRDM9* is implicated in recombination and the formation of hotspots in mammals, including humans [416]. Nevertheless, even without a functioning *PRDM9* gene, the dog genome still contains recombination hotspots [550]. These hotspots are observed to correlate with CpG-rich areas of the genome and transcription start sites (TSS). This relationship has also been observed in chickens [551], who also do not have a functioning *PRDM9* gene, and to a lesser extent in humans [101] and chimpanzees [552].

There is no evidence that recombination rates or hotspots have changed between domestic dogs and wolves [553]. It has been hypothesised that recombination rates should increase as a result of domestication. This is due to the loss of genetic variability caused by bottlenecks [554] and also to avoid Hill-Robertson interference between selected loci [404]. Conversely, recombination rates could be expected to decrease to avoid introducing negative alleles from wild ancestors into beneficial haplotypes [555]. As there is no evidence for either hypothesis, it means a recombination map made for one population of dogs should hold for another. Note that while this holds for now, this may not stay true indefinitely, as the recent severe bottlenecks in dog evolution may have suppressed any potential changes to recombination rates.

Dog autosomes are acrocentric, meaning the centromere is located very close to the start of the chromosome [556], unlike most human autosomes (the exceptions being chromosomes 13, 14, 15, 21 and 22 [557]). Similarly to humans, it has been found that recombination rates in dogs

increase towards the telomeric regions and are lower near the centromere. This is especially evident in male recombination, in both dogs and humans [558]. This behaviour, combined with the acrocentricity, means that recombination rates should increase along the chromosome.

As dogs have always lived and evolved alongside humans, they have been exposed to the same environmental pressures and thus there are similarities in their evolution over this time. There are many similar selected traits in the same orthologous genes, for example in *ABCG5* and *ABCG8* which are involved in digestion [559]. Tibetan Mastiffs have experienced a sweep in *EPAS1* as an adaptation to a high-altitude environment; the same gene has also been identified in studies of humans inhabiting the Tibetan plateau. The adaptations in dogs and in humans were both due to introgression events from Tibetan wolves and Denisovans respectively [218, 560, 561]. As a result of living with humans, dogs have also evolved an enhanced ability to digest starch [562] in parallel to their human counterparts as humans progressed from hunter-gathering to agriculture; however, different genes were selected for in each species [563].

For finding evidence of selection, it is common in published studies to use F_{ST}, or some other cross-population statistic, to compare dogs to grey wolves [559, 564-566] or breeds of dogs to other breeds of dogs [567, 568]. It is fortunate when studying dog evolution that their wild counterpart is still extant; however, many other domestic creatures do not have such an equivalent outgroup, such as cattle whose wild ancestor the aurochs (*Bos primigenius*) became extinct in the 17th century [569]. Thus, for this study a single population statistic will be used to find potential regions under selection, to be validated by previous studies using an outgroup, to show its potential for use on single population datasets.

The aim of this chapter is to analyse the genomes of a population of dogs by building recombination maps and applying the Z_α statistics to find candidate regions for selective sweeps.

8.2 Methods

The code for this chapter can be found at https://github.com/chorscroft/PhD-Thesis/tree/main/Chapter_8.

8.2.1 Data cleaning

The datasets were sourced from Embark, a US company offering DNA test kits for dogs, and have previously been published in Deane-Coe *et al.* [570] and Sams and Boyko [571]. There were three datasets: *discovery*, *at_risk* and *breed_dog*. The datasets are described in Table 8-1. These were merged together and cleaned using PLINK v1.90beta [432] to remove rare variants that could skew the statistics. Only SNPs from the smaller datasets *at_risk* and *breed_dog* were kept, as well

as SNPs with MAF greater than 0.05 and missingness less than 5%. Individual dogs in the dataset with missingness of 5% or more were also filtered out, resulting in a dataset containing 5,641 dogs with 155,402 variants. Hardy-Weinberg tests could not be applied to the datasets as the assumption of random mating is not valid for dogs as a species [572]. In fact, only 0.05% of the SNPs passed the Hardy-Weinberg test at p-value < 0.001.

Table 8-1 Information about the raw data for the dog analysis

Dataset	Dogs	Variants	Reference	Description
Discovery	3,180	213,245	[570]	Mostly mixed-breed dogs, from Embark's customer database whose owners responded to a survey between April and November 2017
At_risk	670	175,123	[571]	Dogs in Embark's customer database in April 2018 homozygous for recessive deleterious conditions
Breed_dog	1,792	175,123	[571]	Purebred dogs from the 11 most common breeds in Embark's database in January 2018

8.2.2 Relatedness

To test for relatedness, first the data were pruned using PLINK, allowing for r^2 values of no more than 0.5, checking window sizes of 50 SNPs in steps of 5, resulting in 113,604 SNPs being kept. PLINK was used to calculate the relatedness between each pair of dogs, using the genome option. There were $\binom{5641}{2} = 15{,}907{,}620$ pairs of dogs. PI_HAT is a value calculated by PLINK that approximates the relationship coefficient. It measures the proportion of the genomes of the pair of dogs that is identical by descent (IBD). Table 8-2 shows the breakdown of PI_HAT values. Over 97% of the pairings had a PI_HAT value less than 5%. Of more concern is the 389 pairs which had a PI_HAT value of 1, indicating clones, monozygotic twins, or in this case, duplicated samples.

Table 8-2 Count of pairwise relationships for each PI_HAT window

This table shows the counts of PI_HAT values falling within these thresholds for each pair of dogs in the sample.

PI_HAT interval	count	PI_HAT interval	count
0 - 0.05	15,562,037	**0.5 - 0.55**	781
0.05 - 0.1	87,927	**0.55 - 0.6**	125
0.1 - 0.15	65,134	**0.6 - 0.65**	31
0.15 - 0.2	45,144	**0.65 - 0.7**	19
0.2 - 0.25	48,723	**0.7 - 0.75**	24
0.25 - 0.3	47,622	**0.75 - 0.8**	7
0.3 - 0.35	27,595	**0.8 - 0.85**	1
0.35 - 0.4	10,854	**0.85 - 0.9**	0
0.4 - 0.45	7,760	**0.9 - 0.95**	0
0.45 - 0.5	3,447	**0.95 - 1**	389
		Total pairs	**15,907,620**

The 147,378 pairwise relationships between dogs with a PI_HAT value greater than 0.2 were selected for further analysis, to ascertain which dogs should be removed to clean the dataset. This value was chosen because it is conservative and should remove all first and second degree relationships from the data [573]. This is stricter than some other studies on dogs [571, 574]. As dogs can be inbred, especially those from pure-breed lines, it is possible that this threshold could remove more dogs than strictly necessary due to them being the same breed. However, as this analysis is focused at the species level not the breed level, this risk was deemed acceptable.

Figure 8-1 shows all 147,378 of the relationships between dogs with a PI_HAT value greater than 0.2. This graph shows that, while the majority of clusters are pairs (71% of clusters) there are multiple large clusters. The graphs of the clusters were generated using R v3.6.0 [480] and the package qgraph v1.6.5 [575].

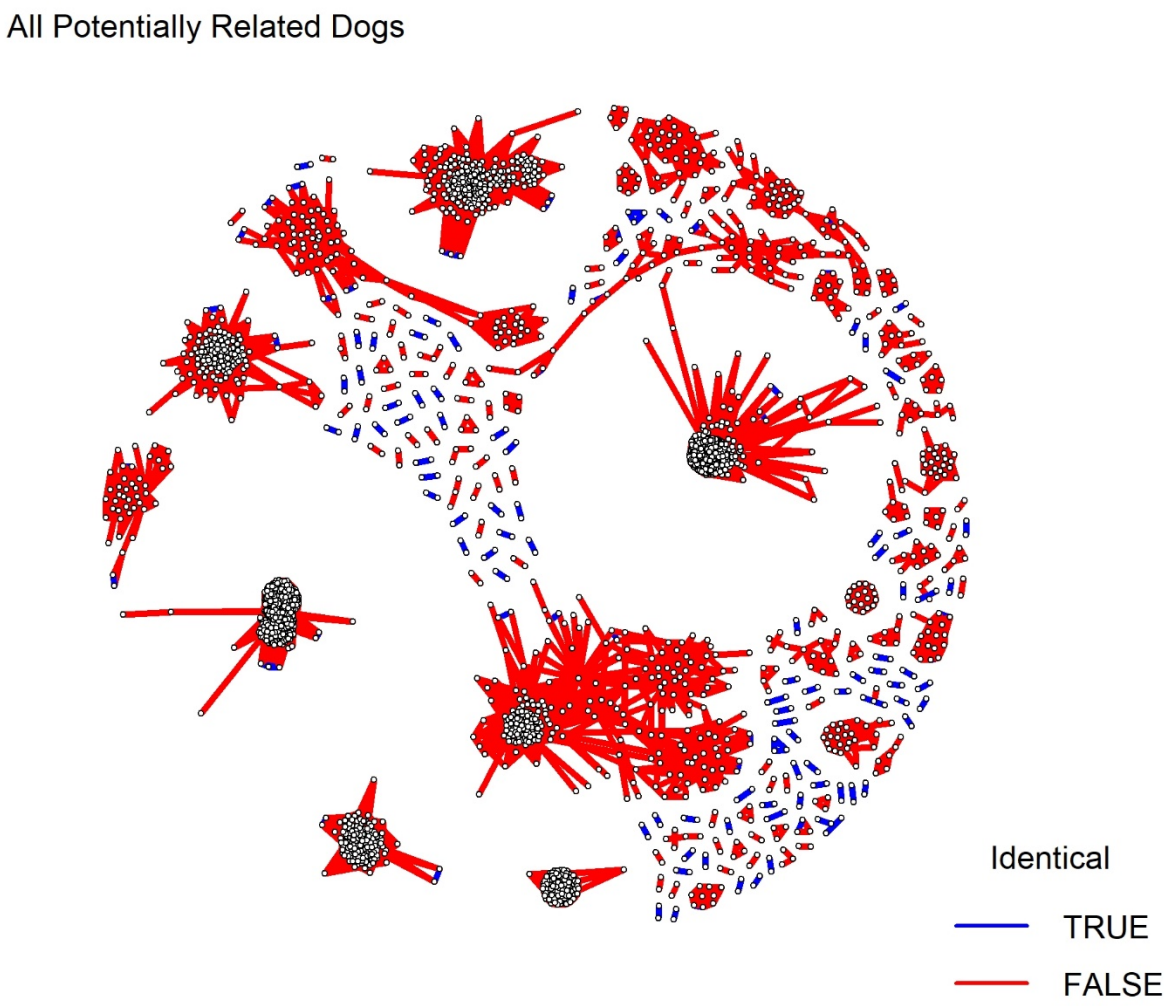


Figure 8-1 Clusters of related dogs

This figure shows the relationships between dogs. Each circle represents a dog with at least one potentially related partner. Each line represents a relationship, where PI_HAT > 0.2. If PI_HAT = 1, the line is blue. The graph shows that the majority of clusters are pairs; however, there are a few very large clusters.

Figure 8-2 shows the relationships between dogs where the relationships are within the same dataset, and between different datasets. This shows that none of the duplicate samples were in the same dataset, and instead were identical dogs called from the Embark database when the dogs were extracted each time. Comparing graphs A, B, and C, there are many more problematic relationships in the *breed_dog* dataset, as expected given purebred dogs are more likely to have shared recent ancestry.

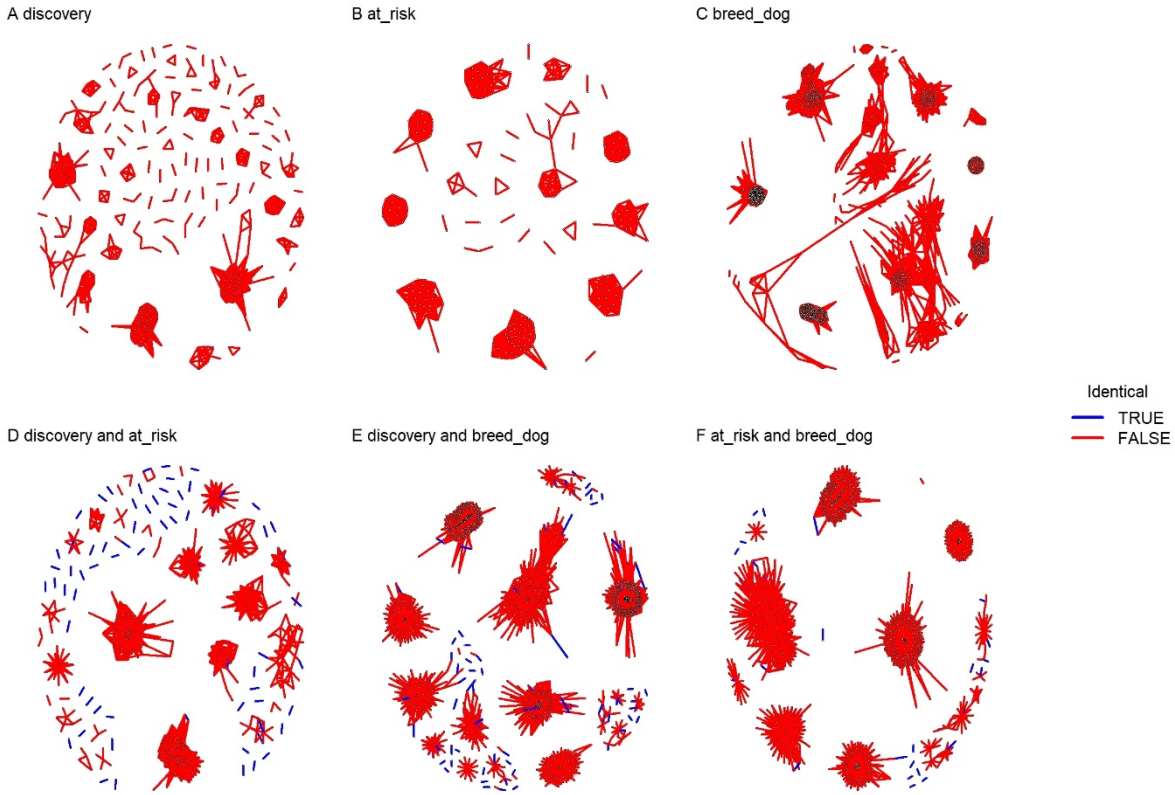


Figure 8-2 Relationships between dogs within and between datasets

This figure shows the relationship between dogs within datasets (A-C) and between datasets (D-F). Each point is a dog, and lines represent relationships between them. This is a line if PI_HAT is greater than 0.2, and it is coloured blue if PI_HAT = 1, i.e. an identical pair. There are no duplicate samples within datasets.

Figure 8-3 shows a flow chart of how the dogs were chosen to be removed from the dataset. A greedy algorithm was implemented, based on the number of problematic links to an individual dog, and the amount of missingness in the data. Firstly, dogs that had an identical pair (or were in a trio) were filtered out, choosing the dog with the least missingness to keep. After this, the dogs were then removed one by one, removing dogs with the most links, breaking ties by keeping the dog with the least missingness. This algorithm was implemented in R v3.6.0 [480].

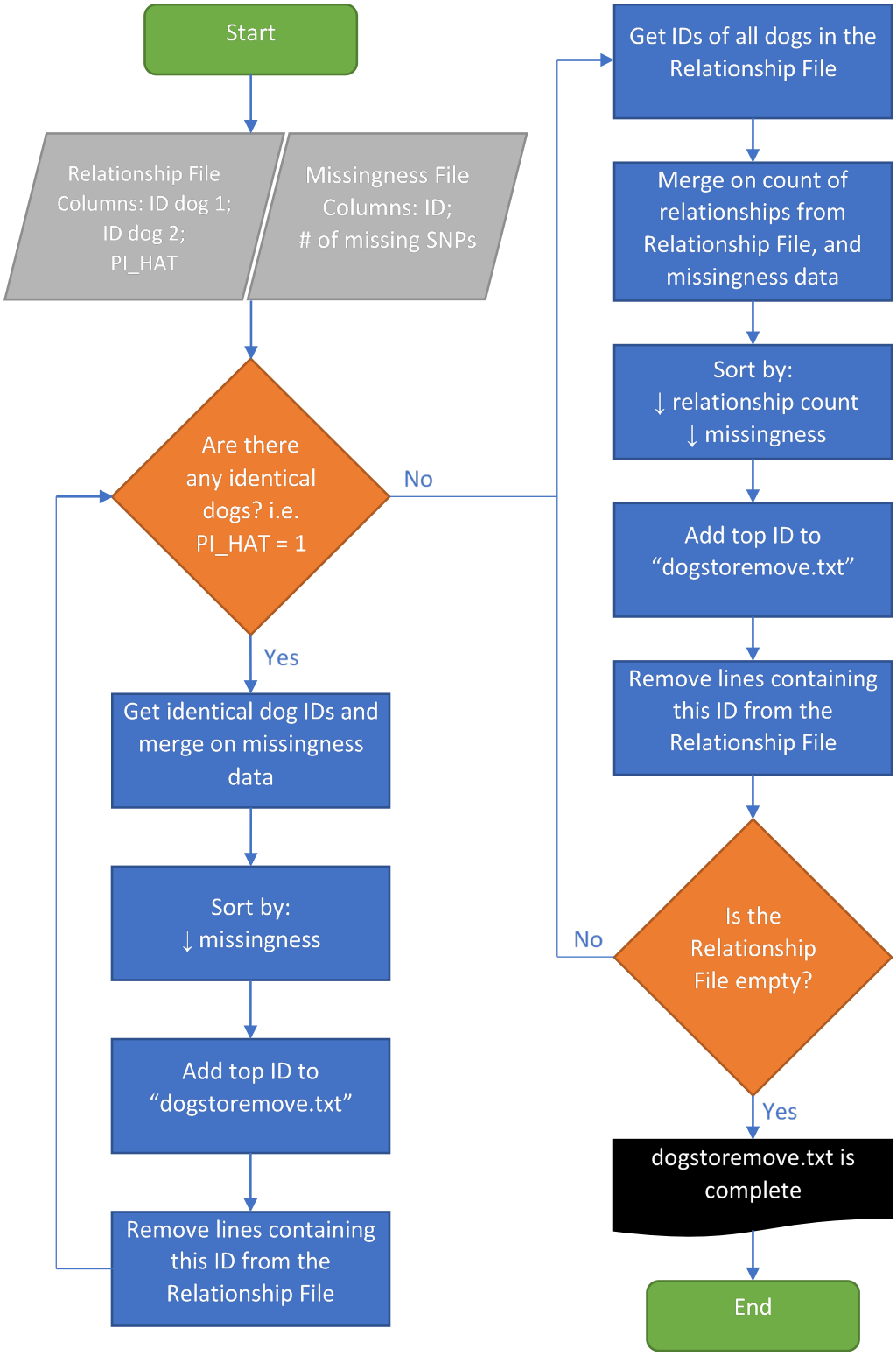


Figure 8-3 Flowchart of removal of related dogs

This figure shows the process of removing dogs one by one until no pairs of dogs remain with a PI_HAT of > 0.2. A greedy algorithm was implemented, removing dogs with the most problematic relationships first, breaking ties by the amount of missingness. Dogs in identical pairs or trios were removed first.

1,940 dogs were removed from the dataset, the majority (65%) were from the *breed_dog* dataset. This is as expected as these dogs are more likely to have a common ancestor due to inbreeding and small population sizes. 70% of the dogs from the *breed_dog* dataset were removed in this step.

The dataset after filtering for relatedness contained 3,701 dogs with 155,402 variants.

8.2.3 PCA

The final step of the data cleaning was to perform a principal components analysis to make sure the dogs are all from the same population and are not skewed by outliers. This was carried out as in section 4.2.1 using the PCA command in PLINK to obtain the eigenvalues and eigenvectors. The SNPs from the pruned dataset were used, excluding those on the sex chromosomes. Figure 8-4A shows the graph of principal components against the variances for the top 20 PCs. The total variance for the dataset was 4052.46 and the variances shown here are the variance explained by each PC. Using the elbow method, seven PCs were selected going forward, explaining around 7% of the variance in the data. The within groups sum of squares were calculated, as shown in Figure 8-4B, and nine clusters were chosen as an appropriate number using the elbow method.

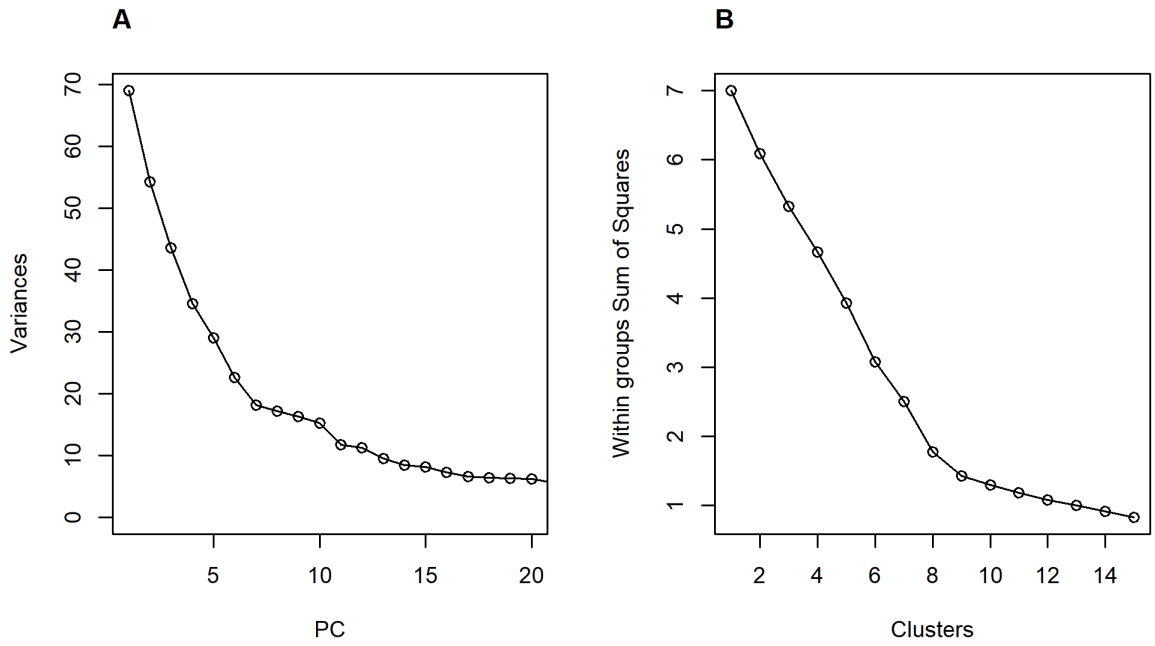


Figure 8-4 PC and cluster ascertainment graphs

Graph A shows the plot of variances for the first 20 (of 3701) PCs calculated using PLINK. The elbow method was used to choose seven PCs as an acceptable number. Graph B shows the within groups sum of squares calculated for each cluster size. Again, using the elbow method, nine clusters were chosen as the appropriate number of clusters.

The graph of the first two principal components is in Figure 8-5. The largest cluster is shown in red and contains 1,452 of the dogs. These will be the dogs kept for the final analysis. A pairwise comparison of all seven of the PCs is available in Appendix A.4.1.

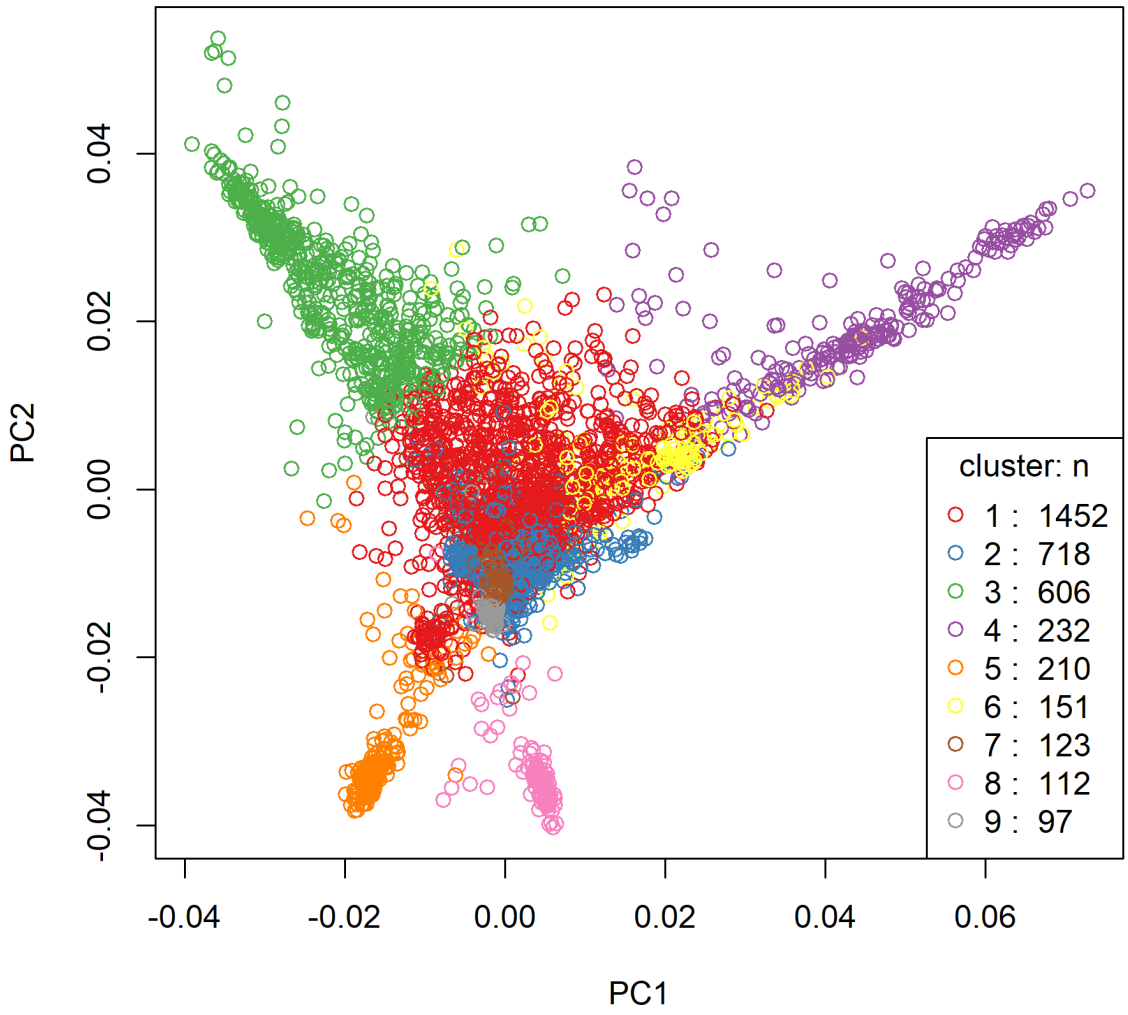


Figure 8-5 PCA plot of dog data

This figure shows a plot of the first two PCs from the analysis of the dog data. The data has been split into 9 clusters that have been colour coded. The largest cluster, in red, contains 1,452 dogs. These are the dogs which will be used in further analysis. A pairwise comparison of all seven of the PCs is available in Appendix A.4.1.

Interestingly, looking at composition of the clusters by the original datasets shows an uneven distribution. Figure 8-6 shows the number of dogs in each cluster, split by which original dataset the individual was from. Clusters 5, 8 and 9 mostly consist of dogs from the *breed_dog* dataset. It is not a surprise that the dogs from the same breed have clustered together. This increases confidence that the dogs forming the final dataset from the largest cluster are indeed mixed breeds from the same population.

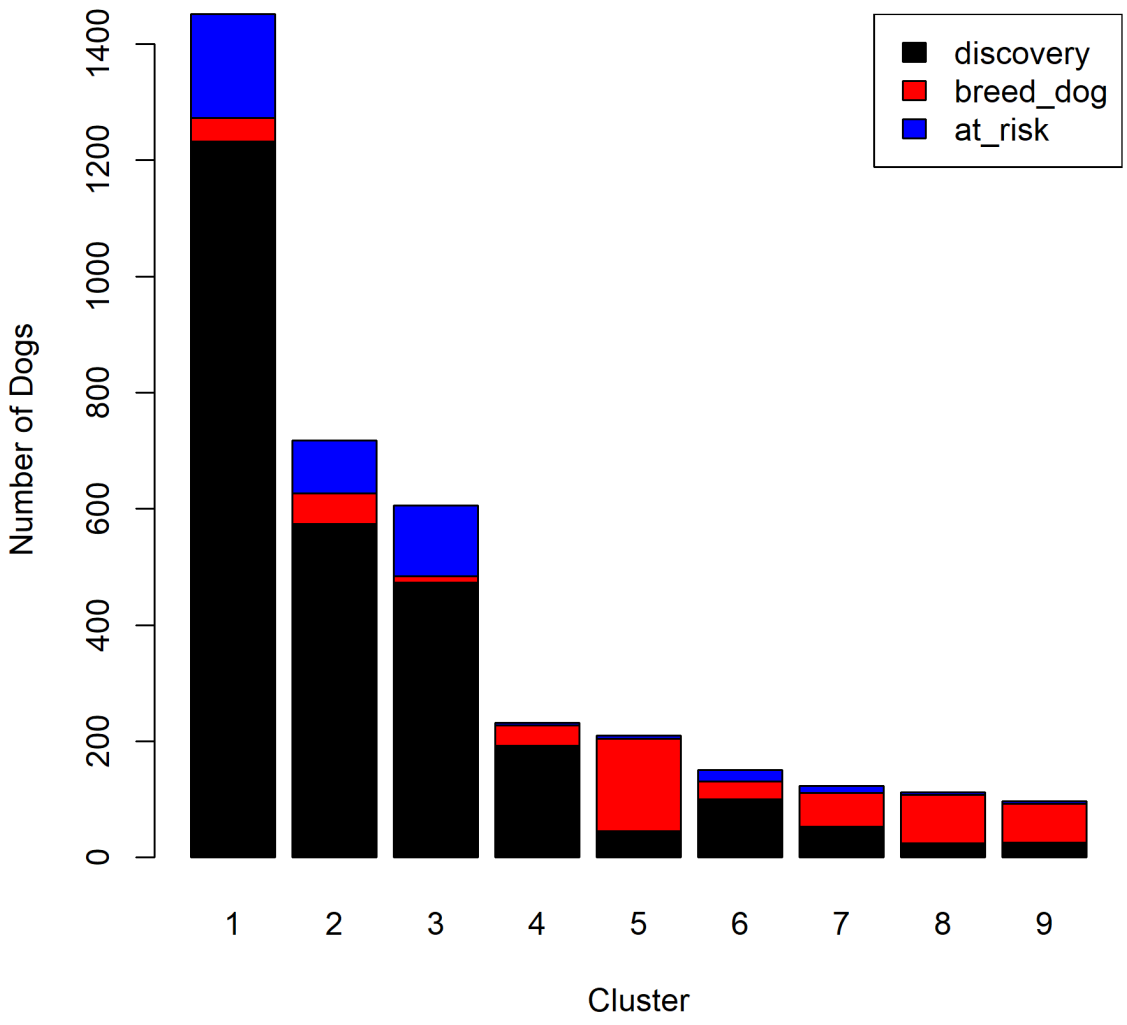


Figure 8-6 Dog clusters by original dataset

This graph shows the number of dogs in each cluster, split by the original dataset each dog came from. There is an over-representation of dogs from the breed_dog dataset in clusters 5, 8 and 9 where they make up the majority of the dogs.

Finally, after filtering for the largest cluster from the principal component analysis, the SNPs were once again filtered for MAF greater than 0.05 and missingness less than 5%. The final dataset contained 1,452 dogs and 153,514 variants.

8.2.4 LDhat

To use zalpha to its fullest extent it is necessary to generate an LD map for the data to find the genetic distance between SNPs and create an LD profile. To do this, the data were formatted and run through the LDhat software. This process is described in full in section 4.2.3. Maps were

created for each of the autosomes. The sex chromosomes were dropped at this point as the later analysis will consider genome-wide patterns and the X and Y chromosomes are not representative due to smaller effective population sizes for X [576] and the use of popular sires in dog breeding affecting the use of Y [577]. As LDhat output is in terms of ρ instead of centimorgans (cM), it is necessary to convert using either an estimate for N_e or an established map length. N_e is difficult to ascertain for dogs due to the intense bottleneck that occurred during domestication [553, 578], so the conversion was carried out using a previously published map length. The maps generated by Auton *et al.* [550] using the CanFam3.1 build were downloaded (19th May 2020) and used to estimate the map lengths for each chromosome. As some of the SNPs in the dog dataset were outside the range of the Auton maps, it was necessary to trim them out. This only affected 24 SNPs in total. Once each map was built it was converted to cM for each chromosome using the cM lengths in the Auton maps, using the exact length if the start and end SNPs were present in both maps or by taking a linear estimate if not. Figure 8-7 shows a comparison of the LDhat map lengths for each chromosome, taken from the LDhat output file, and the Auton maps. The squared correlation is strong (r^2 = 69%) and significantly different from zero (p-value = 1.07×10^{-10}), meaning it is acceptable to use the Auton map lengths to convert to LDhat output into cMs. More comparisons with previously published LD maps and linkage maps for dogs are made in section 8.3.2 of this chapter. The linkage maps created by Campbell *et al.* [558] were downloaded on 4th June 2020. Recombination maps are plotted using the qqman v0.1.4 R package [579] and the intervals v0.15.2 R package [493] for the interpolated map.

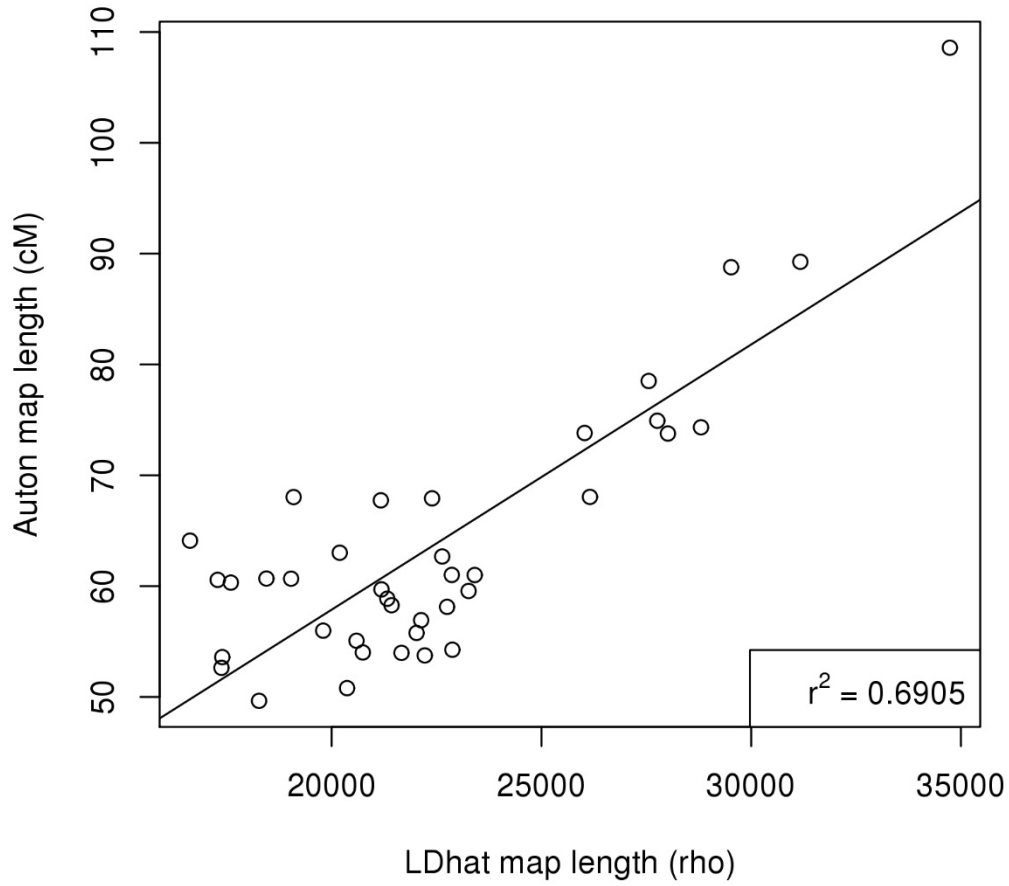


Figure 8-7 Comparison of LDhat and CanFam3.1 map lengths

Each point on this figure represents a chromosome. The map length taken from LDhat is on the x-axis, and the map length in centimorgans ascertained from the Auton recombination map is on the y-axis. The correlation between the maps is high meaning it is acceptable to use the Auton (CanFam3.1 build) map lengths to convert the LDhat output to cMs.

Once each map was built and converted to cM for each chromosome, the maps were then edited into the format required for zalpha by reverting the positions back into physical base pairs rather than kilobases relative to the start.

8.2.5 Wavelets

Comparison of the genetic maps were made using wavelet analysis. This followed the same maximal overlap discrete wavelet transform (MODWT) analysis as described in section 5.2.4. All analysis was carried out in R using the wmtsa v2.0-3 R package [475]. The data were interpolated into bins of size 500,000 bp before the wavelet analysis was applied using the intervals v0.15.2 R

package [493]. The recombination rates were log-transformed using base 10 as per the human analysis due to the extreme nature of recombination hotspots. All wavelet analyses used the Haar wavelet as the base.

8.2.6 zalpha

The zalpha R package described in detail in Chapter 7 was used to apply the Z_α family of statistics to the dog dataset. The cleaned and filtered datasets were used as inputs, as well as the LDhat maps for the genetic distances, and an LD profile. The Zalpha_all function was used to generate all the statistics simultaneously. A window size of 300,000 bp was chosen as this represented a reasonable balance between being as small as possible but still maintaining a large enough number of SNPs to be meaningful for this dataset.

8.2.7 LD profile

The LD profile was generated from the maps created using LDhat. The idea was to find the expected r^2 values for pairs of SNPs given the genetic distance between them. As the aim was to find regions that are different to other regions of the genome, this was a way of calculating the expected values for this population. The create_LDprofile function of the zalpha package was used to create a genome-wide LD profile from a combination of all the LDhat maps. A bin size of 0.0001 cM was chosen, with a maximum distance of 2 cM between SNPs.

8.2.8 Candidate regions

Three statistics were chosen for further analysis: Z_α, $Z_\alpha^{r^2/E[r^2]}$, and $Z_\alpha^{BetaCDF}$. An empirical distribution was fitted to each of them, by ranking and dividing by the number of non-missing values. The top 0.1% of the SNPs from the empirical distribution of the $Z_\alpha^{r^2/E[r^2]}$ and $Z_\alpha^{BetaCDF}$ statistics were identified as candidate SNPs. This was a total of 230 SNPs. Manhattan plots were generated using the qqman v0.1.4 R package [579]. The Venn diagram was created using the VennDiagram v1.6.20 R package [580].

SNP locations were converted to CanFam2 coordinates using the liftOver tool [581] to allow comparison between these results and previously published results using the former system. The previously published results that were used to compare were: Axelsson *et al.* 2013 [562], vonHoldt *et al.* 2010 [566], Akey *et al.* 2010 [567], Vaysse *et al.* 2011 [568], Boyko *et al.* 2010 [582], Wang *et al.* 2013 [559], Freedman *et al.* 2016 [565], and Cagan and Blass 2016 [564]. There were 534 regions and SNPs reported in total across the papers, some of which overlapped. When converting from CanFam2 to CanFam3.1, 14 regions were lost due to regions being split or partially deleted in the new build, resulting in 520 regions.

Candidate SNPs were annotated using Ensembl's variant effect predicator (VEP) [583] and ChIPpeakAnno v3.22.2 [584]. For VEP, the default settings were used and the SNP locations were queried against the CanFam3.1 assembly, For ChIPpeakAnno, the package biomaRt v2.44.1 [585] was used to call the "clfamiliaris_gene_ensembl" dataset (CanFam3.1 assembly). The parameters select="all" and output="both" were chosen to return nearest features as well as features that overlap the SNP location. Where available, the gene symbol was returned; otherwise the Ensembl gene code was returned.

To validate results, the analysis was replicated using the dogs from the second biggest cluster in the principal components analysis. Only SNPs that were in the top 0.1% for the Z_α statistics in both clusters were considered to be replicated.

The figures for the candidate regions (see Figure 8-17 and Appendices A.4.4 and A.4.5) were created using the Gvis R package v1.32.0 [586]. The gene annotations are from the AnnotationHub R package v2.20.1 [587] Ensembl database "AH79907" and formatted using the ensembldb R package v2.12.1 [588]. The regions plotted include a 1 Mb buffer either side of the candidate SNP(s).

8.3 Results

8.3.1 Recombination map

Genetic maps were created for the dog genome using LDhat and converted to cM. Figure 8-8 shows the maps for the 38 autosomes. The map shows that recombination is not uniform across the genome, with clear peaks at recombination hotspots. The smoothed map demonstrates the upward trend predicted given the acrocentric nature of dog chromosomes. Chromosomes 27 and 32 are the exceptions to this trend; however, it is believed this is due to a reversal of these chromosomes in the CanFam3.1 build rather than a physical phenomenon [558, 589]. These results show that LD is not constant across the dog genome, and so when identifying potentially selected regions it would be prudent to take this into account.

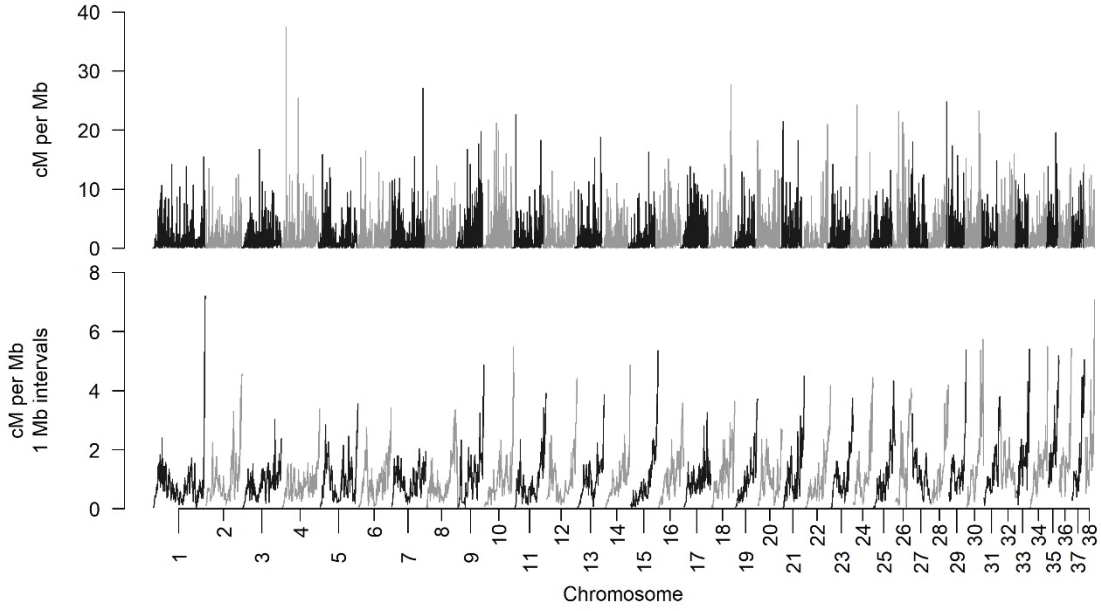


Figure 8-8 Recombination map of the dog genome

The top figure shows the recombination map for each of the 38 autosomes in the dog genome, in terms of centimorgans per megabase. The bottom figure shows the same data but aggregated into 1 Mb bins. Note the upward trend across all chromosomes apart from 27 and 32 - a previously reported reversal in the CanFam3.1 build.

8.3.2 Comparison of genetic maps

To validate the genetic maps generated here, a comparison was made to a previously published LD map created by Auton *et al.* [550] and a linkage map by Campbell *et al.* [558].

Figure 8-9 shows the cM map of chromosome 1 compared to the Auton map and the Campbell map. The Auton map is smoother overall as the SNP density is greater than the other two maps, with a density of on average 1,665 SNPs per Mb for the compared regions opposed to 70 SNPs per Mb in the LDhat output and 6 SNPs per Mb in the Campbell map. Maps for the rest of the chromosomes are supplied in Appendix A.4.2.

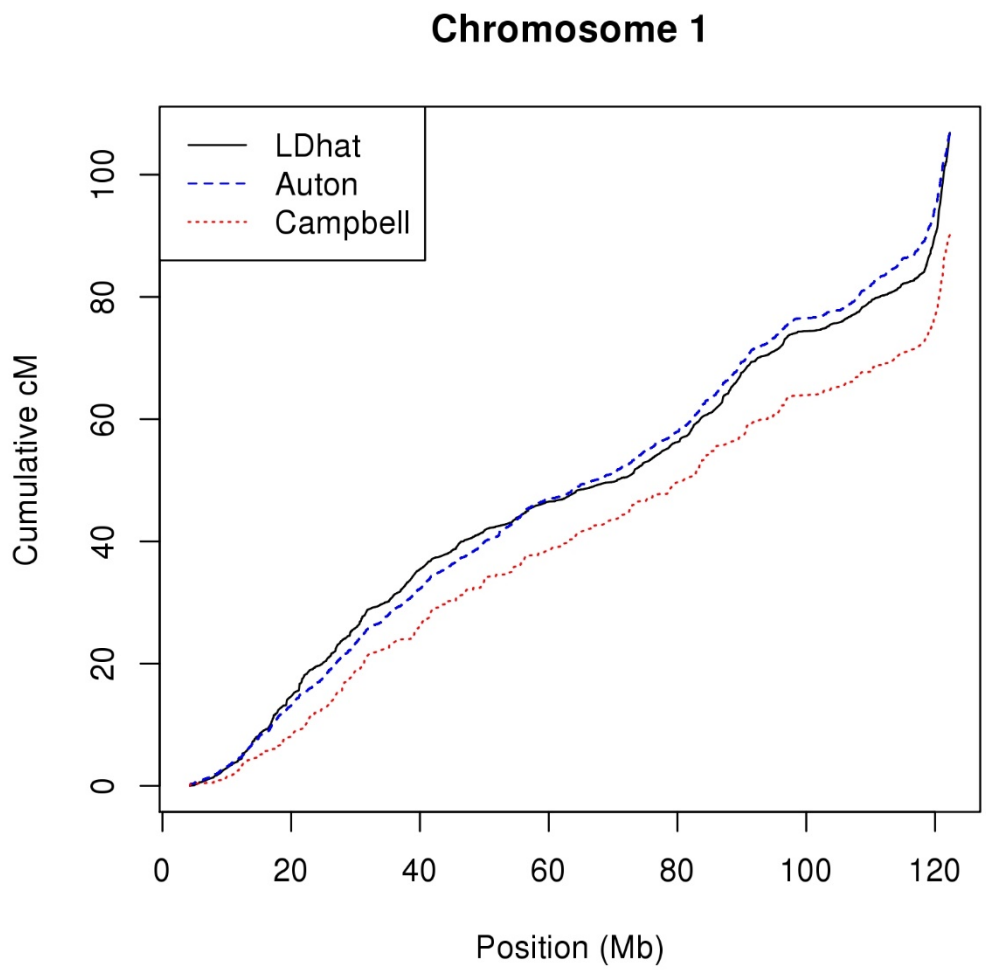


Figure 8-9 Genetic maps of chromosome 1

This figure shows the final LDhat map for chromosome 1 of the dog data. Also shown as a dashed blue line is the Auton LD map, and the Campbell linkage map is the dotted red line. All maps show a similar shape; however, the Auton map is smoother due to a higher SNP density, and the Campbell map is shorter.

The Campbell linkage map is shorter than the two LD maps. This is because linkage maps can only report recombination observed in present day families, whereas LD maps will also contain information on recombination events from previous generations. LD maps can also be affected by other factors such as selection. Table 8-3 shows the published length of these maps, and another LD map and linkage map recently published for dogs, illustrating the differences between the lengths of different types of map.

Table 8-3 Genetic maps for dogs

Authors	Year published	Map type	Total Length (cM)
Wong *et al.* [589]	2010	Linkage	2093
Axelsson *et al.* [548]	2012	LD	3005
Auton *et al.* [550]	2013	LD	2430
Campbell *et al.* [558]	2016	Linkage	1978

The maps generated here were converted to cM using the Auton map lengths for the range of SNP locations in the datasets, as they are the most recently published LD maps.

Wavelet analysis was used to compare the three maps. Firstly, the data were aggregated into bins of size 500,000 bp. This is larger than the bin sizes used for the similar human analysis in Chapter 5 due to the poor SNP density of the Campbell maps, and so a lot of the fine scale recombination information is lost. The data were log-transformed to adjust for extreme differences in recombination rates at hotspots. Finally, the wavelet transform was applied to generate wavelet coefficients.

Before considering the results of the wavelet analysis, correlations were calculated between the rates in the binned maps. Using Kendall's tau, the correlations between the LDhat maps generated in this chapter and the Auton map were 0.65, and compared with the Campbell linkage map, 0.4. The Auton and Campbell maps compared had a correlation of 0.43. These results show, unsurprisingly, that the LD maps are more correlated with each other than the linkage map at the 500 Kb scale. Of more interest is how the change along the chromosomes in the maps correlate, and any difference by scale.

Figure 8-10 shows the proportion of variance for each of the maps at each scale. All the maps show a negative trend as the scale increased. The higher variance in the lowest scale for the Campbell maps is possibly due to the dearth of SNPs in the original data. These results are consistent with findings in humans at these scales, as most of the variance is at the Kb scale and thus is hidden in this analysis.

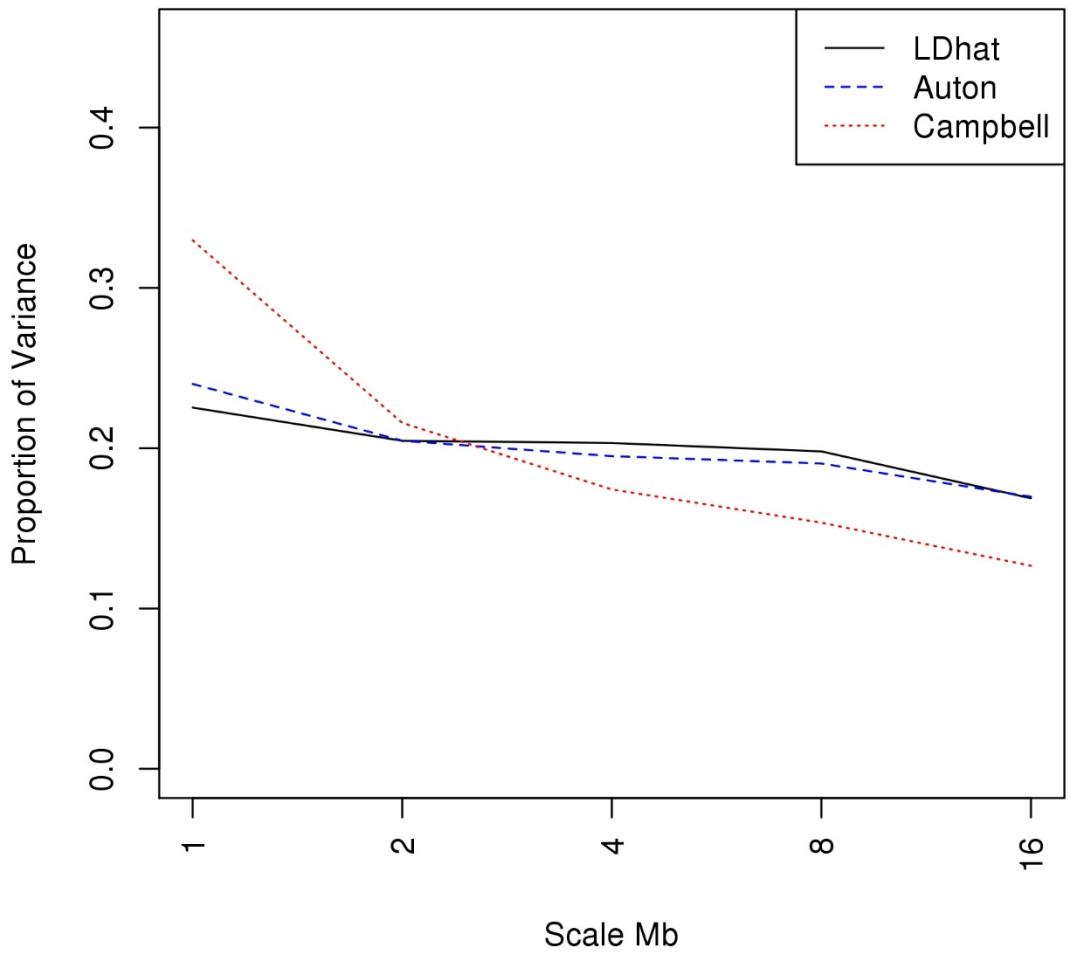


Figure 8-10 Proportion of variance in dog genetic maps

This figure shows the proportion of variance for each of the genetic maps: the maps produced in this chapter using LDhat in black, the Auton LD map in dashed blue, and the Campbell linkage map in dotted red. The proportion of variance is given for each scale from 1 to 16 Mb.

Figure 8-11 shows the correlations between the detail coefficients of each wavelet transformed map. Detail coefficients are proportional to the change in rates along a chromosome. Predictably, the correlations between detail coefficients are higher between the two LD maps than comparisons including the linkage maps at all scales. Each of the plots shows a positive trend as the scale increases, meaning the change on the wide scale is more consistent across maps than finer scales. This positive trend is congruous with the findings in humans.

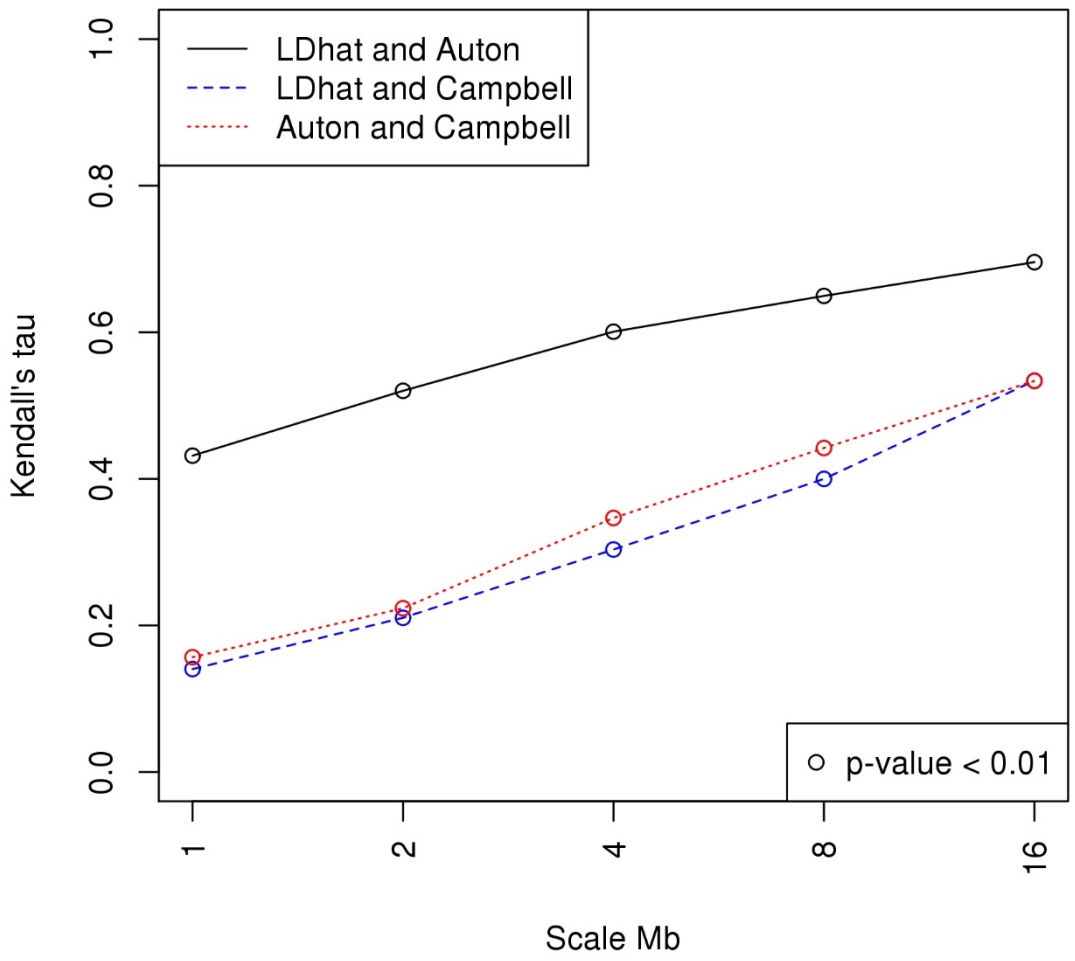


Figure 8-11 Correlation between wavelet coefficients of three genetic maps for dogs

This figure shows the correlation between the detail coefficients of the MODWT genetic maps. The black line shows the correlation between the maps generated in this chapter using LDhat, and the Auton LD map. The other two lines show the comparisons with the Campbell linkage map. Kendall's tau was used to assess correlation between detail coefficients.

This comparison has shown that the maps generated in this chapter are correlated with previously generated maps and show the expected patterns and changes at multiple scales. However, there are still questions around whether LD maps generated using the LDhat software are "true" recombination maps or whether they are significantly affected by other forces. Simulations and linkage maps created with larger datasets and thus at a finer scale may shed further light.

8.3.3 Candidate regions

To find candidate regions for selection, the Z_α statistics were applied to the autosomes. Figure 8-12 shows the results for Z_α. Highlighted on the graph above the line are all the SNPs who fall in the top 0.1% of the empirical distribution for the Z_α statistic for these dogs.

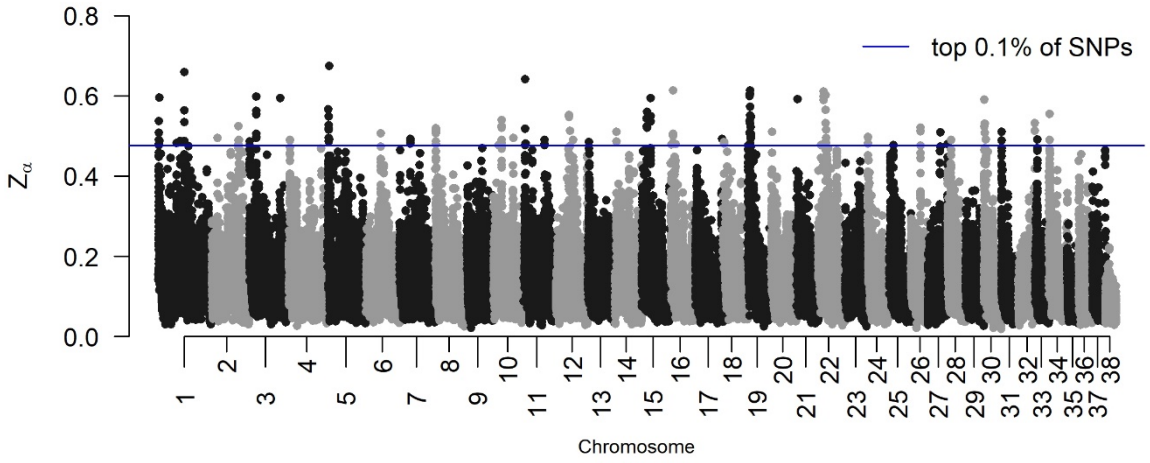


Figure 8-12 Z_α across the dog genome

This figure shows the results of the Z_α analysis across the 38 dog autosomes. SNPs above the blue line are in the top 0.1% of the Z_α empirical distribution and are classified as outliers.

There are some clear peaks that may indicate a selective event. However, this is before adjusting for expected squared correlations between SNPs. Regions of low recombination activity could imitate a selective sweep, and conversely in some areas high levels of recombination could mask a selective sweep.

An LD profile was created from the data. This contains the expected squared correlation between a pair of SNPs, given the genetic distance between them. Figure 8-13 shows a plot of the expected squared correlations from the LD profile. As expected, the larger the centimorgan distance, the less correlated SNPs are on average.

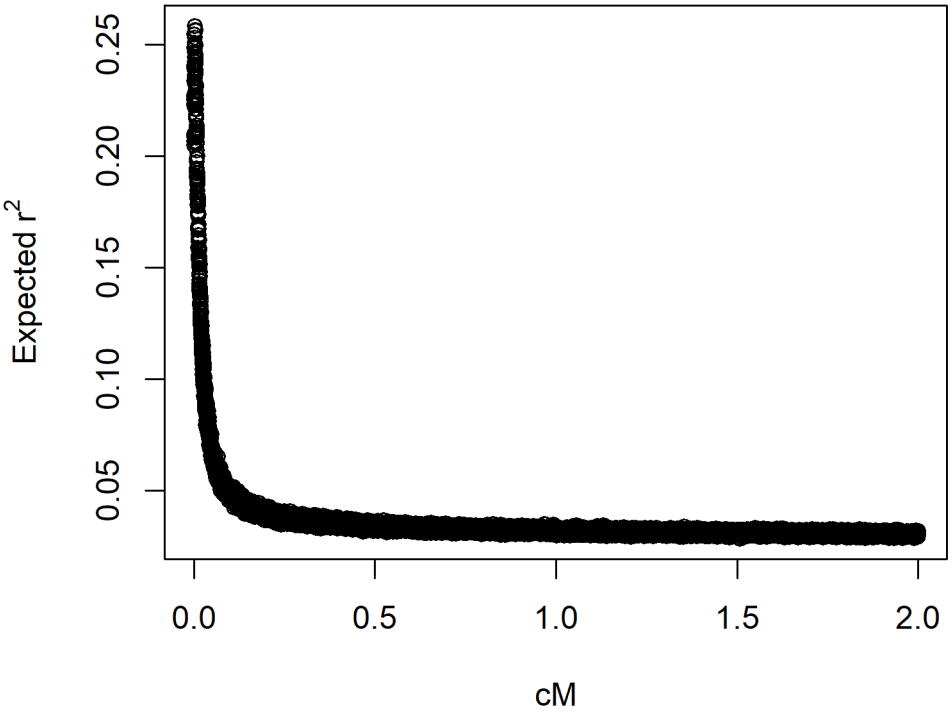


Figure 8-13 Plot of the LD profile

This figure shows the expected squared correlation (r^2) between SNPs given the genetic distance in cM between them. It is calculated using 20,000 bins of size 0.0001 between 0 and 2 cM. As the cM distance increases, the expected squared correlation decreases.

Using the LD profile, two more statistics were considered: $Z_\alpha^{r^2/E[r^2]}$ and $Z_\alpha^{BetaCDF}$. For definitions of these statistics see section 7.2. These statistics were chosen to cover a range of results: basic Z_α, Z_α with a simple adjustment for expected r^2, and Z_α with an adjustment based on the distribution of r^2. Statistics such as Z_α^{Zscore} were not considered as it is not believed the r^2 values for the dog data follow a normal distribution and are better represented by the Beta distribution. For SNPs to be considered as outlying candidates for this analysis, they must be in the top 0.1% of the empirical distribution for the two adjusted Z_α statistics.

The effect of adjusting for expected r^2 resulted in the loss of some of the candidate SNPs from the Z_α results. For example, Figure 8-14 shows the first 20 Mb section of chromosome 3. The first highlighted section A in green shows a region where a SNP that was previously a candidate has been adjusted downwards. The next region B highlighted in blue shows a SNP that previously was not a candidate but is now an outlier after adjusting for expected r^2. The final region C highlighted in pink shows a region containing SNPs that have remained candidates after adjusting for expected r^2.

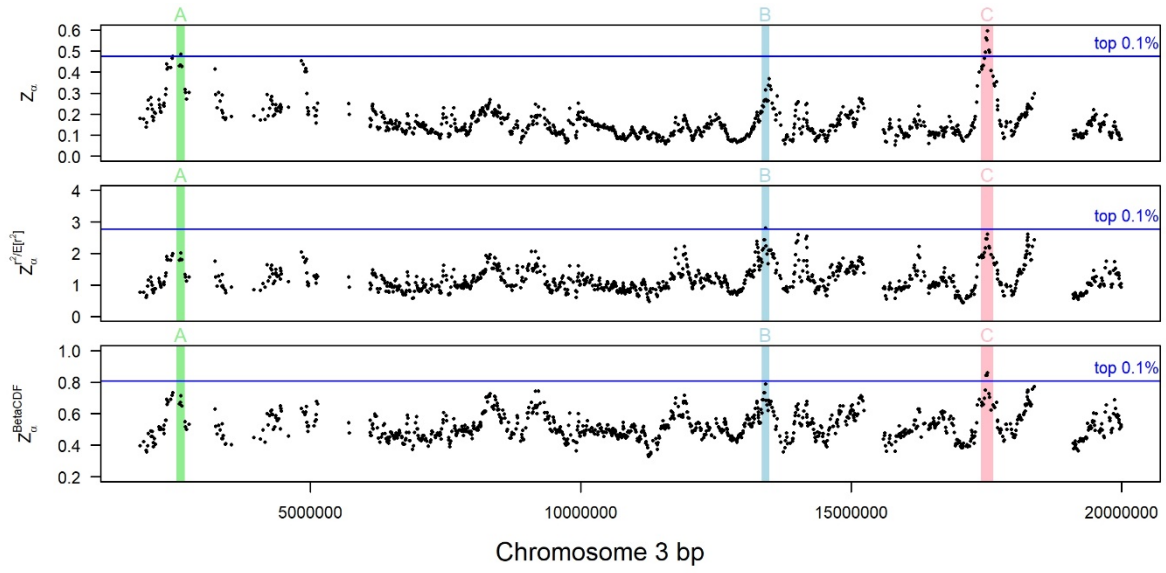


Figure 8-14 Plot of chromosome 3 for three statistics

This plot shows the values for the three statistics Z_α, $Z_\alpha^{r^2/E[r^2]}$ and $Z_\alpha^{BetaCDF}$ for the first 20 Mb region of chromosome 3. Any SNPs above the blue line are in the top 0.1% of SNPs in the whole genome for that statistic. Region A highlighted in green shows where a SNP that was an outlier for Z_α is not after adjusting for expected r^2. Region B highlighted in blue shows a SNP that has become a candidate after adjusting for expected r^2. The region in pink shows SNPs that have remained candidates.

There were 230 SNPs left in the candidate pool after adjusting for expected r^2, shown in Figure 8-15 below. Figure 8-16 shows the number of the 230 candidate SNPs that were identified by each statistic: 24% of SNPs were identified by both statistics, with the rest only identified by one or the other. The strict thresholding limited the overlap: the 87 SNPs identified only by $Z_\alpha^{BetaCDF}$ were all within the top 1% of the $Z_\alpha^{r^2/E[r^2]}$ results, and all but 3 SNPs of the 87 SNPs identified by $Z_\alpha^{r^2/E[r^2]}$ were in the top 5% of the $Z_\alpha^{BetaCDF}$ results. These candidate SNPs were then annotated according to overlapping or nearby genes using VEP and the R package ChIPpeakAnno. Thus, some SNPs have multiple genes associated with them. All 230 SNPs and relevant information can be found in Appendix A.4.3.

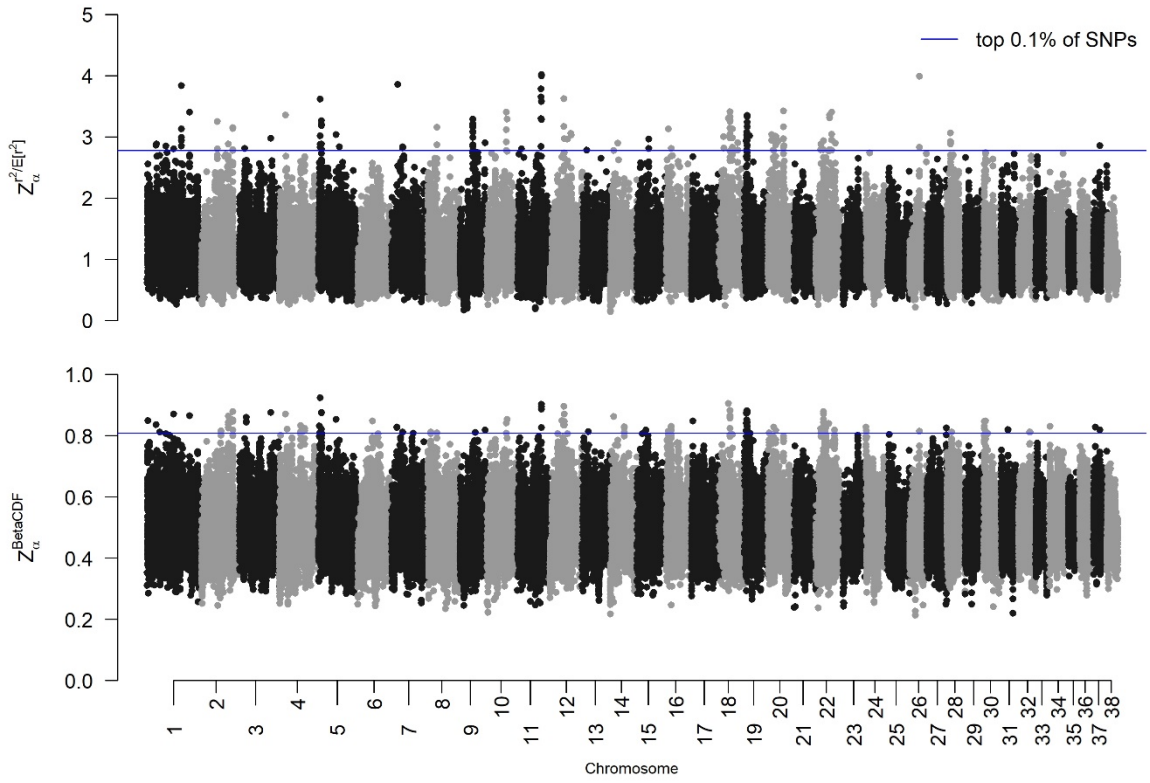


Figure 8-15 Manhattan plot of the final candidate SNPs

This figure shows the results for the $Z_\alpha^{r^2/E[r^2]}$ and $Z_\alpha^{BetaCDF}$ statistics across the 38 dog autosomes. SNPs above the blue line are in the top 0.1% of the empirical distribution for the statistic and are classified as outliers. There are 230 candidate SNPs after collating the outliers from both statistics.

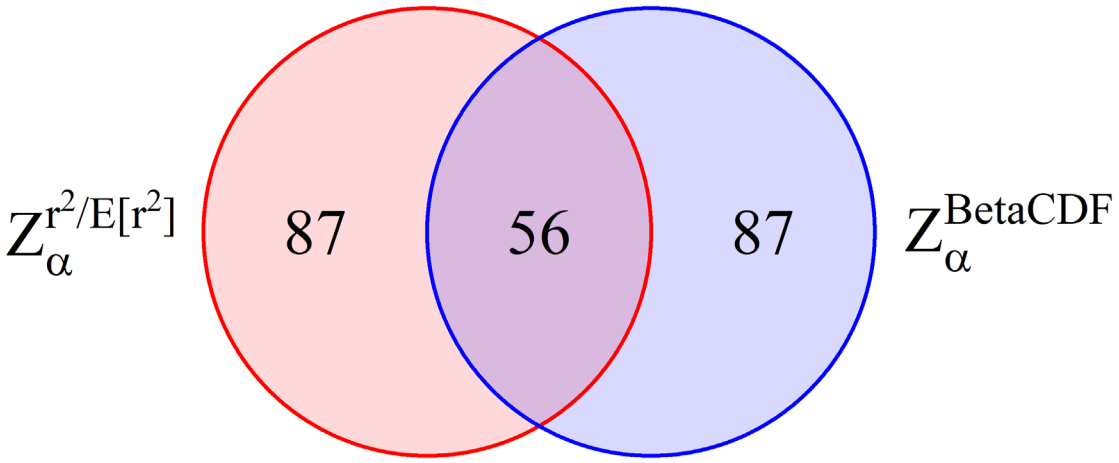


Figure 8-16 Count of the candidate SNPs identified by each statistic

This Venn diagram shows the 230 candidate SNPs split by whether they were in the top 0.1% of $Z_\alpha^{r^2/E[r^2]}$, $Z_\alpha^{BetaCDF}$, or both. 24% of the candidate SNPs were identified by both statistics.

Two approaches were used to validate the findings: the first was to compare the results to previous studies to find overlaps with previously published selected regions, and the second was to compare to a replication study. For replication, the dogs from the second cluster of the principal components analysis were processed in the same way as the cluster one dogs. If the same signal was found in the cluster two dogs, this gave more confidence that the results were not just random noise or exclusive to the dogs in the first cluster.

The results in this study were compared to other studies by finding overlapping regions, first converting the other studies to CanFam3.1 if applicable. Table 8-4 shows the other studies that the results were compared against. As can be seen from the table, most of the other studies compared dogs with wolves or across breeds to find candidate regions for selection. Thus, they are appropriate for validating this study as the methodology was very different.

Table 8-4 List of previous studies where selection was identified in the dog genome

First Author	Year	Build	Method	Ref
Axelsson	2013	CanFam2	Pooled heterozygosity (H_P) and F_{ST} with wolves	[562]
vonHoldt	2010	CanFam2	XP-EHH and F_{ST}, both with wolves	[566]
Akey	2010	CanFam2	d_i - function of pairwise F_{ST} between multiple breeds	[567]
Vaysse	2011	CanFam2	S_i, d_i and XP-EHH between multiple breeds	[568]
Boyko	2010	CanFam2	F_{ST} across breeds	[582]
Wang	2013	CanFam2	F_{ST} with wolves	[559]
Freedman	2016	CanFam3.1	$\Delta\pi$, F_{ST} and ΔTajima's D, all with wolves	[565]
Cagan	2016	CanFam3.1	F_{ST} with wolves	[564]

Table 8-5 shows contingency tables for all the SNPs, counting whether they were in the top 0.1% for the statistic or not, and whether they were overlapped by a previously published region or not. The table shows that when the SNPs are in the top 0.1% for either of the statistics, they have a significantly higher chance of overlap with a previously published study. It may be surprising that there were 8,981 SNPs overlapped by a previous study; however, several of the studies only published large windows. The mean window size reported was over 400,000 bp, with a maximum window of almost 2 Mb. Therefore, a large number of SNPs were overlapped by these wide regions where realistically only a few would be expected to be outliers. As well as this, the strict top 0.1% threshold for outliers means that many of the previously reported regions could well show evidence of sweeps in these statistics, but not quite to this extent.

Table 8-5 Contingency tables of SNPs in the top 0.1% and overlap with previously published regions

Each contingency table shows the number of SNPs split by the top 0.1% of the empirical distribution for the statistic. The columns show the number of SNPs falling within a candidate region from a previously published study. The brackets show the row percentages. The χ^2 test was conducted in R and resulted in a p-value < 2.2e-16 for both contingency tables.

$Z_\alpha^{r^2/E[r^2]}$	No overlap	Overlap
Bottom 99.9%	133,117 (94%)	8,936 (6%)
Top 0.1%	98 (69%)	45 (31%)
	133,215 (94%)	8,981 (6%)

$Z_\alpha^{BetaCDF}$	No overlap	Overlap
Bottom 99.9%	133,111 (94%)	8,942 (6%)
Top 0.1%	104 (73%)	39 (27%)
	133,215 (94%)	8,981 (6%)

There were 11 regions in total that were validated by both methods (replication in cluster two from the PCA and overlapped by a previous study). These can be found in Table 8-6.

Table 8-6 Regions containing signals of selection

This table shows the regions containing least one outlying SNP, where the region has been supported twofold by also containing an outlying SNP in the replication group (the dogs in cluster two from the PCA) and having been previously published. The genes listed are those within 250 Kb either side of the region or SNP, to match the 500 Kb windows reported by vonHoldt *et al*. [566] and Cagan and Blass [564]. Only genes with symbols are listed; genes with only Ensembl names were discarded.

Chromosome	SNP/Region (bp)	Named Genes (up to 250 Kb away)
2	61876498-61901702	*RPGRIP1L, FTO*
4	57366377	*G3BP1, ATOX1, GLRA1*
5	4064061-4093514	*U6, SNX19*
6	33510473	*CARHSP1, PMM2, ABAT, METTL22, USP7, LITAFD*
7	24652821-24664438	*U6, RABGAP1L, GPR52*
10	44372549-44388924	*INPP4A, VWA3B, TMEM131, CNGA3*
11	54324689-54391443	*SNORD42, SHB, EXOSC3, DCAF10, ALDH1B1*
15	20317533	
16	7462818	*U4, U6, CLEC5A, PRSS37, SSBP1, DENND11, AGK, MGAM, TAS2R38, TAS2R3, WEE2, TMEM178B, TAS2R4*
19	4813917-6590666	*PCDH18*
30	1558195-1732646	*RYR3, FMN1*

Figure 8-17 shows the region in chromosome 11 with a 1 Mb extension either side. This region contains eight SNPs that were outliers in the $Z_{\alpha}^{r^2/E[r^2]}$ statistic, including the most outlying SNP for this statistic at 54,368,623 bp, and four for the $Z_{\alpha}^{BetaCDF}$ statistic. There is a clear, strong signal for selection; however, there are many genes in the region, so it is hard to ascertain which, if any, is being selected for. There are 11 known genes in the region, five of which have symbols and the other six identifiable only by their Ensembl gene name. Three of the genes are RNA genes, and the other eight are protein coding.

The empty region to the left contains a dearth of SNPs. The SNPs that were present were either filtered out because their MAF was below 5% or were not calculated, due to the restrictions on the minimum number of SNPs needed in the left and right sets when calculating the Z_{α} statistics. This could be due to random chance, or that the region is challenging to sequence e.g. due to repeated elements [590]; however, it could indicate that the region is highly conserved i.e. the region contains code vital to life and so any mutations are fatal and removed from the population. The sweep itself could be the reason the region is so sparse, as when nearby variants hitchhike along with the beneficial variant this reduces the diversity of the surrounding regions, see section 2.2 where this was discussed. The low level of recombination estimated in the area supports this theory. LDhat was tested on simulations with low density SNPs and it was not found to under-estimate recombination in these scenarios [348].

To assess the progression of the sweep, two different Z_{β} statistics were calculated based on the Z_{α} statistics. They were applied in two different ways, one by taking away from the equivalent Z_{α} statistic and one by dividing. The final two statistics presented are $Z_{\alpha}^{r^2/E[r^2]} - Z_{\beta}^{r^2/E[r^2]}$ and $\frac{Z_{\alpha}^{BetaCDF}}{Z_{\beta}^{BetaCDF}}$. The expectation for these statistics is that while a sweep is in progress, they should return values that are similar to or less than neutral regions. However, when a sweep is nearing fixation, they should return values that are much higher than neutral regions. In this graph a line has been drawn for the median value for these statistics as a proxy for neutrality. For both statistics, the values do deviate from the median towards the positive end of the scale, indicating that the sweep in the region is near fixation.

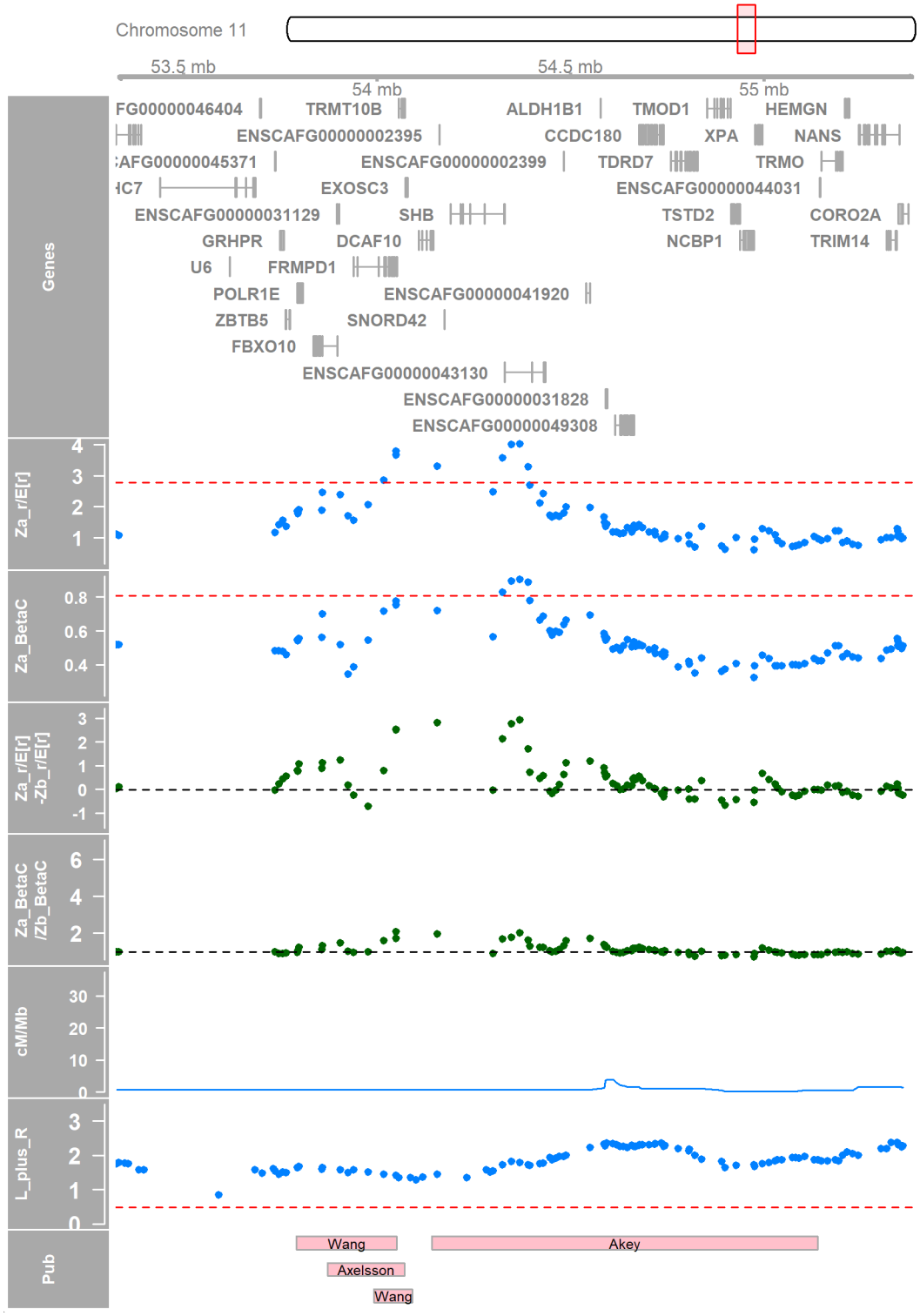


Figure 8-17 A candidate region of chromosome 11

This combined plot shows a region of chromosome 11 containing evidence for a selective sweep. From the top: An ideogram of chromosome 11 highlighting the region; the genome axis track; the genes in the region from Ensembl; plot of $Z_{\alpha}^{r^2/E[r^2]}$ with a red line indicating the top 0.1% of values; plot of $Z_{\alpha}^{BetaCDF}$ with a red line indicating the top 0.1% of values; plot of $Z_{\alpha}^{r^2/E[r^2]} - Z_{\beta}^{r^2/E[r^2]}$ with a black line indicating the median; plot of $\frac{Z_{\alpha}^{BetaCDF}}{Z_{\beta}^{BetaCDF}}$ with a black line indicating the median; the recombination rate in cM/Mb; plot of $\log_{10}\left(\binom{|L|}{2} + \binom{|R|}{2}\right)$ with a red line showing the bottom 0.1%; previously published candidate regions.

A further 26 regions containing signals were found in this study and replicated in the second cluster from the PCA but are unique to this study compared to the others. These are listed in Table 8-7.

Some of the regions (for example 17:3753156, see Appendix Figure 37) do not have any named genes within 250 Kb either side. This could be because one of the genes with only an Ensembl designation currently is under selection, or the gene under selection is linked but is further away, or that there is an epistatic or epigenetic effect of a variant in this region.

The small nuclear RNA (snRNA) gene *U6* appears many times in the results; however, this gene appears multiple times in the dog genome, and most are likely to be pseudogenes [591]. There is also evidence that it is highly conserved when the mammalian version is compared with yeast [592]. Thus, it is unlikely that the *U6* gene is under positive selection, and only appears in the results multiple times due to its ubiquity.

Figures for each of the regions in Table 8-6 and Table 8-7 can be found in Appendices A.4.4 and A.4.5 respectively.

The diversity statistics may also contain information on selection. For this analysis the statistic $\binom{|L|}{2} + \binom{|R|}{2}$ was used, which can be interpreted as the sum of the number of pairs of SNPs on each side of the target SNP. An example of where this is useful is around the *MBP* gene on chromosome 1, see Figure 8-18. This figure shows that there is a large region where no Z_{α} statistics could be calculated due to the dearth of SNPs. The diversity statistic shows an outlier, which is replicated in the second cluster, and many of the previous studies have also identified this region as potentially under selection. The *MBP* gene is implicated in behavioural changes, with a fixed variant between dogs and wolves the suggested cause [564].

Table 8-7 Regions unique to this study containing a selection signal

These are the regions containing at least one outlying SNP and an outlying SNP in the replication group but were not published in any of the studies in Table 8-4. The genes listed are those within 250 Kb either side of the region. Only genes with symbols are listed; genes with only Ensembl names were discarded.

Chromosome	SNP/Region (bp)	Named Genes (up to 250 Kb away)
1	43001368	*SYNE1, VIP, MTRF1L, RGS17, FBXO5, MYCT1*
1	96115461	*SYK*
2	71434345	*U6, EPB41, PTPRU, MECR, SRSF4*
3	17490492-17516194	*U6, U1, ARRDC3*
3	72708942	*PDS5A, UBE2K, UGDH, LIAS*
4	17518453	*CTNNA3*
4	57345395	*G3BP1, GLRA1*
5	6838932-6859691	
5	40202215	*U6, SPECC1, AKAP10, ULK2*
8	7735497	
9	29752455	*U6*
10	46053118	*THADA, PLEKHH2, DYNC2LI1, ABCG8, ABCG5*
12	26284264	*KHDRBS2*
12	31691990-31835704	*U6, ADGRB3*
14	8117811	*MIR129-1, FAM71F1, FAM71F2, IMPDH1, RBM28, LEP, PRRT4, SND1, OPN1SW, LRRC4*
17	3753156	
18	29595073	
19	7095253-7122489	*U4*
20	13387022	*SUMF1, SETMAR, LRRN1*
22	11073667-12039716	*U6*
22	18774821-19925395	*U6*
22	31194138-31347124	*SCEL, EDNRB, SLAIN1*
26	22151015-22156289	*U6, TTC28, HSCB, CCDC117, ZNRF3, C26H22orf31*

27	44328723	*CACNA1C*
30	4822803	*MEIS2*
32	24657487-25070561	*U6, U2, SLC9B2, BDH2, CENPE, TACR3*

.

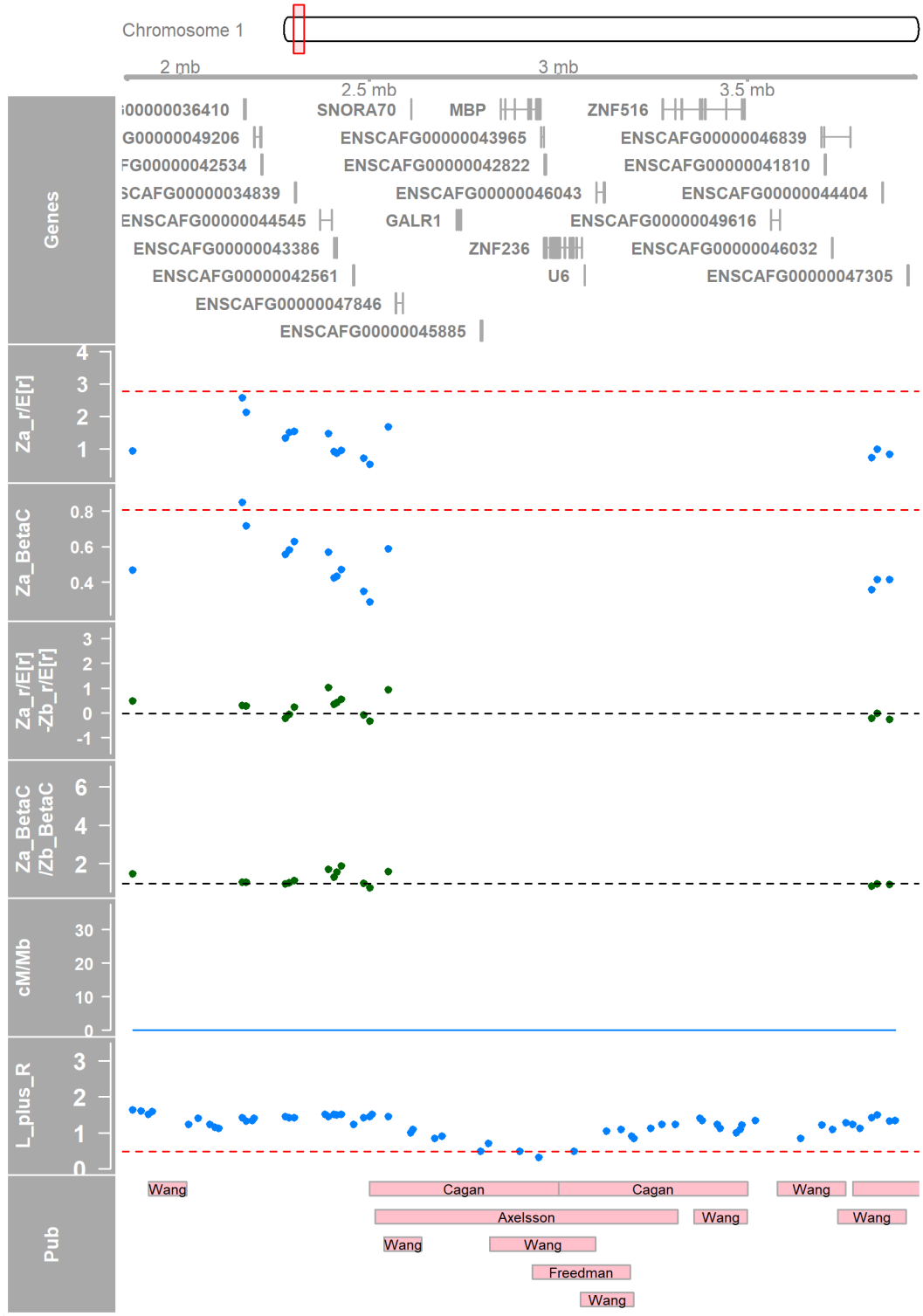


Figure 8-18 The region of chromosome 1 around the *MBP* gene

This combined plot shows a region of chromosome 1 containing evidence for a selective sweep. From the top: An ideogram of chromosome 1 highlighting the region; the genome axis track; the genes in the region from Ensembl; plot of $Z_{\alpha}^{r^2/E[r^2]}$ with a red line indicating the top 0.1% of values; plot of $Z_{\alpha}^{BetaCDF}$ with a red line indicating the top 0.1% of values; plot of $Z_{\alpha}^{r^2/E[r^2]} - Z_{\beta}^{r^2/E[r^2]}$ with a black line indicating the median; plot of $\frac{Z_{\alpha}^{BetaCDF}}{Z_{\beta}^{BetaCDF}}$ with a black line indicating the median; the recombination rate in cM/Mb; plot of $\log_{10}\left(\binom{|L|}{2} + \binom{|R|}{2}\right)$ with a red line showing the bottom 0.1%; previously published candidate regions.

8.4 Discussion

The results showed that it is possible to find candidate regions for selective sweeps using the Z_{α} statistics. Using these statistics, multiple regions have been found containing extreme outlying values, both replicated from previous studies and novel to this one. Recombination was shown to be variable in the dog genome, with clear hotspots even without a functioning *PRDM9* gene, and a general increase in recombination activity along each of the chromosomes due to their acrocentricity. This study also showed that it was important to correct for this variable recombination rate, as these adjustments changed which SNPs were considered extreme outliers.

Setting the threshold at 0.1% for outliers was a deliberate choice to only return the extreme values of the statistics and thus reduce the chance of reporting false positives. Requiring replication of the signals in the second cluster of dogs further shrunk the pool of results. Therefore, while confidence in the reported regions is high, the results here are almost certainly an under-estimate of the number of regions under selection within the dog genome. Weakly selected regions are unlikely to have been identified, nor selection that has recently begun to sweep. Polygenic traits under selection, where multiple variants across the genome may be sweeping synchronously but weakly, are also likely to have been dismissed. The region around *SEMA3D* gene has been identified in multiple studies as a candidate region. This region was identified as a candidate for selection in this study, see Appendix A.4.3, but it was not in the top 0.1% of the second cluster and so was deemed not replicated. However, the region did fall within the top 0.4% of the second cluster results, and so is almost certainly an example of where the strict thresholding decision has resulted in type II errors.

While this study found many candidate regions showing the pattern expected from a selective sweep, it is a hard to know the exact cause of the sweep. The SNP with the highest Z_{α} value in a region is not necessarily the cause of the sweep: genetic drift means there will always be a

random element creating noise in population statistics. There could be nearby SNPs that are not present in the sample or sweeps that have reached fixation and so will not be polymorphic anymore. Hitchhiking means any one of the nearby variants could be the beneficial variant. Epistatic effects mean that the variant could be affecting the regulation of a gene in an entirely different region. It is possible to look at the genes nearby the novel regions identified by this study, look up their function and apply a biological meaning as to why they have been selected for; however, this could be misleading and storytelling should be avoided [377]. Further study would be required to fully characterise the sweeps.

Complex demography can alter patterns in the genome, and this is true of dogs. Inbreeding can cause an increase in LD and runs of homozygosity where large haplotypes are inherited from both parents originally from a common ancestor and thus are identical by descent [571]. However, an arguably bigger effect on LD is the bottlenecks that some dog breeds have experienced, decreasing the effective population size N_e [536]. For some breeds the effective population size has been estimated as less than twenty [593, 594]. When admixture first occurs between inbred lines, the amount of LD will increase [143]. However, this is expected to reduce over a short amount of time [149]. For this study efforts were made to ensure dogs were mixed breed and from the same population to reduce the risk of these effects altering the results.

Given the variety of phenotypes observable in present day dogs it is not surprising that strong selection signals can be found for some of the major differences between breeds. Because of the short time since breeds were formed by breeders strongly selecting for particular traits, these traits are often governed by few genes with considerable effects [595, 596]. For example, there is evidence the *IGF1* gene has been selected for in small dogs [597], and the variation in coat types between many breeds of dogs can be reduced to variants in just three genes [598].

Finding evidence of phenotypic traits that modern domestic dogs have in common but are unique to dogs is perhaps a harder task; however, previously published studies have found evidence to suggest sweeps have occurred in genes involved in behavioural and neurological traits, as well as in digestion and metabolism to adapt to an omnivorous diet [559, 562, 564]. One gene identified in the literature is *MGAM* which is involved in digestion, specifically in converting starch into glucose [599]. This gene is located on chromosome 16 and was identified in Table 8-6 as being within 250 Kb of a candidate SNP from this study. While the candidate SNP identified does not overlap this gene, it is likely in linkage with the *MGAM* gene and has hitchhiked along. The region overlapping the *MGAM* gene was SNP-poor for this study (see Appendix Figure 19) so it would have been impossible using this data to pinpoint this exact gene under selection. As this area of the genome was identified by this study, as well as three of the other studies compared here

(Axelsson *et al.* [562], Cagan and Blass [564], and Wang *et al.* [559]), there is strong evidence that there has been a selective event in this region.

Another pair of genes mentioned in the literature as being selected for are *ABCG5* and *ABCG8*. These genes are involved in the digestion of dietary cholesterol and the removal of unusable components from ingested plants [600]. There is evidence for selection in both humans and dogs [559, 601]; potentially an example of parallel evolution as both lived in the same environments and consumed the same plants. These genes were identified in Table 8-7 as being nearby a candidate SNP on chromosome 10. While the candidate SNP itself was not overlapped by any previously published studies, it is clear when viewing the SNP in context (see Appendix Figure 33) that the nearby SNPs show a pattern reminiscent of a selective sweep and were overlapped by a previous study. The strict criteria used in this study means that, even though this region was identified by proximity, some other regions with sweeps previously identified will have been missed.

A recurring problem identified in this chapter are SNP-poor regions where Z_α cannot be calculated. This is somewhat mitigated by using diversity statistics to identify regions where the SNP count drops and thus there may be a sweep. However, this relies on there being at least one SNP to assign the statistic to. The diversity statistics also showed a bias towards the front of the chromosomes: for $\binom{|L|}{2} + \binom{|R|}{2}$ almost half (46%) of the 139 outliers (the bottom 0.1% of SNPs) were in the first 5% of the chromosome. SNP density is correlated with recombination rate through the influence of selection [602, 603], and as shown in Figure 8-8, recombination increases along the chromosome in dogs. SNP density is also known to be lower in regions of repeat DNA [604]. Thus, the diversity statistics cannot be used in isolation to find evidence of sweeps; however, they are useful in conjunction with other information, as was shown when considering the region around the *MBT* gene.

Most of the other studies used F_{ST} to compare dogs to wolves. This study has shown that many of the signals identified using this method can be recovered using a single-population method. This is encouraging as this means these statistics would be suitable for populations that do not have an appropriate outgroup or data available. Discrepancies were often due to a lack of SNPs in the region, which will become less of a problem as higher density SNP arrays and whole genome sequencing become more accessible. Discrepancies could also occur due to false positives within the F_{ST} results, for example F_{ST} has been shown to be inflated in regions with low recombination, a confounder specifically adjusted for by the Z_α statistics used here [605, 606]. Some of the studies also used very small sample sizes, especially for their wolf comparison population, which could affect results. Other differences could be explained by Z_α identifying adaptations present in both dogs and wolves, which would not be identified using the F_{ST} method [324].

The Z_α family is a development on other LD-based statistics such as Kelly's Z_{nS} [258] and ω [327] as it adjusts for recombination by assessing expected squared correlations based on the cM distance between SNPs. However, it does not explicitly consider other confounders that are considered in other methods, for example background selection in SweepFinder2 [277]. While Z_α examines fluctuations in LD, other methods identify different signals of selection, for example the singleton density score (SDS) which calculates the distance to the nearest singleton mutation [268]. Therefore, it is advisable to use a variety of methods, including multiple populations if available, to obtain the fullest picture of selection from the data.

Dogs are a model organism and as such are well studied and there are many genetic datasets for them [607]. However, it would be useful to see how well the Z_α statistic performs with smaller datasets as may be typical for non-model organisms being studied today. Non-model organisms are also more likely to not have a suitable outgroup available for analysis, making a single population statistic like Z_α more attractive. Future work could involve simulating datasets and taking smaller samples to see how well the Z_α statistic performs. The work could also include taking a subset of the dog data from this chapter to see if the same conclusions can be drawn, and at what point there are too few genomes to generate meaningful results.

In conclusion, this chapter has shown that the Z_α statistics can be used to find candidate regions for selective sweeps in the genomes of dogs. It has been shown that it is possible and important to adjust for recombination when using statistics based on LD. The use of multiple statistics using different methods would improve the results and help to further characterise the sweeps found. Further work would be required to fully investigate the candidate regions. Increased resolution in the regions and refinement of genes is necessary to find the specific evolutionary benefit and prove a selective event has occurred.

Chapter 9 Conclusion

The aim of this thesis was to develop and apply methods designed to detect regions in the genome under selection. This was achieved by reviewing the current field and identifying a method that could be developed further: the Z_α family of statistics. A tangible difference has been made to the field via the development of new software for applying the Z_α statistics. While investigating LD and recombination, new information was found regarding recombination rates across the genomes of different populations. Finally, the Z_α statistics could then be applied to real-world data, and new candidate regions of the genome of the domestic dog were identified.

Reviewing the methods made it clear that there are many different signals in the genome that can be detected that could be indicative of selective activity, for example, linkage disequilibrium patterns, fluctuations in the site frequency spectrum and patterns of singleton mutations. Many methods rely on comparisons with other species or populations, and so methods that are effective with single population samples are advantageous as they avoid the need to acquire extra data from outgroups, if indeed a suitable outgroup exists. There are also a variety of requirements around the data, for example requiring phased data, knowledge of the ancestral or derived status of SNPs, and B-value maps. These varying requirements necessitate many different methods that can be used in different scenarios. Some methods are more effective at detecting specific types of sweeps, such as hard, soft, in progress, or complete. By combining statistics or using multiple statistics researchers should be able to extract the most information from the genome.

While publishing methods and statistics is worthwhile, without corresponding software and applications it can be challenging for other researchers to apply the method to their own data. For reproducibility, reliability, and transparency, it is worth creating publicly available, open source software for new methods. The Z_α statistic was identified as a method worth developing further because it performed well and had potential, as shown in Chapter 3, but did not have any publicly available software. The novel use of an LD profile is its distinguishing feature as this is an effective way of adjusting for variable recombination rates, which can otherwise confound results.

Before adjusting for recombination using the Z_α statistic, it was important to understand it first. Chapter 4 and Chapter 5 were dedicated to analysing recombination rates, and especially to ascertaining whether recombination maps are portable between populations or whether they are private to the population. The conclusion from this work was that recombination rates are highly variable, both across the genome and within populations, and therefore it would be more accurate to create recombination maps for each population rather than using one as a proxy for another. It was also interesting to note that when analysing recombination rate by scale,

populations were most correlated at the wider scale as opposed to fine scales. The finer scale the analysis, the less likely it would be that a recombination map from another population would be fit for purpose; however, if analysis is on the wide scale, i.e. >1 Mb, maps have broadly similar patterns. This fits with what is known about the general patterns of recombination and the evolution of recombination hotpots.

The Z_α statistics were coded into a new R package called zalpha. R was chosen as it is a popular, open source software environment commonly used for biostatistics. The software was written with the goals of being transparent, reliable, and reproducible. For transparency, the software is all open source and the development version is hosted publicly for feedback and collaboration. The package has been peer-reviewed, both by CRAN where it is hosted, and by reviewers for the paper announcing the package. The functions in the package have been thoroughly tested, and the tests are available in the GitHub repository and through the Travis CI service. CRAN also sends warnings if a package may lose or change functionality, due to modifications in base R for example, so it should always be reliable and consistent. By submitting the package to CRAN, it is now fully accessible to any R user who wishes to use it, and is fully documented, with a manual and a vignette containing worked examples.

The zalpha R package is a useful contribution to the field as it is the first publicly available software for using the Z_α family of statistics. It also allows the user to create and use an LD profile, one of the key benefits of the Z_α method. By taking variable recombination rates into account, predictions regarding regions undergoing sweeps can be improved, as shown in Chapter 7. The new ability to easily generate Z_α statistics in R means that it is now straightforward to compare and combine these results with those from other sweep-detecting methods.

The next logical step was to apply the Z_α statistics to a real-life dataset. For this, the domestic dog was chosen as they have a different evolutionary path to humans yet are complementary as they have lived in the same environments. Using the zalpha package on this data garnered results that affirmed regions previously published, both providing replication support for the previous results, and confirming the proficiency of the Z_α statistics to identify selected regions. The results also highlighted novel candidate regions, suggesting that these regions way warrant further study.

Finding evidence of selective sweeps is hard, as confounders such as demography and recombination inhibit the ability to easily spot patterns in the genome. However, as technology continues to improve, so does the ability to store and process larger datasets in increasingly innovative ways. More data not only means larger datasets with more samples and at higher resolution, but also more populations and species. There is a clear need for simple, easy to interpret and easily accessible methods to analyse these new datasets. The zalpha software is

ideal for this analysis, as the software is open-source and publicly available, and thus is easily applied and results are replicable and reproducible.

The increased understanding of how LD methods can adjust for variable recombination rates will be useful for analysing these datasets. It was shown that each population must be treated as independent from others when utilising recombination maps, and that maps are not interchangeable. Using the information contained within an appropriate recombination map, for example via an LD profile, is a valid method for avoiding the confounding of results.

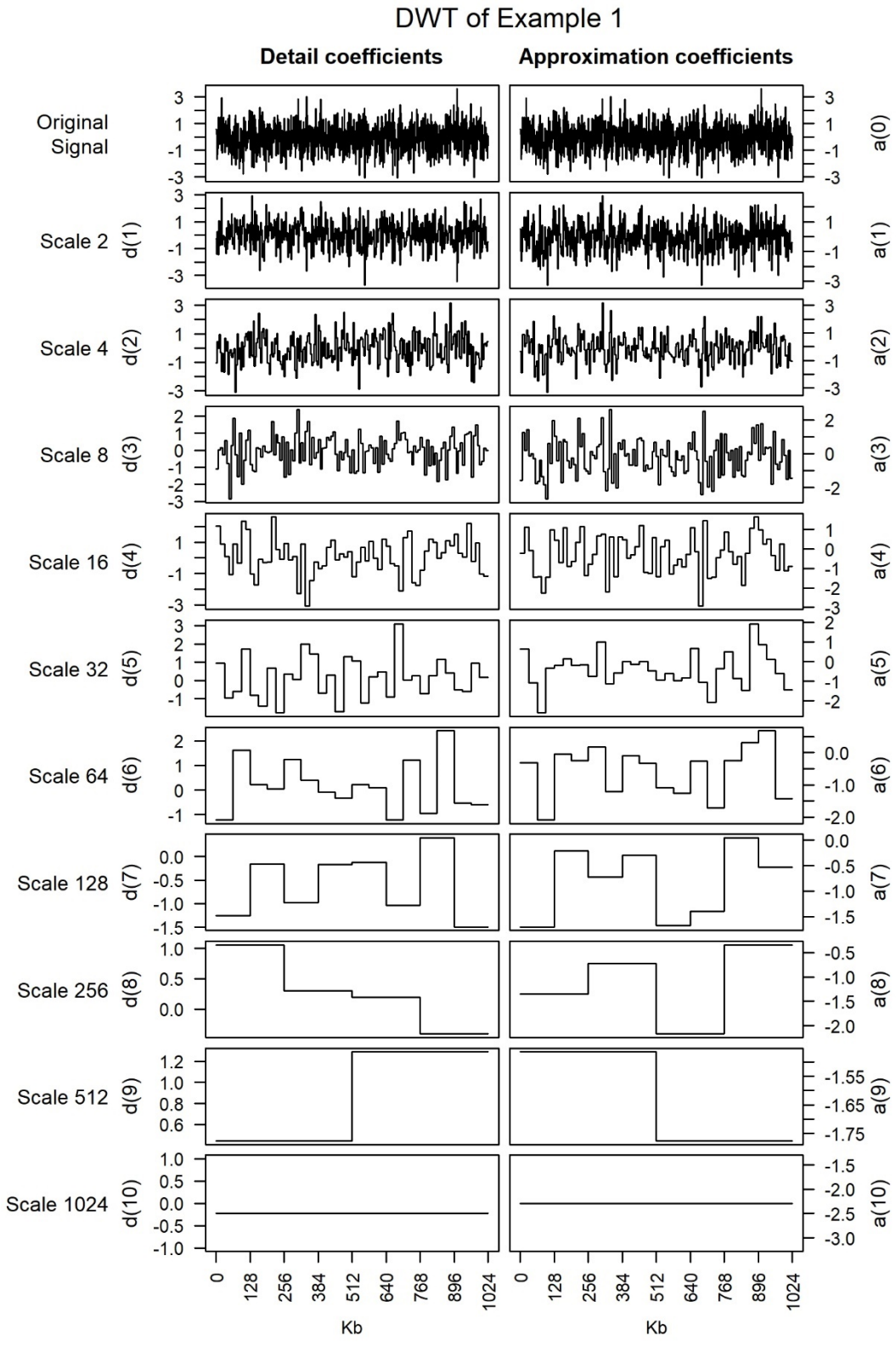
This thesis showed that the Z_{α} statistics could be used to add specific knowledge to the literature about candidate regions for selection in the domestic dog genome. As more species and populations are sequenced, more knowledge can be gained about their evolutionary processes and the way their phenotypes developed by applying these methods. Given the vast diversity of life on Earth, methods such as these will be vital to understanding all our differences, similarities, and origins. As new selective pressures arise, such as climate change and global pandemics, methods like these will aid us in understanding how humans, our food, and our animal companions can adapt and ultimately overcome the challenges the future may bring.

Appendix A Supplementary material

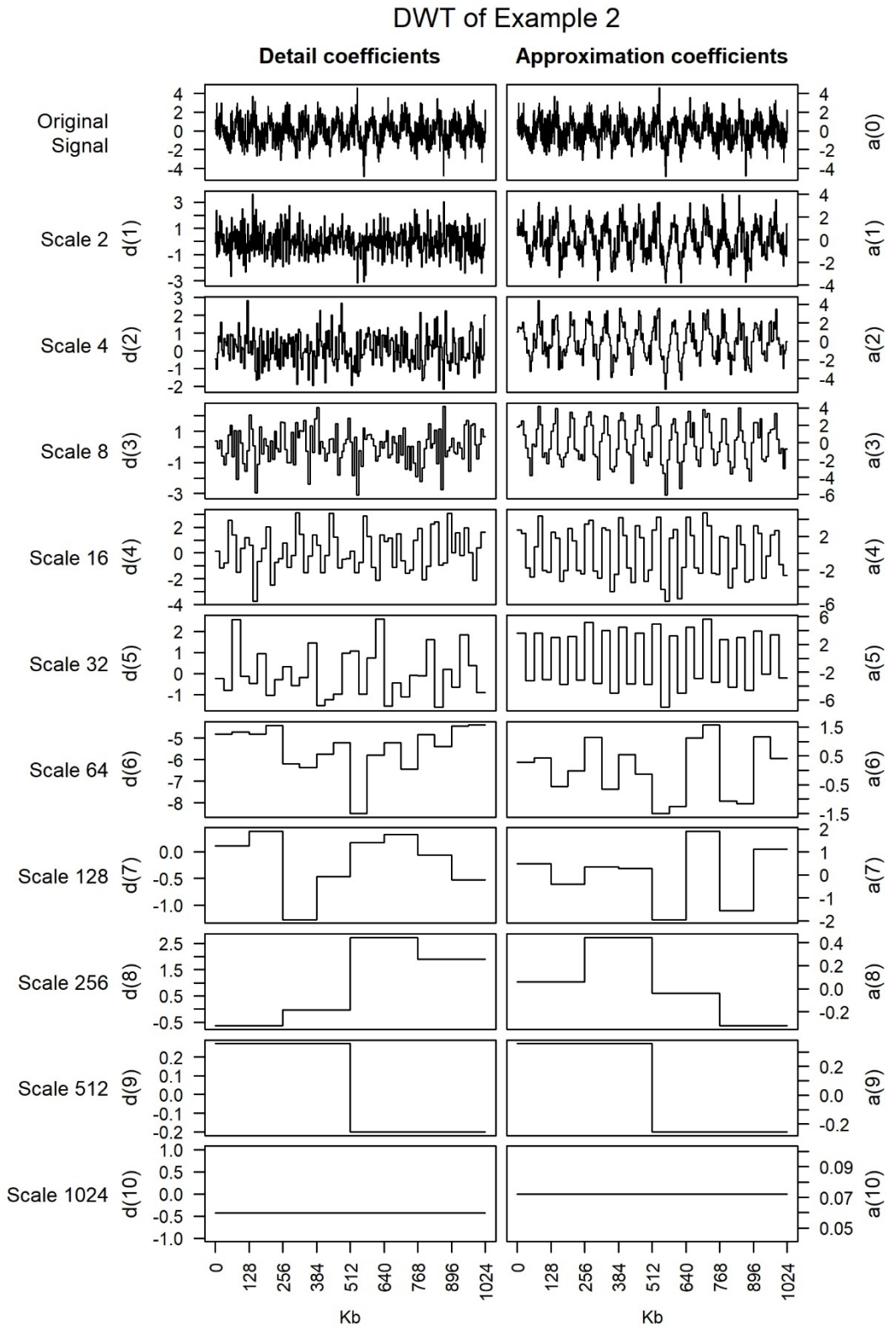
A.1 Supplementary material for Chapter 5

A.1.1 Discrete wavelet transform

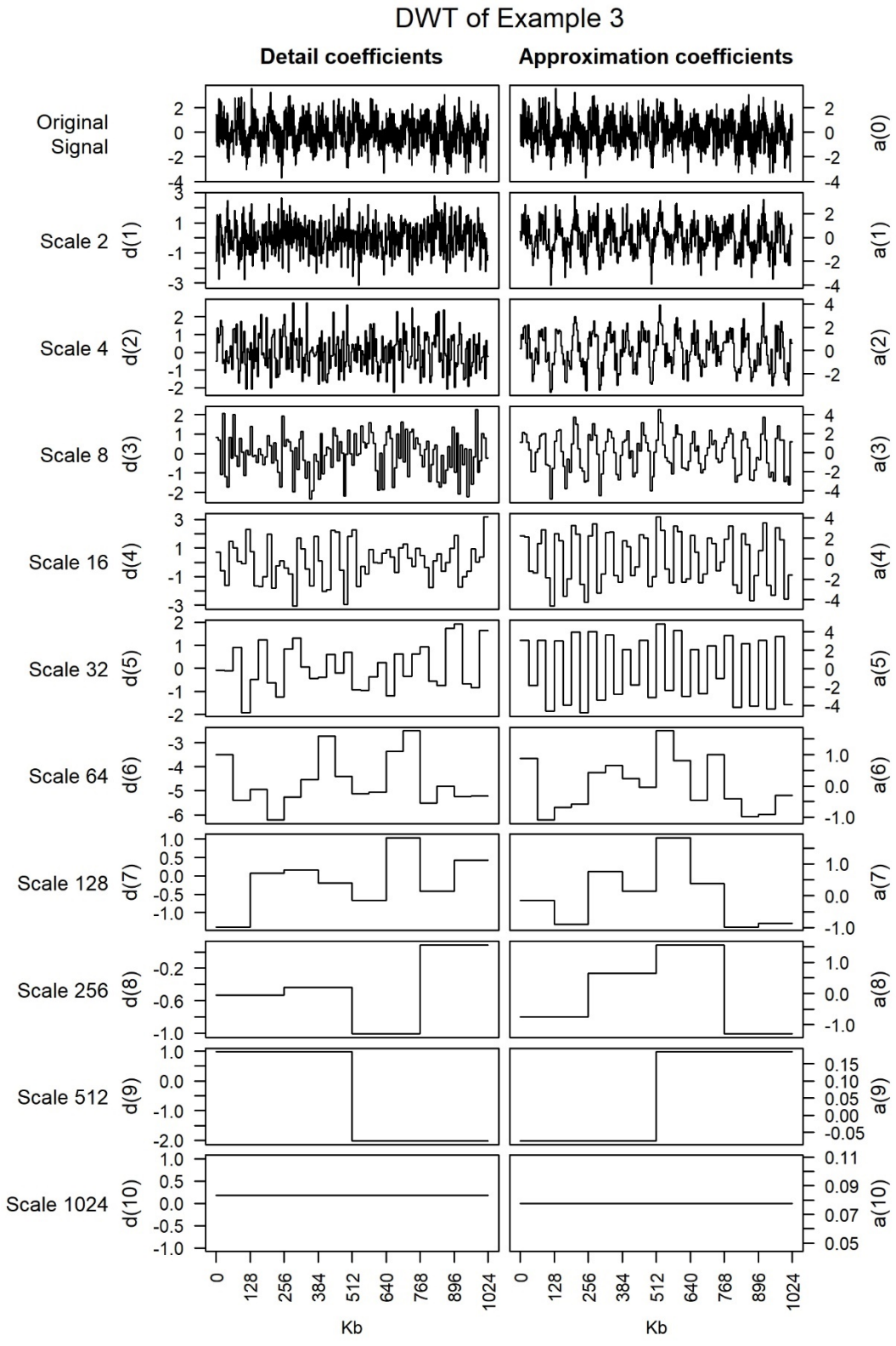
These figures show the wavelet decomposition for each example signal using the Haar DWT. The top graphs on each side show the original signal, with each graph below showing the detail and approximation coefficients for that scale. Approximation coefficients are proportional to the averages of the original signal, and detail coefficients are proportional to the changes in the averages. Note the y-axis is scaled for each graph individually so the detail can be seen. This analysis was performed in R using the wmtsa package [475].



Appendix Figure 1 Wavelet decomposition of example 1



Appendix Figure 2 Wavelet decomposition of example 2



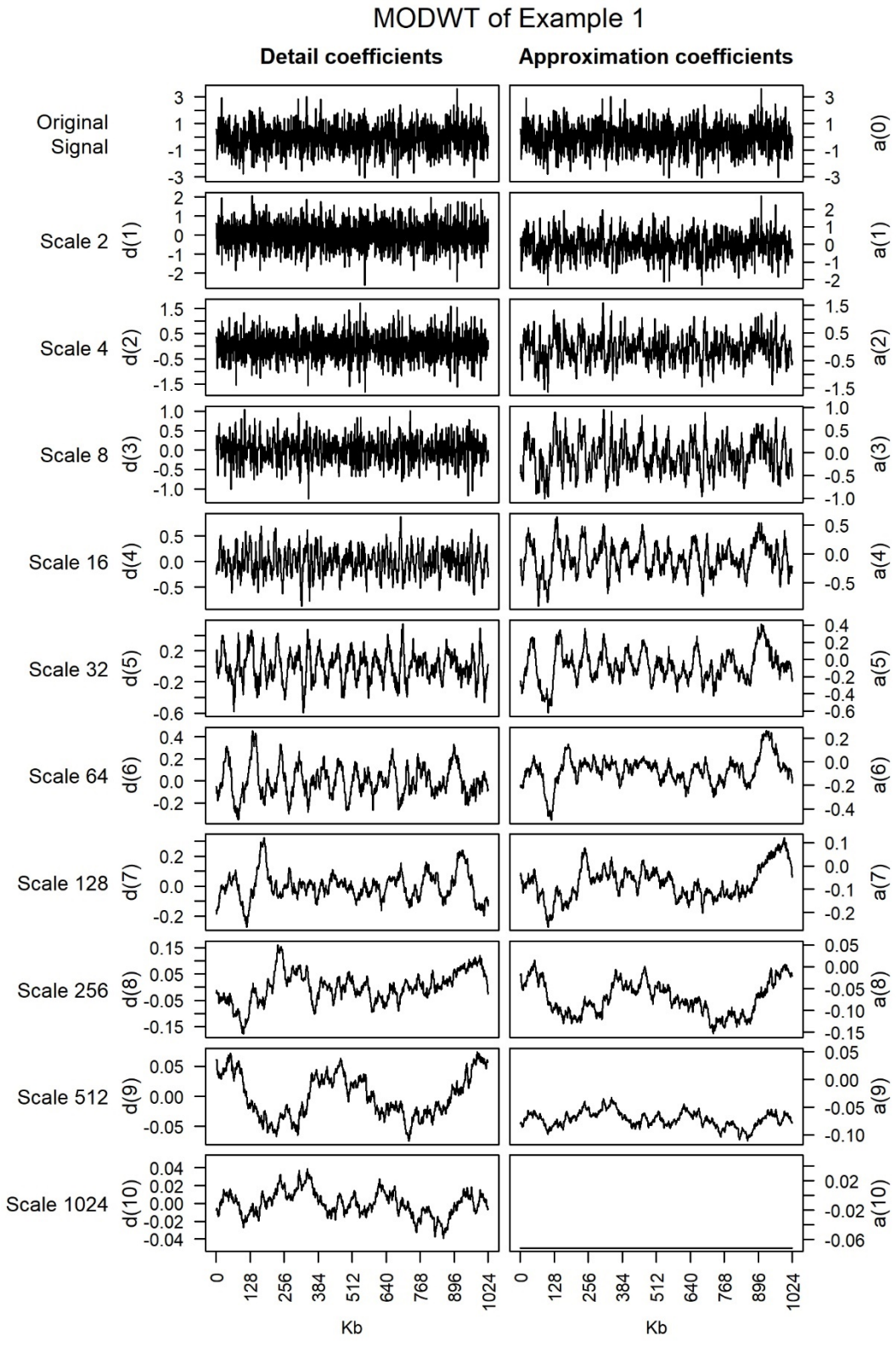
Appendix Figure 3 Wavelet decomposition of example 3

A.1.2 Maximal overlap DWT

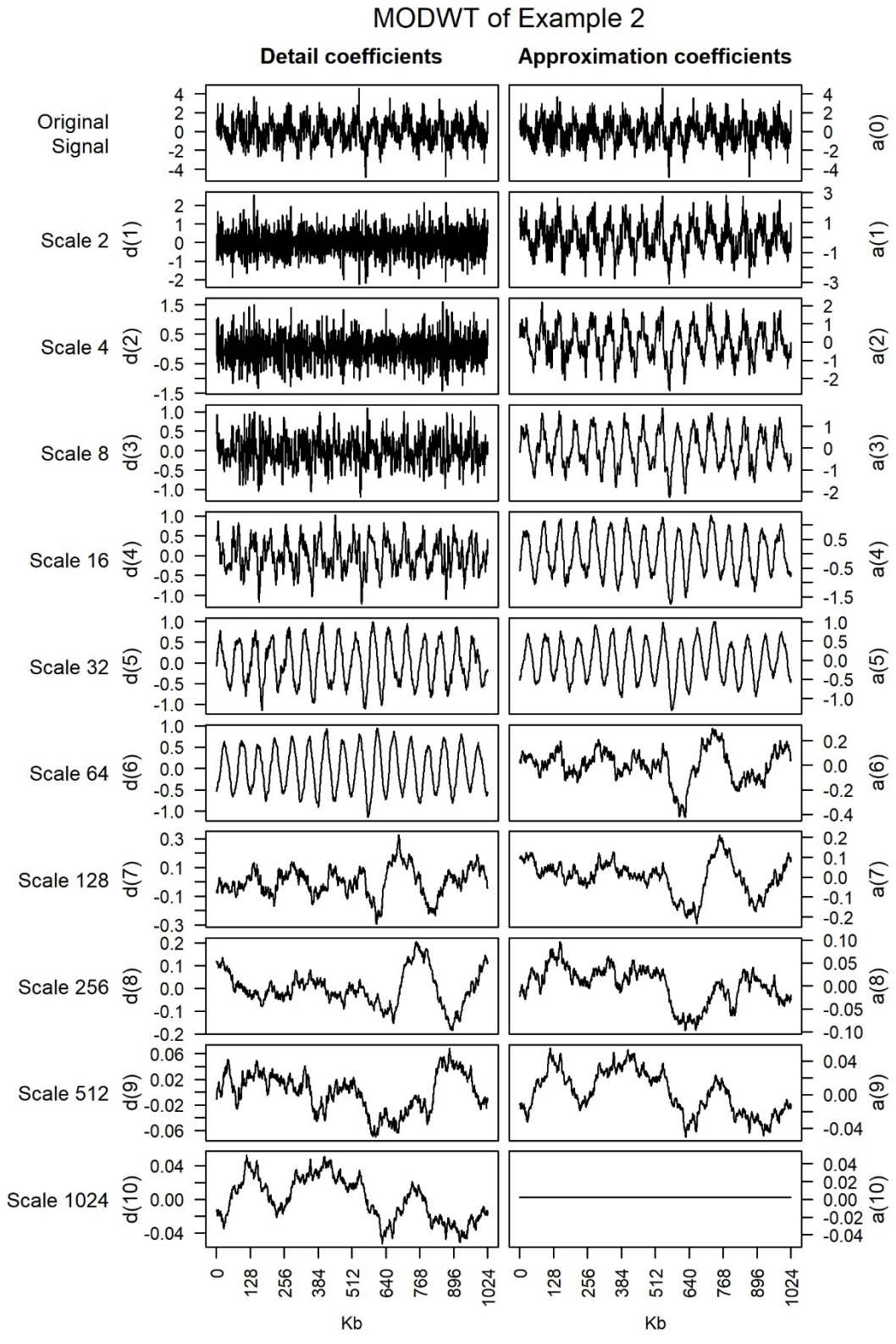
These figures show the wavelet decomposition for each example signal using the Haar MODWT.

The top graphs on each side show the original signal, with each graph below showing the detail

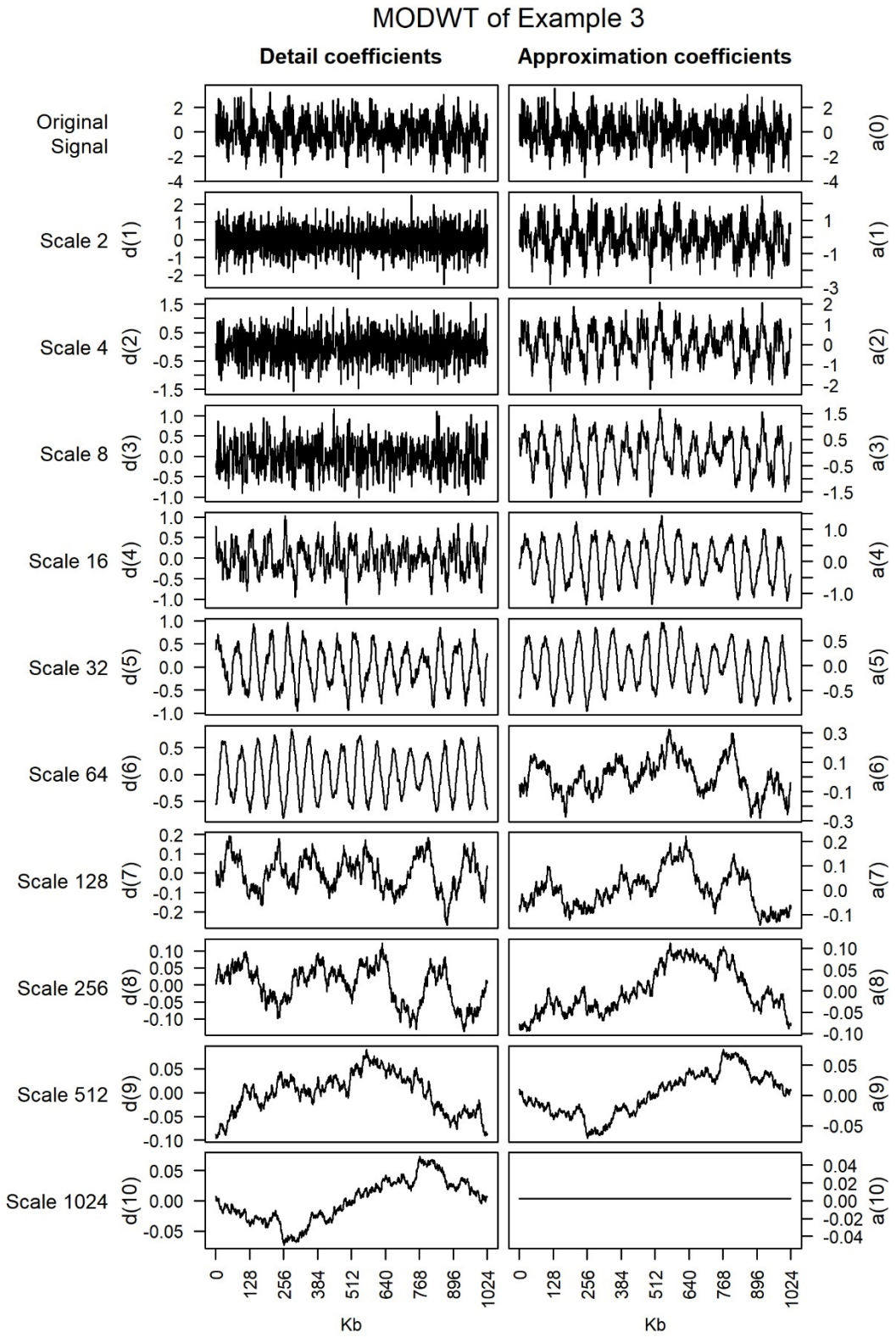
and approximation coefficients for that scale. Approximation coefficients are proportional to the averages of the original signal, and detail coefficients are proportional to the changes in the averages. Note the y-axis is scaled for each graph individually so the detail can be seen. This analysis was performed in R using the wmtsa package [475].



Appendix Figure 4 MODWT decomposition of example 1



Appendix Figure 5 MODWT decomposition of example 2

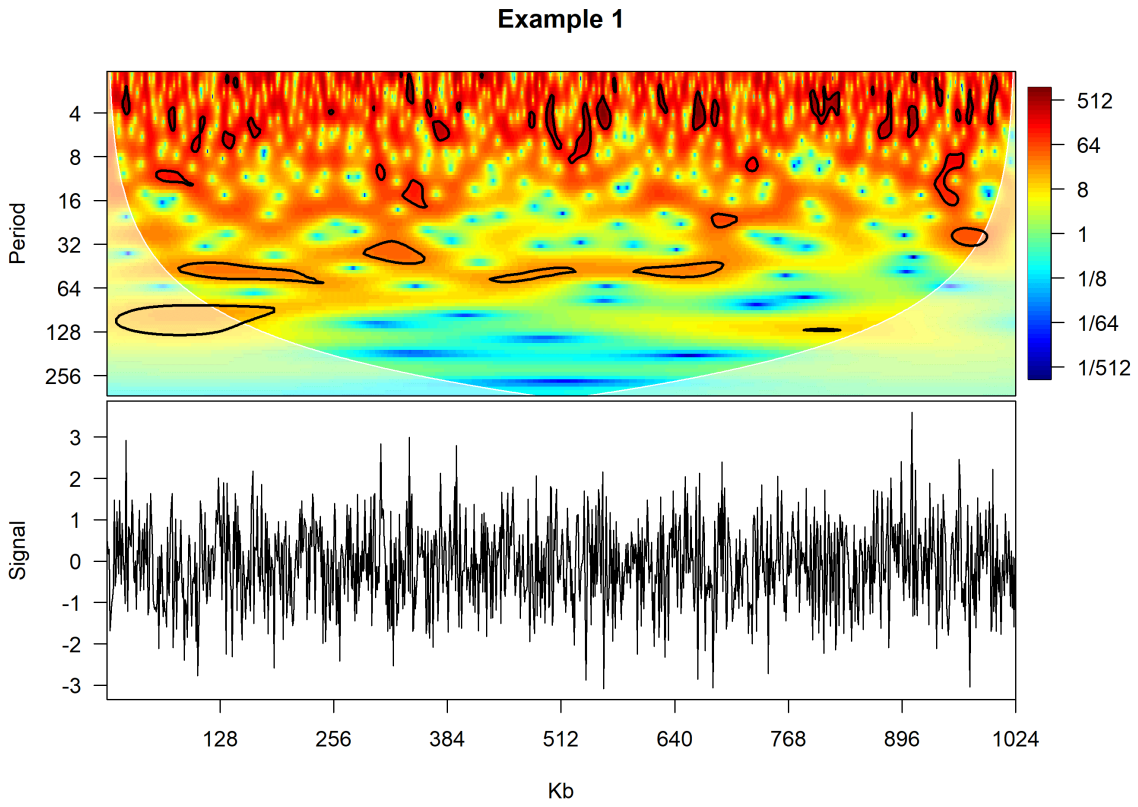


Appendix Figure 6 MODWT decomposition of example 3

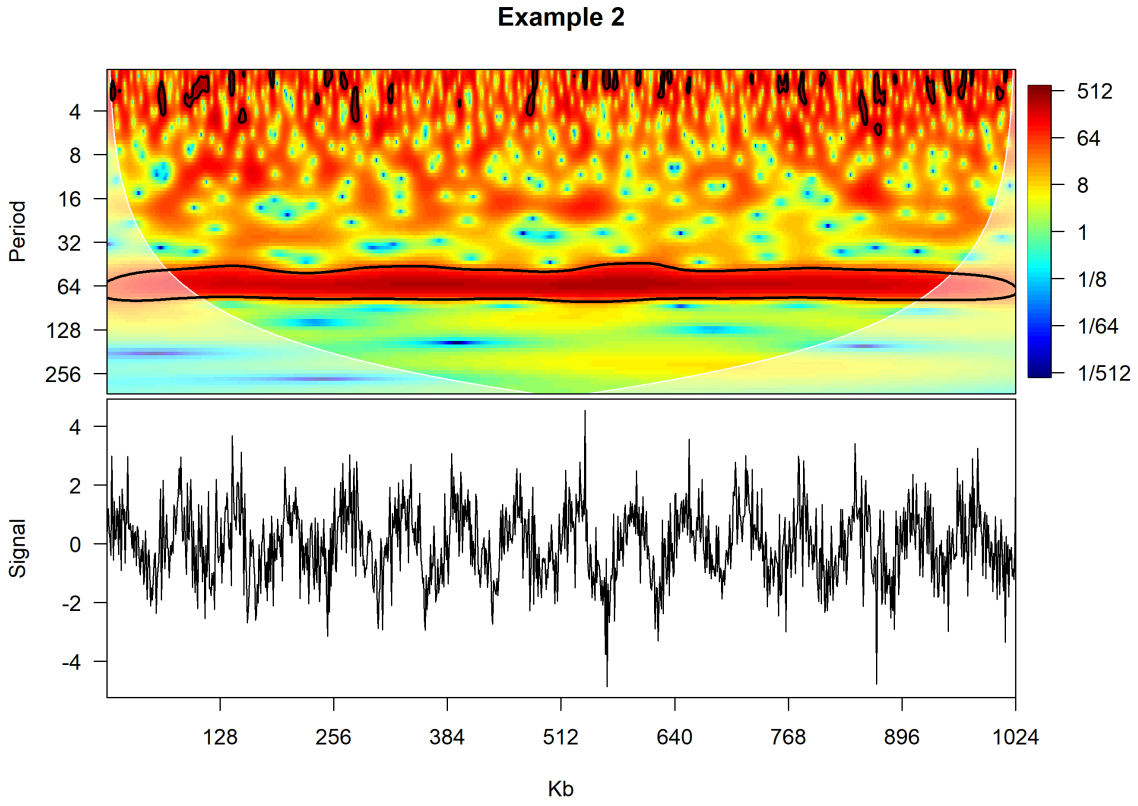
A.1.3 Continuous wavelet transform

These plots show the continuous wavelet transform for each example signal, with the original signals plotted below. The colours represent the wavelet power at that location and scale. The

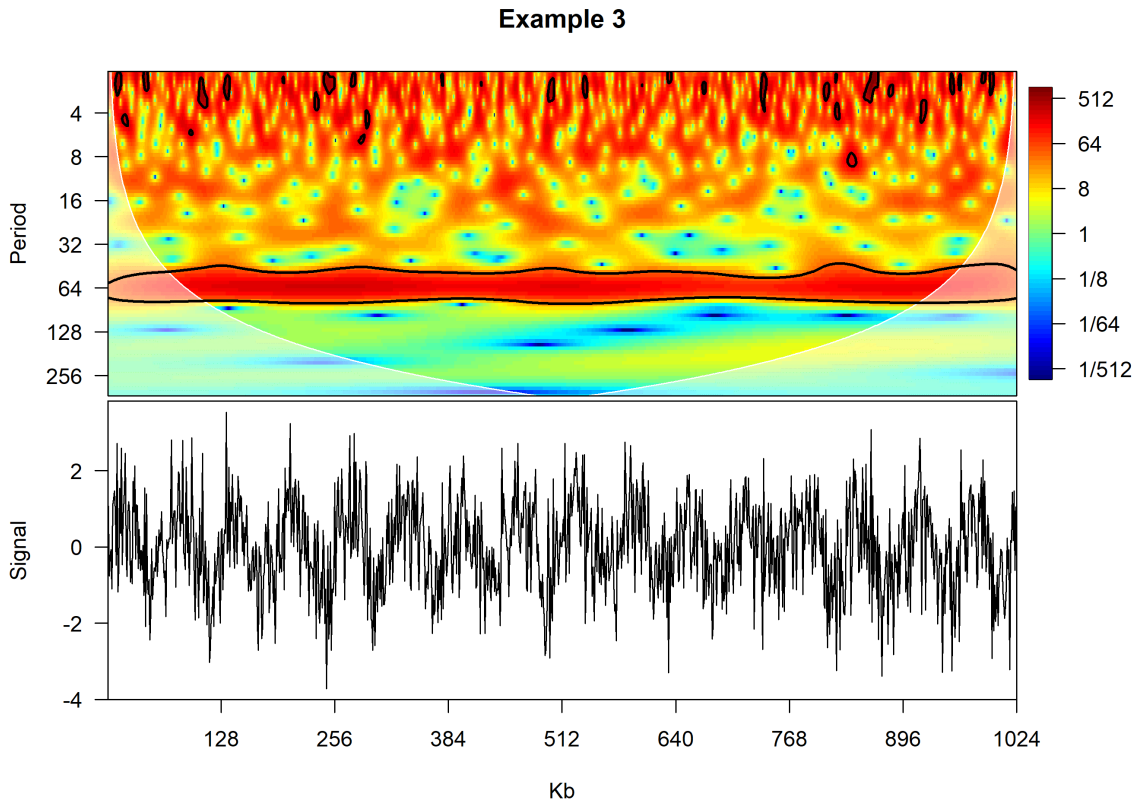
higher the power, the more energy the original signal had at that scale and location. Low power is coloured in blue (if the original signal had been completely flat, the plot would be a solid dark blue) and high power in red. Regions significant at the 95% level are surrounded by a black contour line. Regions within the cone of influence, i.e. affected by edge effects and thus should be discarded, are shown in the shaded white areas. This analysis was performed in R using the biwavelet package [481].



Appendix Figure 7 Continuous wavelet transform of example 1



Appendix Figure 8 Continuous wavelet transform of example 2



Appendix Figure 9 Continuous wavelet transform of example 3

A.2 Supplementary material for Chapter 6

The following is SLiM code for the effective bottleneck time and gene density simulations. The code below is for chromosomes with a density of 5 per Mb.

```
initialize()
{
 initializeMutationRate(1e-8);                  // overall mutation rate
 initializeMutationType("m1", 0.5, "f", 0.0);   // neutral mutation
 initializeMutationType("m2", 0.5, "f", -0.03); // deleterious mutation
 m2.convertToSubstitution = F;      // stops mutation disappearing if fixed
 initializeRecombinationRate(1e-8);           // overall recombination rate
 initializeGenomicElementType("g1", c(m1,m2), c(2,8));  // Genes
 initializeGenomicElementType("g2", m1, 1);     // Non-coding regions

 // Calculate gene lengths and non-coding region lengths
 noGenesPerMb = 5;                                        // genes per Mb
 geneInterval = asInteger(round(1000000/noGenesPerMb));  // bp between genes
```

```
 geneLength=10000;                                          // length of genes
 base = asInteger(round(geneInterval/2));             // First gene starts here

 initializeGenomicElement(g2,0,base-1);              // First non-coding region

 // Assign genes to locations along the chromosome
 geneCount=0;
 while (geneCount < noGenesPerMb*10-1) {
        initializeGenomicElement(g1,base,base+geneLength-1);
        geneCount=geneCount+1;
        initializeGenomicElement(g2,base+geneLength,base+geneInterval-1);
        base=base+geneInterval;
 }

 // last gene
 initializeGenomicElement(g1,base,base+geneLength-1);
 initializeGenomicElement(g2,base+geneLength,9999999);
}

// create a population of 10000 individuals
1 {
 sim.addSubpop("p1", 10000);
}
// run to generation 5000
5000 late() {
 p1.outputVCFSample(100, filePath="simGene5.vcf");
}
```

A.3 Supplementary material for Chapter 7

A.3.1 Uniform recombination rate ROC curve summary table

Appendix Table 1 ROC curve summary table for uniform recombination rate

A summary of the ROC curve values AUC and pAUC for each of the statistics as described in section 7.4.3 and discussed in section 7.5.1. Numbers in brackets are the 95% confidence intervals.

	Neutral vs End of sweep		Neutral vs Mid-way through sweep		Mid-way vs End of sweep	
Statistic	AUC	pAUC	AUC	pAUC	AUC	pAUC
Z_α	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.6983 (0.6251 - 0.7715)	0.146 (0.058 - 0.2841)
Z_β	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.5376 (0.4568 - 0.6184)	0.042 (0.01 - 0.092)
$\binom{\lvert L\rvert}{2} + \binom{\lvert R\rvert}{2}$	0.9999 (0.9996 - 1)	0.998 (0.988 - 1)	0.9975 (0.9943 - 1)	0.965 (0.923 - 0.998)	0.7386 (0.6677 - 0.8094)	0.1015 (0.044 - 0.197)
$\lvert L\rvert\lvert R\rvert$	0.9998 (0.9993 - 1)	0.996 (0.98 - 1)	0.9965 (0.9923 - 1)	0.957 (0.916 - 0.991)	0.7368 (0.6661 - 0.8075)	0.104 (0.0422 - 0.236)
$Z_\alpha + Z_\beta$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.5992 (0.5198 - 0.6786)	0.07 (0.02 - 0.148)
$Z_\alpha - Z_\beta$	1 (1 - 1)	1 (1 - 1)	0.9978 (0.9936 - 1)	0.956 (0.85 - 1)	0.9219 (0.8852 - 0.9586)	0.536 (0.388 - 0.706)
$Z_\alpha Z_\beta$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.5871 (0.5072 - 0.667)	0.05 (0.014 - 0.124)
$\frac{Z_\alpha}{Z_\beta}$	0.7667 (0.7013 - 0.8321)	0.216 (0.138 - 0.334)	0.8563 (0.8051 - 0.9075)	0.53 (0.39 - 0.662)	0.9512 (0.924 - 0.9784)	0.598 (0.43 - 0.82)
$Z_\alpha^{r^2/E[r^2]}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.6626 (0.5861 - 0.7391)	0.124 (0.036 - 0.254)
$Z_\alpha^{\log(r^2/E[r^2])}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.7707 (0.7054 - 0.836)	0.232 (0.114 - 0.3821)
Z_α^{ZScore}	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.6683 (0.5923 - 0.7443)	0.132 (0.04 - 0.2561)

Appendix A

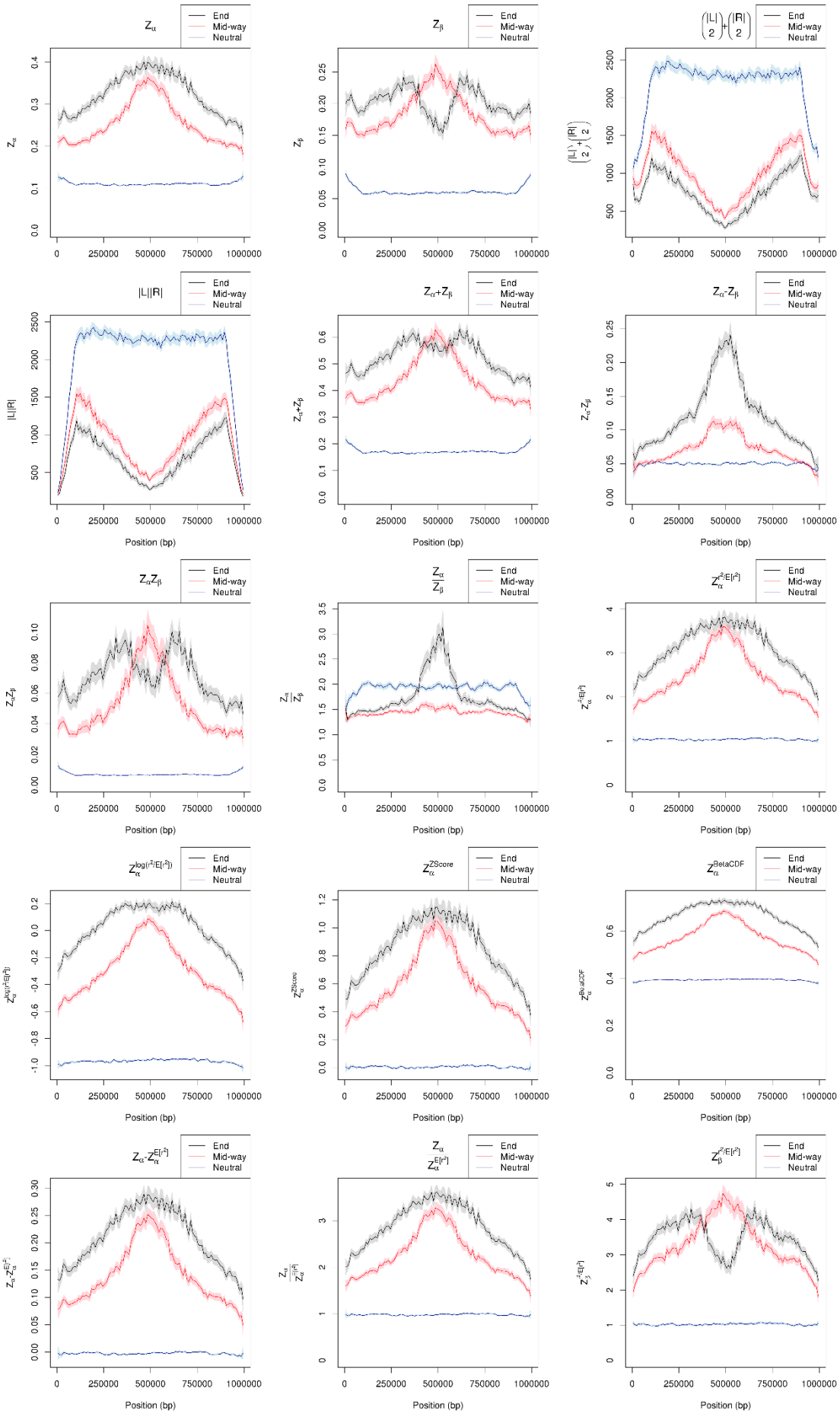
	Neutral vs End of sweep		Neutral vs Mid-way through sweep		Mid-way vs End of sweep	
Statistic	AUC	pAUC	AUC	pAUC	AUC	pAUC
$Z_{\alpha}^{BetaCDF}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.7769 (0.7127 - 0.8411)	0.234 (0.11 - 0.4021)
$Z_{\alpha} - Z_{\alpha}^{E[r^2]}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.6965 (0.623 - 0.77)	0.146 (0.058 - 0.2861)
$\frac{Z_{\alpha}}{Z_{\alpha}^{E[r^2]}}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.6873 (0.6127 - 0.7619)	0.136 (0.05 - 0.274)
$Z_{\beta}^{r^2/E[r^2]}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.4732 (0.3923 - 0.5541)	0.038 (0.006 - 0.088)
$Z_{\beta}^{\log(r^2/E[r^2])}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.6866 (0.6117 - 0.7615)	0.102 (0.026 - 0.2621)
Z_{β}^{ZScore}	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.4797 (0.3986 - 0.5608)	0.042 (0.01 - 0.092)
$Z_{\beta}^{BetaCDF}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.739 (0.6698 - 0.8082)	0.126 (0.03 - 0.344)
$Z_{\beta} - Z_{\beta}^{E[r^2]}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.5364 (0.4555 - 0.6173)	0.044 (0.008 - 0.098)
$\frac{Z_{\beta}}{Z_{\beta}^{E[r^2]}}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.5073 (0.4261 - 0.5885)	0.048 (0.012 - 0.1)
$Z_{\alpha}^{r^2/E[r^2]} + Z_{\beta}^{r^2/E[r^2]}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.5297 (0.4485 - 0.6109)	0.05 (0.012 - 0.112)
$Z_{\alpha}^{\log(r^2/E[r^2])} + Z_{\beta}^{\log(r^2/E[r^2])}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.713 (0.6404 - 0.7856)	0.126 (0.034 - 0.282)
$Z_{\alpha}^{ZScore} + Z_{\beta}^{ZScore}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.5362 (0.455 - 0.6174)	0.05 (0.012 - 0.112)
$Z_{\alpha}^{BetaCDF} + Z_{\beta}^{BetaCDF}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.7431 (0.6742 - 0.812)	0.142 (0.046 - 0.322)
$(Z_{\alpha} - Z_{\alpha}^{E[r^2]}) + (Z_{\beta} - Z_{\beta}^{E[r^2]})$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.5973 (0.5177 - 0.6769)	0.066 (0.02 - 0.15)
$\frac{Z_{\alpha}}{Z_{\alpha}^{E[r^2]}} + \frac{Z_{\beta}}{Z_{\beta}^{E[r^2]}}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.5552 (0.4743 -	0.052 (0.016 -

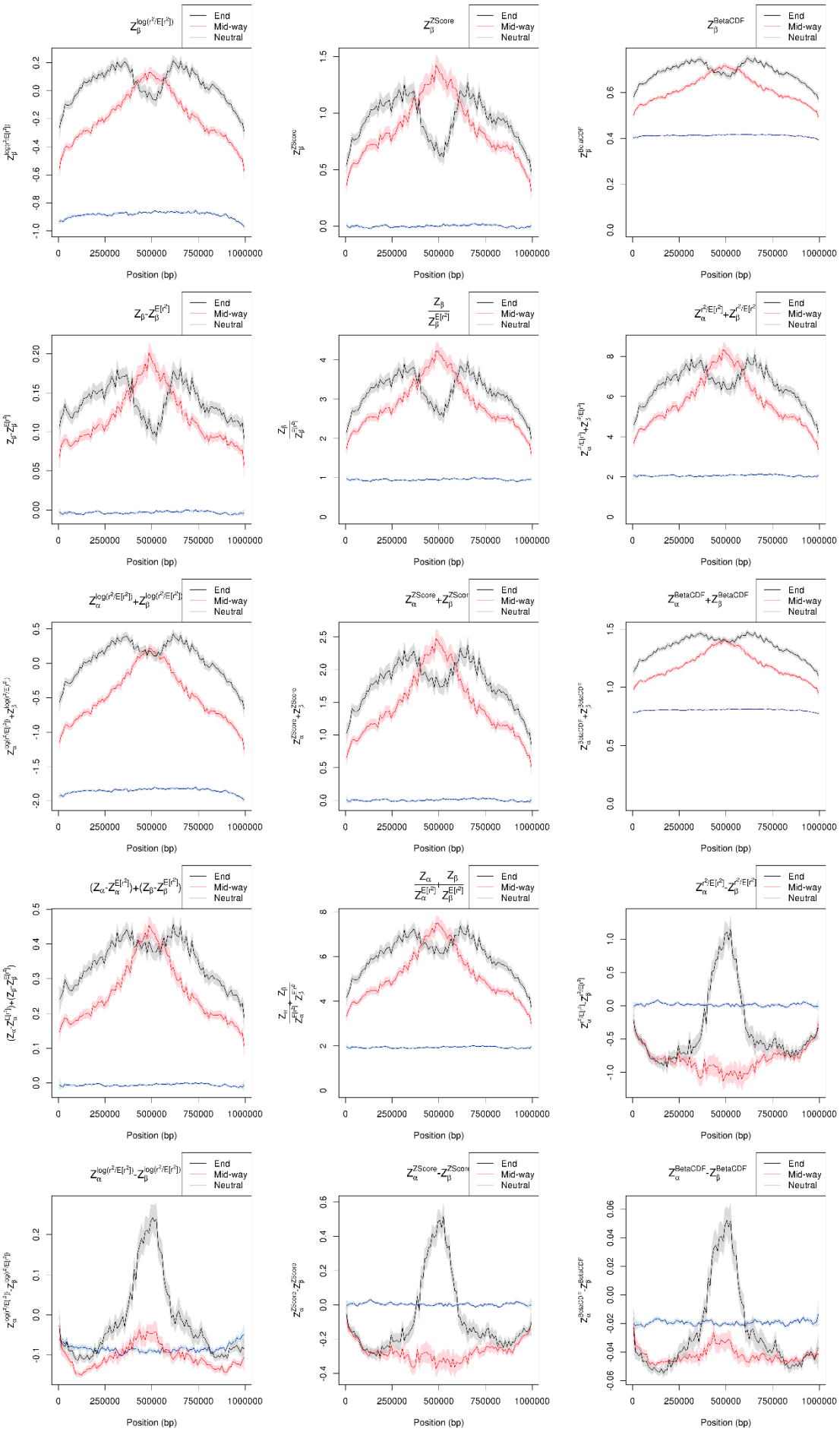
	Neutral vs End of sweep		Neutral vs Mid-way through sweep		Mid-way vs End of sweep	
Statistic	AUC	pAUC	AUC	pAUC	AUC	pAUC
					0.6361)	0.114)
$Z_{\alpha}^{r^2/E[r^2]} - Z_{\beta}^{r^2/E[r^2]}$	0.9956 (0.9875 - 1)	0.984 (0.956 - 1)	0.5668 (0.4817 - 0.6519)	0.282 (0.184 - 0.39)	0.9585 (0.9342 - 0.9828)	0.684 (0.494 - 0.858)
$Z_{\alpha}^{\log(r^2/E[r^2])} - Z_{\beta}^{\log(r^2/E[r^2])}$	0.9927 (0.9859 - 0.9995)	0.916 (0.846 - 0.976)	0.775 (0.7108 - 0.8392)	0.29 (0.2 - 0.416)	0.8831 (0.8353 - 0.9309)	0.522 (0.382 - 0.694)
$Z_{\alpha}^{ZScore} - Z_{\beta}^{ZScore}$	0.9971 (0.9919 - 1)	0.984 (0.956 - 1)	0.6598 (0.5801 - 0.7395)	0.33 (0.228 - 0.444)	0.9569 (0.932 - 0.9818)	0.704 (0.5239 - 0.862)
$Z_{\alpha}^{BetaCDF} - Z_{\beta}^{BetaCDF}$	0.9862 (0.9725 - 0.9999)	0.926 (0.87 - 0.978)	0.6224 (0.5413 - 0.7035)	0.28 (0.188 - 0.378)	0.9053 (0.8635 - 0.9471)	0.518 (0.35 - 0.712)
$(Z_{\alpha} - Z_{\alpha}^{E[r^2]}) - (Z_{\beta} - Z_{\beta}^{E[r^2]})$	1 (1 - 1)	1 (1 - 1)	0.9971 (0.9921 - 1)	0.942 (0.816 - 1)	0.9234 (0.8873 - 0.9595)	0.538 (0.404 - 0.706)
$\frac{Z_{\alpha}}{Z_{\alpha}^{E[r^2]}} - \frac{Z_{\beta}}{Z_{\beta}^{E[r^2]}}$	0.9967 (0.9912 - 1)	0.978 (0.94 - 1)	0.5935 (0.5094 - 0.6776)	0.274 (0.182 - 0.384)	0.9535 (0.9279 - 0.9791)	0.694 (0.534 - 0.84)
$Z_{\alpha}^{BetaCDF} Z_{\beta}^{BetaCDF}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.7423 (0.6734 - 0.8112)	0.14 (0.042 - 0.3201)
$\frac{Z_{\alpha}^{BetaCDF}}{Z_{\beta}^{BetaCDF}}$	0.9509 (0.9225 - 0.9793)	0.704 (0.586 - 0.8261)	0.5315 (0.4498 - 0.6132)	0.128 (0.062 - 0.21)	0.8904 (0.8448 - 0.936)	0.474 (0.312 - 0.686)
$\frac{Z_{\alpha}}{Z_{\beta}} - \frac{Z_{\alpha}^{E[r^2]}}{Z_{\beta}^{E[r^2]}}$	0.7711 (0.7059 - 0.8363)	0.226 (0.14 - 0.336)	0.863 (0.8133 - 0.9127)	0.558 (0.454 - 0.67)	0.954 (0.9282 - 0.9798)	0.634 (0.52 - 0.786)

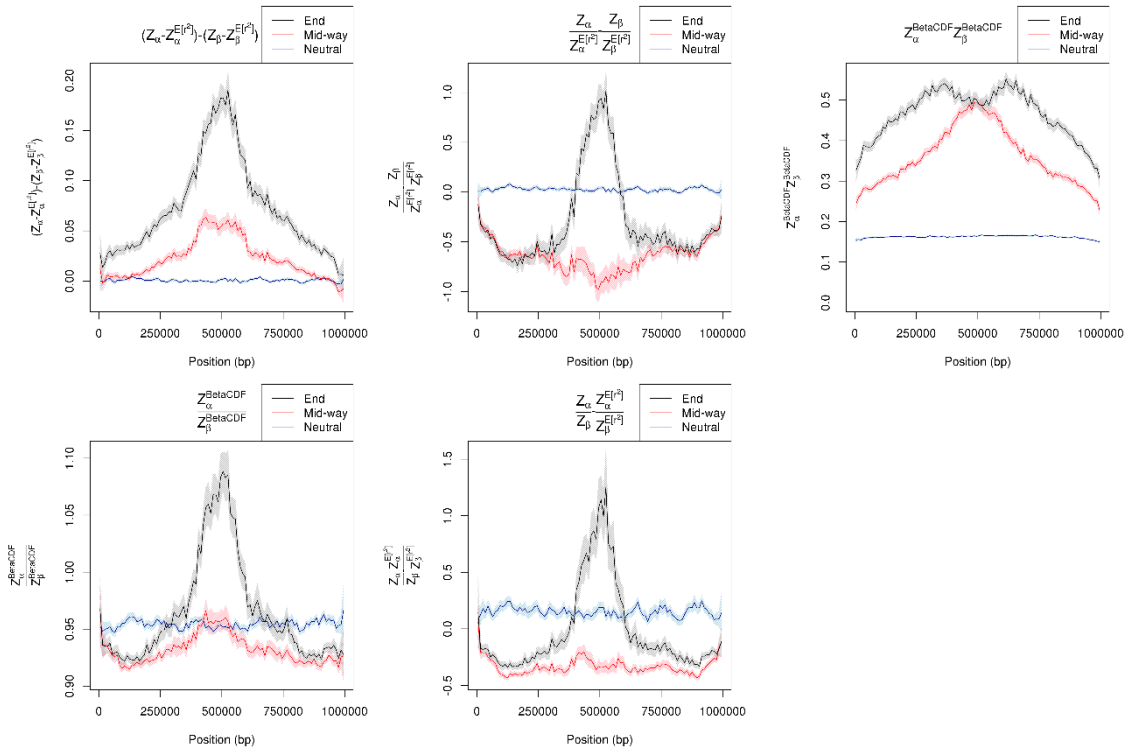
A.3.2 Uniform recombination rate aggregate graphs

These figures show the graphs for each of the statistics as described in section 7.4.2 and discussed in section 7.5.1.

Appendix A


Z_α
Z_β
$\binom{|L|}{2}+\binom{|R|}{2}$
$|L||R|$
$Z_\alpha+Z_\beta$
$Z_\alpha-Z_\beta$
$Z_\alpha Z_\beta$
$\frac{Z_\alpha}{Z_\beta}$
$Z_\alpha^{r^2/E[r^2]}$
$Z_\alpha^{\log(r^2/E[r^2])}$
Z_α^{ZScore}
$Z_\alpha^{BetaCDF}$
$Z_\alpha-Z_\alpha^{E[r^2]}$
$\frac{Z_\alpha}{Z_\alpha^{E[r^2]}}$
$Z_\beta^{r^2/E[r^2]}$
End
Mid-way
Neutral
Position (bp)


End
Mid-way
Neutral
Position (bp)
0
250000
500000
750000
1000000



A.3.3 Variable recombination rate ROC curve summary table

Appendix Table 2 ROC curve summary table for variable recombination rate

A summary of the ROC curve values AUC and pAUC for each of the statistics as described in section 7.4.3 and discussed in section 7.5.2. Numbers in brackets are the 95% confidence intervals.

	Neutral vs End of sweep		Neutral vs Mid-way through sweep		Mid-way vs End of sweep	
Statistic	AUC	pAUC	AUC	pAUC	AUC	pAUC
Z_α	0.9866 (0.9747 - 0.9985)	0.862 (0.784 - 0.954)	0.9691 (0.9489 - 0.9893)	0.712 (0.606 - 0.854)	0.6346 (0.5572 - 0.712)	0.062 (0.014 - 0.198)
Z_β	0.9737 (0.9558 - 0.9916)	0.71 (0.526 - 0.894)	0.9215 (0.8847 - 0.9583)	0.466 (0.308 - 0.7221)	0.6785 (0.6047 - 0.7523)	0.11 (0.05 - 0.21)
$\binom{\lvert L\rvert}{2}+\binom{\lvert R\rvert}{2}$	0.9703 (0.9484 - 0.9921)	0.687 (0.394 - 0.92)	0.9191 (0.8772 - 0.9609)	0.444 (0.141 - 0.7501)	0.7016 (0.6294 - 0.7738)	0.08 (0.032 - 0.17)
$\lvert L\rvert\lvert R\rvert$	0.977 (0.9586 - 0.9953)	0.722 (0.458 - 0.9361)	0.9268 (0.8877 - 0.9658)	0.485 (0.203 - 0.763)	0.7127 (0.6416 - 0.7837)	0.094 (0.038 - 0.2121)
$Z_\alpha + Z_\beta$	0.9807 (0.9653 - 0.9961)	0.774 (0.572 - 0.944)	0.9474 (0.9171 - 0.9777)	0.576 (0.388 - 0.8041)	0.6577 (0.5824 - 0.733)	0.082 (0.026 - 0.228)

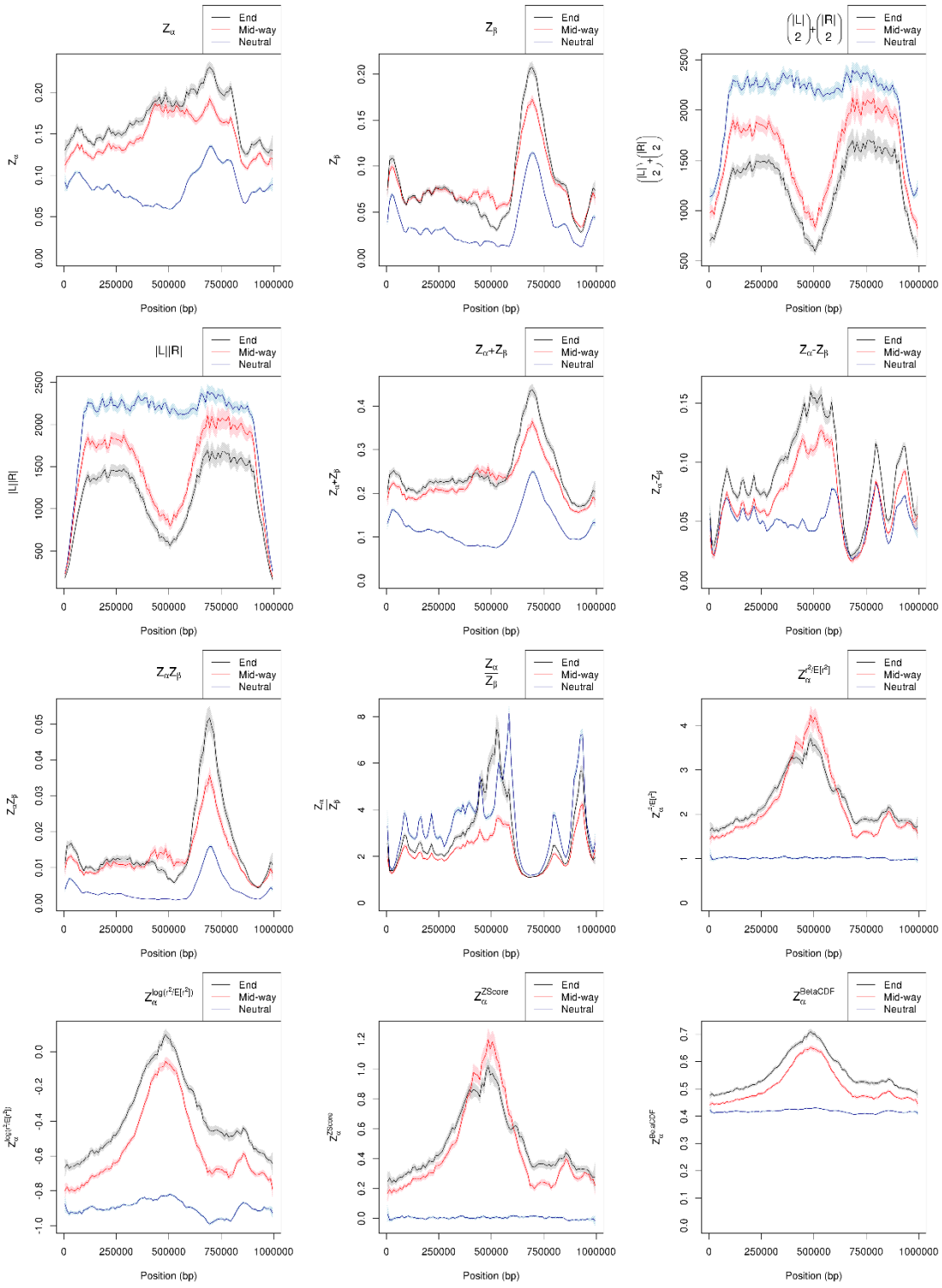
	Neutral vs End of sweep		Neutral vs Mid-way through sweep		Mid-way vs End of sweep	
Statistic	AUC	pAUC	AUC	pAUC	AUC	pAUC
$Z_\alpha - Z_\beta$	0.9925 (0.9851 - 0.9999)	0.882 (0.776 - 0.976)	0.9535 (0.9278 - 0.9792)	0.592 (0.43 - 0.7781)	0.7163 (0.6453 - 0.7873)	0.2 (0.098 - 0.3361)
$Z_\alpha Z_\beta$	0.9789 (0.9628 - 0.995)	0.762 (0.558 - 0.942)	0.9413 (0.9093 - 0.9733)	0.548 (0.33 - 0.7801)	0.6709 (0.5965 - 0.7453)	0.086 (0.032 - 0.224)
$\frac{Z_\alpha}{Z_\beta}$	0.5922 (0.5124 - 0.672)	0.144 (0.064 - 0.2381)	0.9517 (0.9243 - 0.9791)	0.714 (0.598 - 0.85)	0.9466 (0.9186 - 0.9746)	0.582 (0.378 - 0.768)
$Z_\alpha^{r^2/E[r^2]}$	0.9998 (0.9993 - 1)	0.996 (0.982 - 1)	0.9996 (0.9988 - 1)	0.992 (0.968 - 1)	0.6086 (0.5303 - 0.6869)	0.18 (0.096 - 0.3)
$Z_\alpha^{\log(r^2/E[r^2])}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.6864 (0.6126 - 0.7602)	0.182 (0.104 - 0.29)
Z_α^{ZScore}	0.9999 (0.9996 - 1)	0.998 (0.988 - 1)	0.9995 (0.9985 - 1)	0.99 (0.964 - 1)	0.5934 (0.5143 - 0.6725)	0.172 (0.09 - 0.284)
$Z_\alpha^{BetaCDF}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.7221 (0.6521 - 0.7921)	0.196 (0.11 - 0.3041)
$Z_\alpha - Z_\alpha^{E[r^2]}$	0.9981 (0.9956 - 1)	0.962 (0.918 - 0.996)	0.9949 (0.9898 - 1)	0.91 (0.852 - 0.9801)	0.6068 (0.5281 - 0.6855)	0.086 (0.04 - 0.18)
$\frac{Z_\alpha}{Z_\alpha^{E[r^2]}}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.5549 (0.4744 - 0.6354)	0.086 (0.04 - 0.166)
$Z_\beta^{r^2/E[r^2]}$	0.9983 (0.9958 - 1)	0.966 (0.904 - 1)	0.9996 (0.9988 - 1)	0.992 (0.968 - 1)	0.7563 (0.6905 - 0.8221)	0.242 (0.138 - 0.38)
$Z_\beta^{\log(r^2/E[r^2])}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.4454 (0.3652 - 0.5256)	0.018 (0 - 0.058)
Z_β^{ZScore}	0.9976 (0.9945 - 1)	0.952 (0.89 - 0.998)	0.9995 (0.9986 - 1)	0.99 (0.968 - 1)	0.7757 (0.7123 - 0.8391)	0.286 (0.182 - 0.408)
$Z_\beta^{BetaCDF}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.5158 (0.4348 - 0.5968)	0.032 (0.004 - 0.074)
$Z_\beta - Z_\beta^{E[r^2]}$	0.9886 (0.9772 - 1)	0.804 (0.614 - 0.97)	0.9853 (0.9704 - 1)	0.726 (0.496 - 0.968)	0.5984 (0.5198 - 0.677)	0.02 (0 - 0.154)
$\frac{Z_\beta}{Z_\beta^{E[r^2]}}$	0.9987 (0.9964 - 1)	0.974 (0.914 - 1)	0.9995 (0.9986 - 1)	0.99 (0.966 - 1)	0.7112 (0.6405 -	0.212 (0.096 -

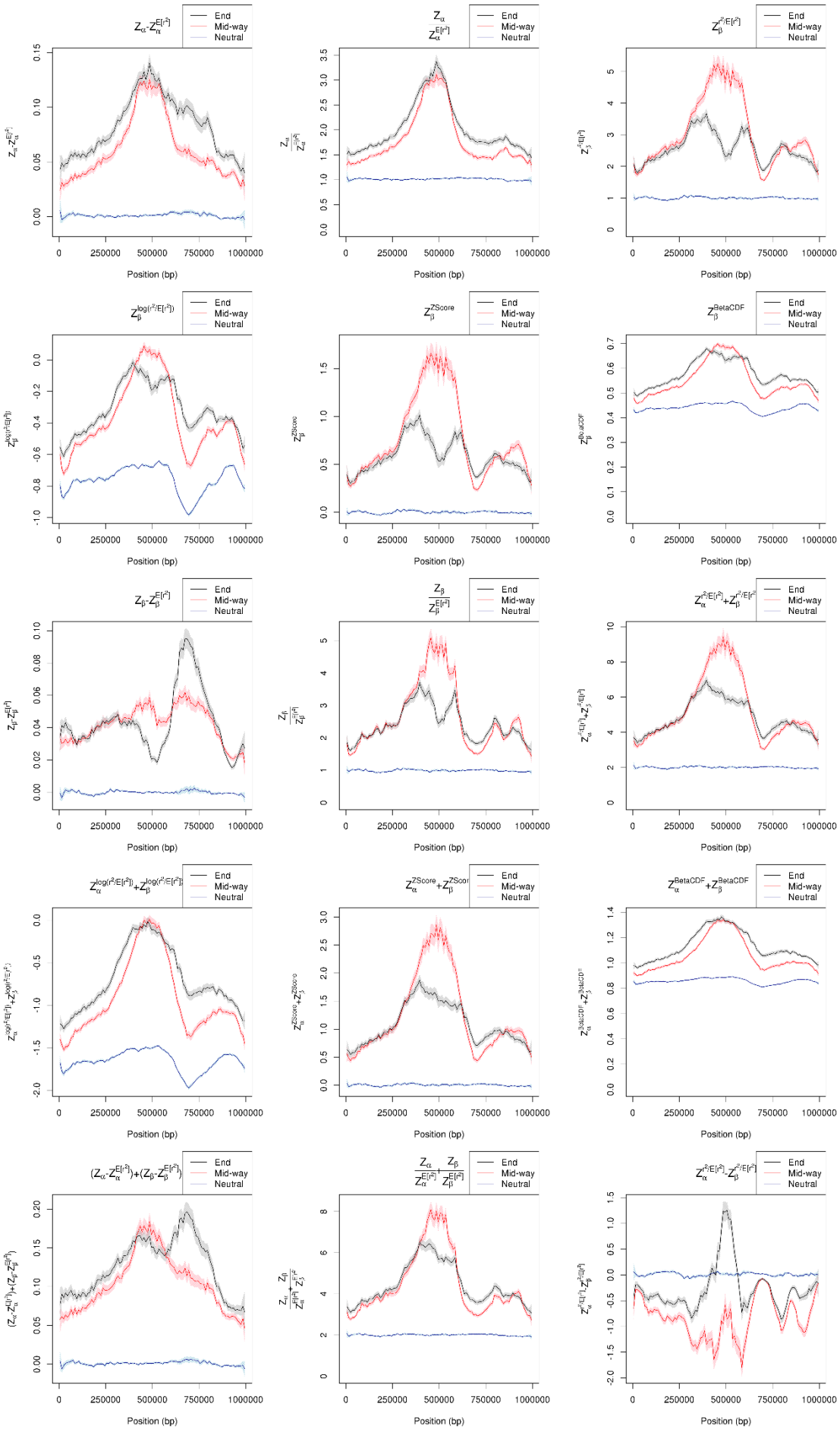
Appendix A

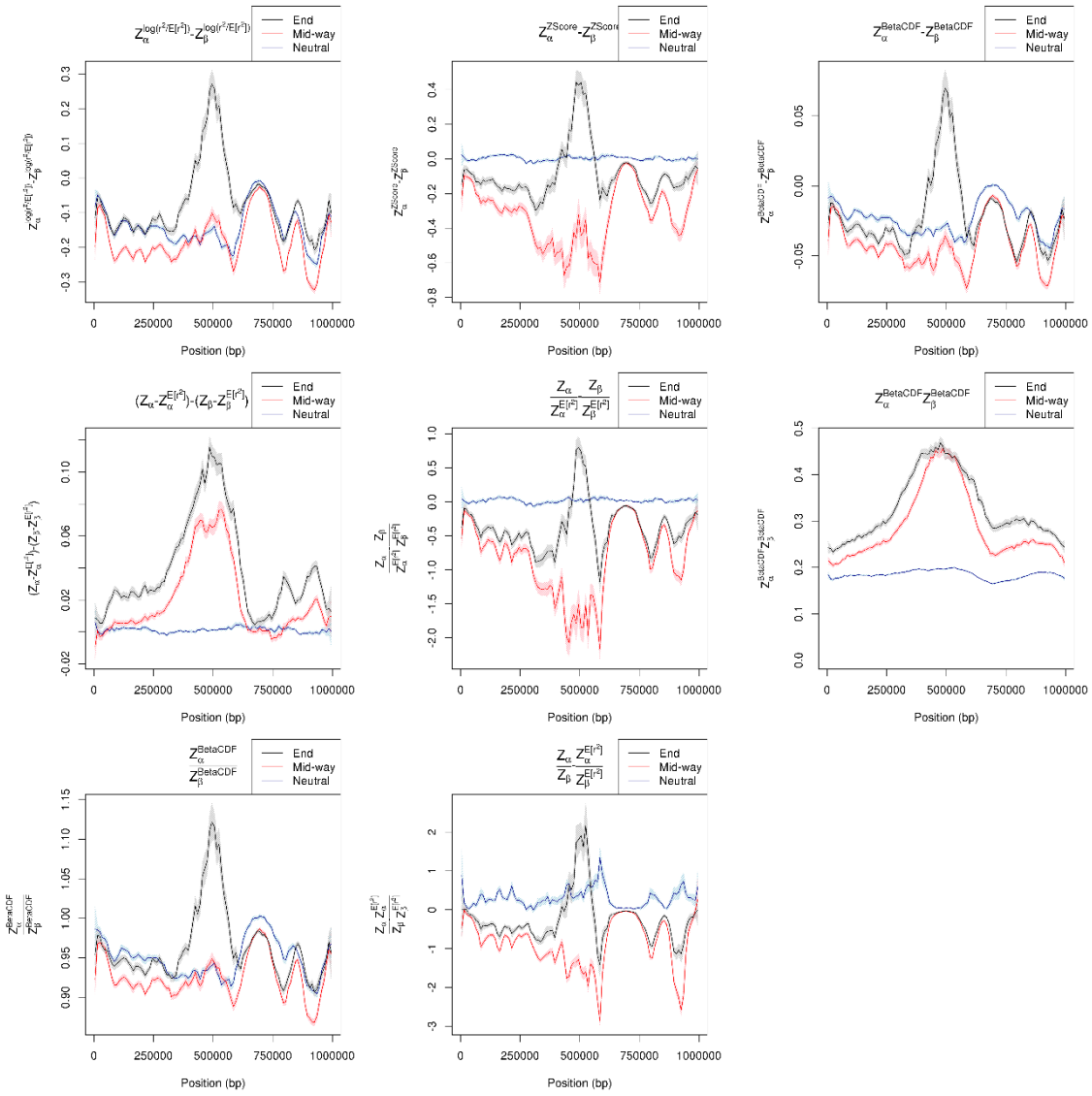
	Neutral vs End of sweep		Neutral vs Mid-way through sweep		Mid-way vs End of sweep	
Statistic	AUC	pAUC	AUC	pAUC	AUC	pAUC
					0.7819)	0.346)
$Z_{\alpha}^{r^2/E[r^2]} + Z_{\beta}^{r^2/E[r^2]}$	0.9995 (0.9984 - 1)	0.99 (0.96 - 1)	0.9998 (0.9993 - 1)	0.996 (0.982 - 1)	0.7326 (0.6641 - 0.8011)	0.252 (0.146 - 0.38)
$Z_{\alpha}^{\log(r^2/E[r^2])} + Z_{\beta}^{\log(r^2/E[r^2])}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.521 (0.4399 - 0.6021)	0.062 (0.02 - 0.126)
$Z_{\alpha}^{ZScore} + Z_{\beta}^{ZScore}$	0.9995 (0.9985 - 1)	0.99 (0.964 - 1)	0.9998 (0.9993 - 1)	0.996 (0.98 - 1)	0.743 (0.6756 - 0.8104)	0.27 (0.166 - 0.402)
$Z_{\alpha}^{BetaCDF} + Z_{\beta}^{BetaCDF}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.574 (0.4938 - 0.6542)	0.056 (0.018 - 0.134)
$(Z_{\alpha} - Z_{\alpha}^{E[r^2]}) + (Z_{\beta} - Z_{\beta}^{E[r^2]})$	0.995 (0.9892 - 1)	0.906 (0.798 - 0.992)	0.9921 (0.9824 - 1)	0.846 (0.662 - 0.994)	0.5808 (0.5013 - 0.6603)	0.034 (0.002 - 0.112)
$\frac{Z_{\alpha}}{Z_{\alpha}^{E[r^2]}} + \frac{Z_{\beta}}{Z_{\beta}^{E[r^2]}}$	0.9996 (0.9987 - 1)	0.992 (0.968 - 1)	0.9997 (0.999 - 1)	0.994 (0.976 - 1)	0.6615 (0.5865 - 0.7365)	0.198 (0.106 - 0.306)
$Z_{\alpha}^{r^2/E[r^2]} - Z_{\beta}^{r^2/E[r^2]}$	0.9532 (0.9216 - 0.9848)	0.81 (0.718 - 0.902)	0.585 (0.4981 - 0.6719)	0.308 (0.212 - 0.4261)	0.8406 (0.7853 - 0.8959)	0.318 (0.142 - 0.51)
$Z_{\alpha}^{\log(r^2/E[r^2])} - Z_{\beta}^{\log(r^2/E[r^2])}$	0.9629 (0.9412 - 0.9846)	0.71 (0.5579 - 0.854)	0.5451 (0.4645 - 0.6257)	0.026 (0 - 0.146)	0.9455 (0.9181 - 0.9729)	0.58 (0.474 - 0.708)
$Z_{\alpha}^{ZScore} - Z_{\beta}^{ZScore}$	0.9418 (0.9058 - 0.9778)	0.794 (0.698 - 0.884)	0.5547 (0.469 - 0.6404)	0.258 (0.168 - 0.362)	0.8522 (0.7987 - 0.9057)	0.29 (0.11 - 0.508)
$Z_{\alpha}^{BetaCDF} - Z_{\beta}^{BetaCDF}$	0.9216 (0.8807 - 0.9625)	0.772 (0.678 - 0.866)	0.5905 (0.5104 - 0.6706)	0.174 (0.1 - 0.278)	0.9279 (0.8924 - 0.9634)	0.368 (0.21 - 0.596)
$(Z_{\alpha} - Z_{\alpha}^{E[r^2]}) - (Z_{\beta} - Z_{\beta}^{E[r^2]})$	0.9948 (0.9883 - 1)	0.898 (0.776 - 0.992)	0.9723 (0.9534 - 0.9912)	0.68 (0.48 - 0.878)	0.7089 (0.6374 - 0.7804)	0.16 (0.088 - 0.31)
$\frac{Z_{\alpha}}{Z_{\alpha}^{E[r^2]}} - \frac{Z_{\beta}}{Z_{\beta}^{E[r^2]}}$	0.9073 (0.858 - 0.9566)	0.776 (0.676 - 0.874)	0.7116 (0.6332 - 0.79)	0.452 (0.342 - 0.58)	0.9155 (0.8764 - 0.9546)	0.39 (0.222 - 0.6541)
$Z_{\alpha}^{BetaCDF} Z_{\beta}^{BetaCDF}$	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	0.5719 (0.4917 - 0.6521)	0.054 (0.012 - 0.128)
$\frac{Z_{\alpha}^{BetaCDF}}{Z_{\beta}^{BetaCDF}}$	0.876 (0.8252 - 0.9268)	0.632 (0.534 - 0.7401)	0.6749 (0.6007 - 0.7491)	0.192 (0.112 - 0.3)	0.9224 (0.8853 - 0.9595)	0.342 (0.186 - 0.562)

	Neutral vs End of sweep		Neutral vs Mid-way through sweep		Mid-way vs End of sweep	
Statistic	AUC	pAUC	AUC	pAUC	AUC	pAUC
$\frac{Z_\alpha}{Z_\beta} - \frac{Z_\alpha^{E[r^2]}}{Z_\beta^{E[r^2]}}$	0.5989 (0.5192 - 0.6786)	0.118 (0.054 - 0.224)	0.9689 (0.9479 - 0.9899)	0.798 (0.7 - 0.894)	0.9393 (0.907 - 0.9716)	0.414 (0.256 - 0.632)

A.3.4 Variable recombination rate aggregate graphs

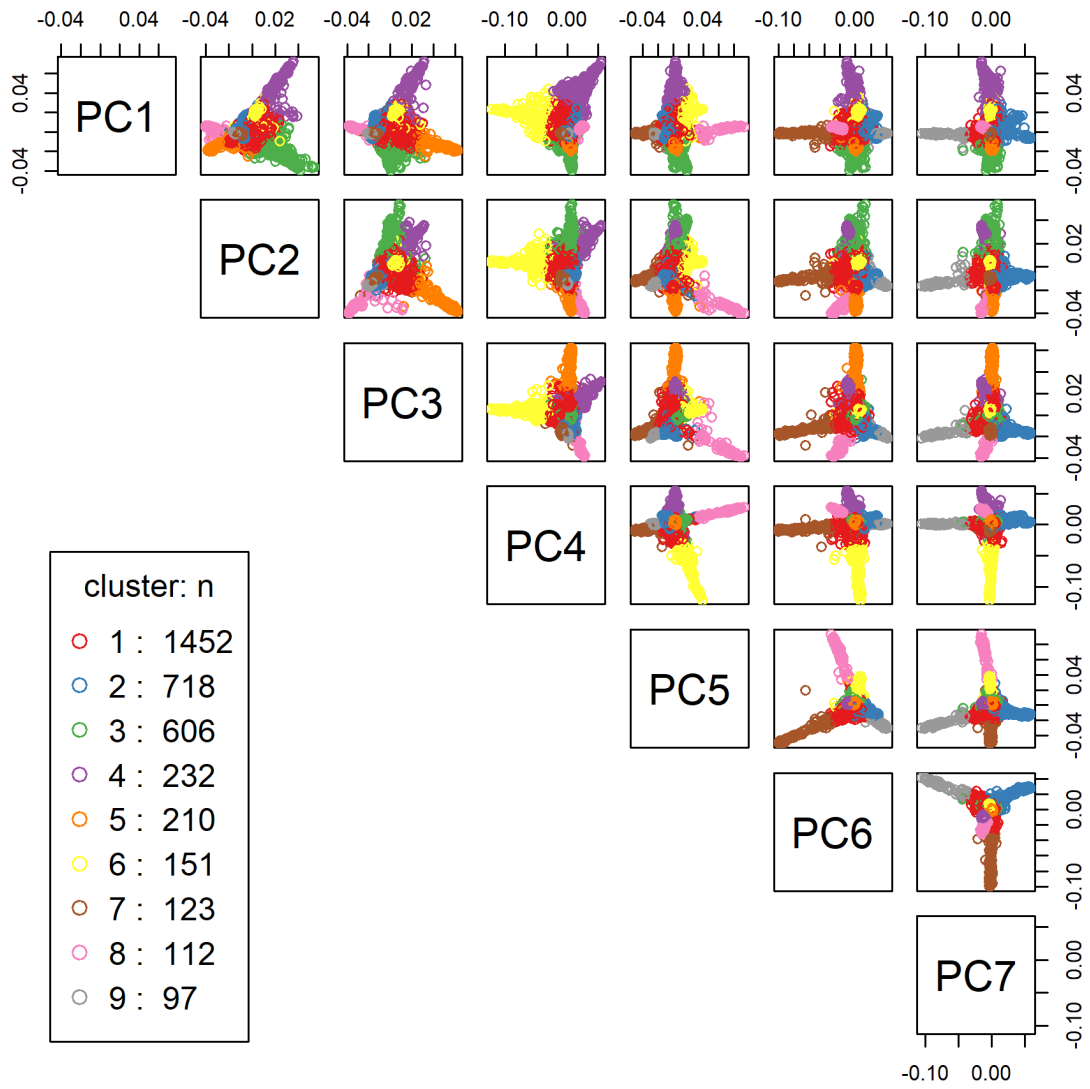



End
Mid-way
Neutral
Position (bp)
0
250000
500000
750000
1000000


End
Mid-way
Neutral
Position (bp)
0
250000
500000
750000
1000000

A.4 Supplementary material for Chapter 8

A.4.1 PCA

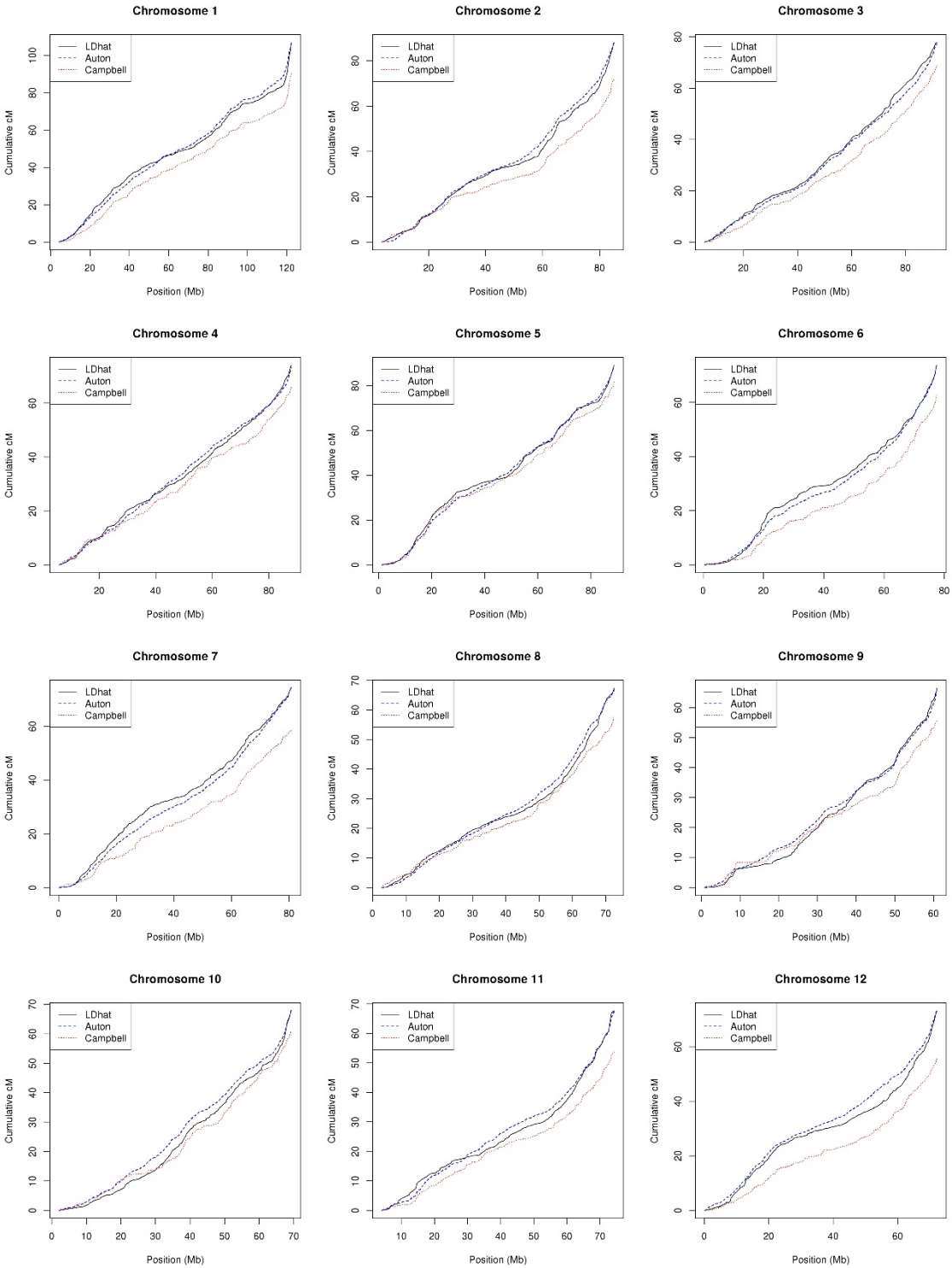


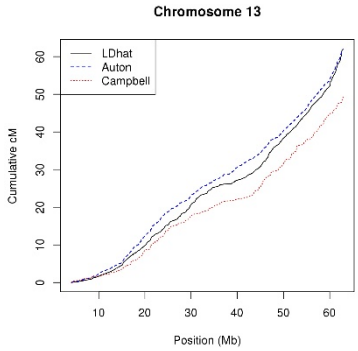
Appendix Figure 10 Pairwise PC plots of the dog data

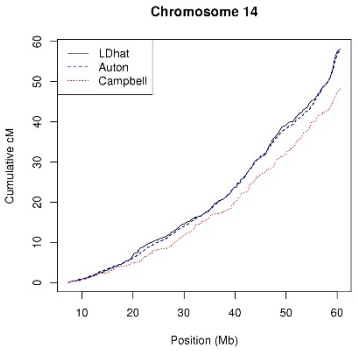
The principal components analysis of the dog data returned nine clusters of similar dogs. These are represented by different colours for each cluster. The analysis suggested seven PCs were appropriate to use. These graphs show a pairwise comparison of each PC.

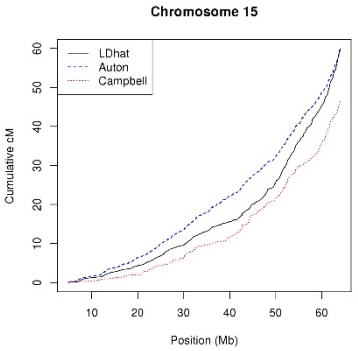
A.4.2 LD Maps

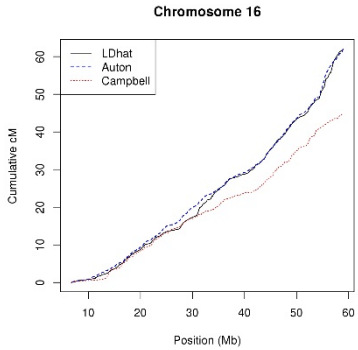
These figures show the final LDhat map for each autosome in the dog data. Also shown as a dashed blue line is the Auton LD map, and the Campbell linkage map is the dotted red line.

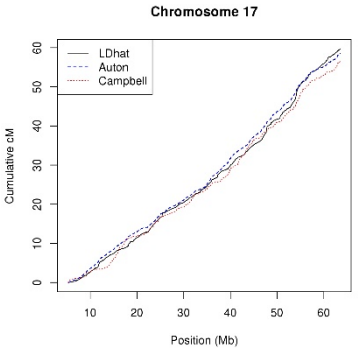


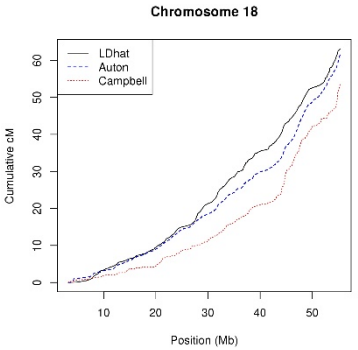

Chromosome 13
LDhat
Auton
Campbell
Cumulative cM
Position (Mb)

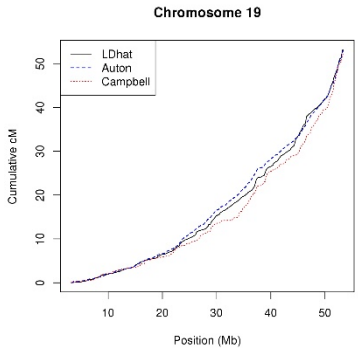

Chromosome 14
LDhat
Auton
Campbell
Cumulative cM
Position (Mb)

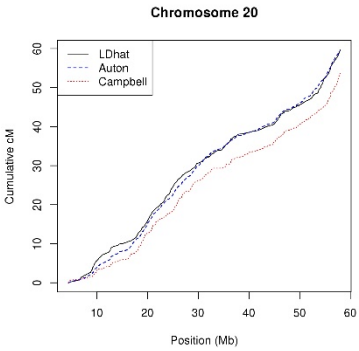

Chromosome 15
LDhat
Auton
Campbell
Cumulative cM
Position (Mb)

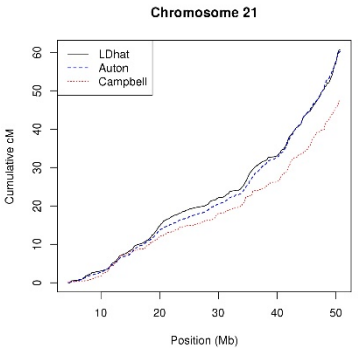

Chromosome 16
LDhat
Auton
Campbell
Cumulative cM
Position (Mb)

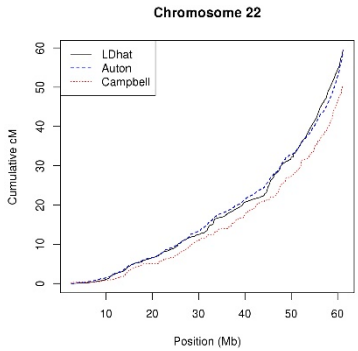

Chromosome 17
LDhat
Auton
Campbell
Cumulative cM
Position (Mb)

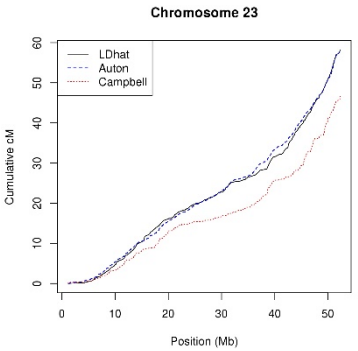

Chromosome 18
LDhat
Auton
Campbell
Cumulative cM
Position (Mb)

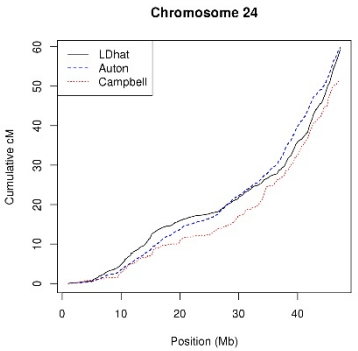

Chromosome 19
LDhat
Auton
Campbell
Cumulative cM
Position (Mb)

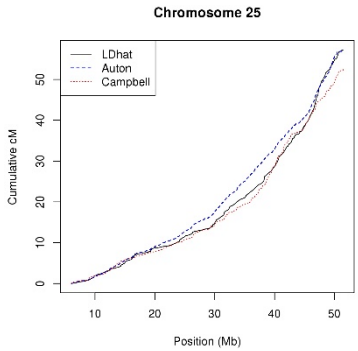

Chromosome 20
LDhat
Auton
Campbell
Cumulative cM
Position (Mb)

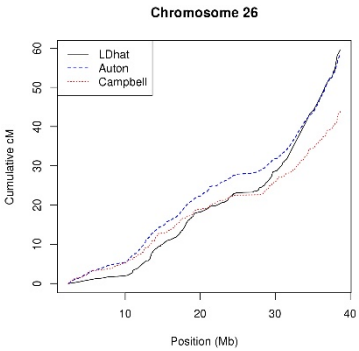

Chromosome 21
LDhat
Auton
Campbell
Cumulative cM
Position (Mb)

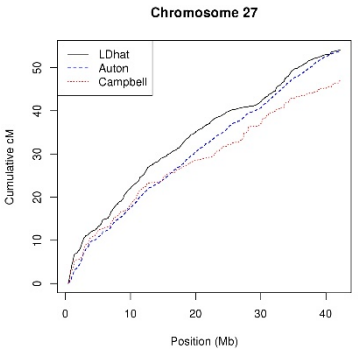

Chromosome 22
LDhat
Auton
Campbell
Cumulative cM
Position (Mb)

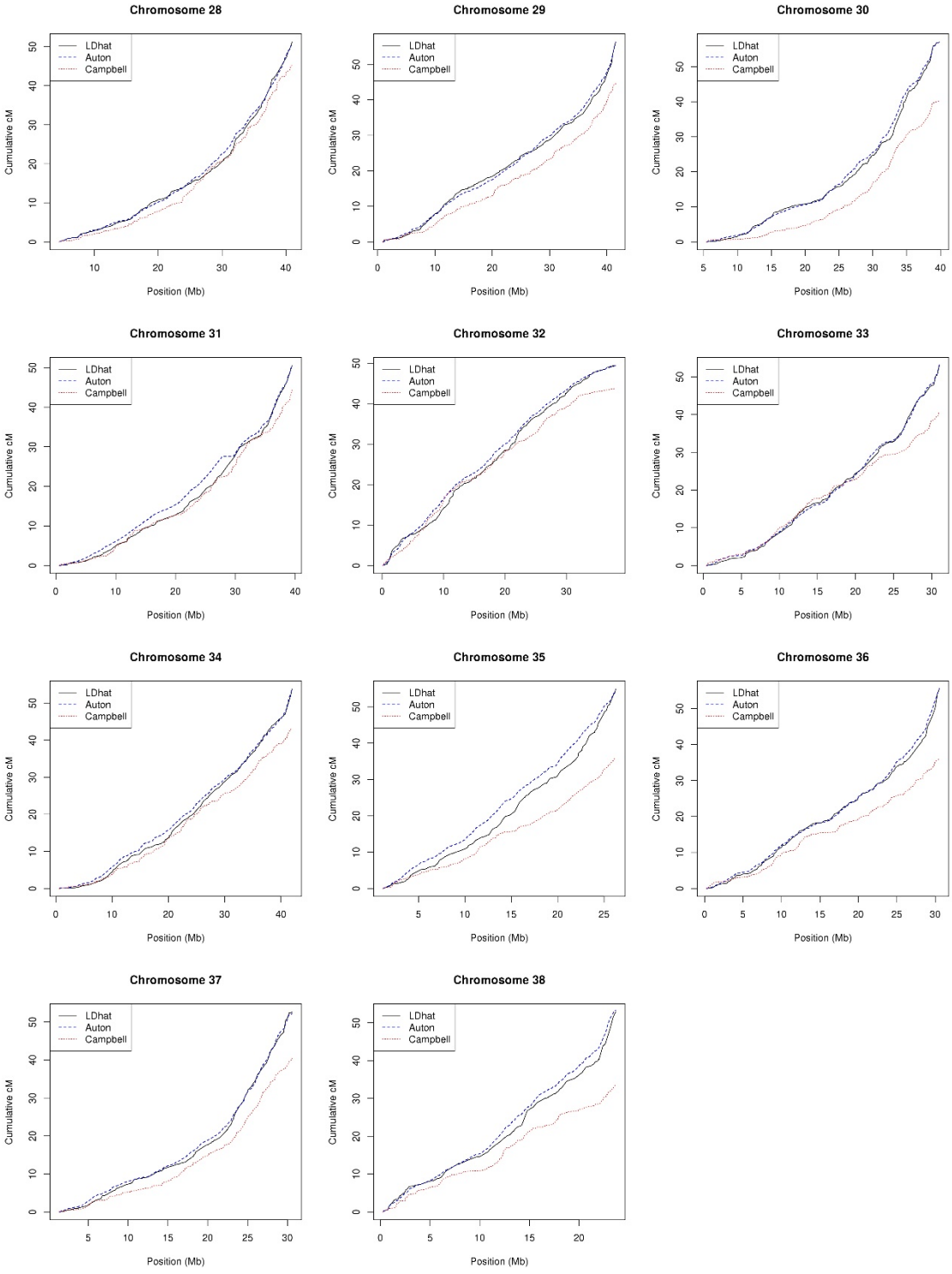

Chromosome 23
LDhat
Auton
Campbell
Cumulative cM
Position (Mb)


Chromosome 24
LDhat
Auton
Campbell
Cumulative cM
Position (Mb)


Chromosome 25
LDhat
Auton
Campbell
Cumulative cM
Position (Mb)


Chromosome 26
LDhat
Auton
Campbell
Cumulative cM
Position (Mb)


Chromosome 27
LDhat
Auton
Campbell
Cumulative cM
Position (Mb)


Chromosome 28
Chromosome 29
Chromosome 30
Chromosome 31
Chromosome 32
Chromosome 33
Chromosome 34
Chromosome 35
Chromosome 36
Chromosome 37
Chromosome 38
LDhat
Auton
Campbell
Cumulative cM
Position (Mb)

A.4.3 Candidate SNPs

Appendix Table 3 Candidate SNPs

This table shows all of the SNPs from the analysis where their result for $Z_\alpha^{r^2/E[r^2]}$ or $Z_\alpha^{BetaCDF}$ was in the top 0.1% of the values. The first two columns of the table show the chromosome and base pair location of each SNP. The next column is the Z_α value for the SNP, followed by the value in the observed empirical distribution for the Z_α statistic. This is followed by the values and empirical distribution values for the $Z_\alpha^{r^2/E[r^2]}$ and $Z_\alpha^{BetaCDF}$ statistics. A1 is the minor allele and A2 is the major allele, followed by the minor allele frequency (MAF). Ensembl's Variant Effect Predictor was used to find the consequences of the variants and the gene(s) involved, see method section 8.2.8. ChIPpeakAnno was used to find the nearest gene(s) either side of the variants. Finally, if the SNP was overlapped by a region previously published in one of the papers in Table 8-4 then the paper's first author is named in the final column.

Chr	bp	Z_α	Dist	$Z_\alpha^{r^2/E[r^2]}$	Dist	$Z_\alpha^{BetaCDF}$	Dist	A1	A2	MAF	VEP consequences	VEP Genes	ChIPpeakAnno Genes	Overlap
1	2,164,283	0.60	0.9999	2.56	0.9980	0.85	0.9997	A	G	8%	downstream_gene_variant	*ENSCAFG00000036410*	*ENSCAFG00000036410*	
1	20,947,142	0.42	0.9971	2.86	0.9993	0.84	0.9996	A	G	16%	intron_variant, non_coding_transcript_variant	*ENSCAFG00000042880*	*ENSCAFG00000042880*	Akey
1	21,141,240	0.28	0.9796	2.89	0.9994	0.60	0.9242	A	G	11%	intron_variant	*STARD6*	*STARD6*	Akey
1	28,449,172	0.45	0.9981	2.70	0.9988	0.81	0.9991	A	G	7%	intergenic_variant	-	*PDE7B*	Freedman
1	43,001,368	0.38	0.9951	2.85	0.9992	0.81	0.9989	A	C	14%	intergenic_variant	-	*ENSCAFG00000000540*	
1	60,262,327	0.66	1.0000	2.81	0.9991	0.87	0.9999	G	A	23%	intergenic_variant	-	*U4*	
1	77,413,224	0.30	0.9843	2.99	0.9995	0.73	0.9939	A	G	31%	intron_variant, non_coding_transcript_variant	*ENSCAFG00000046035*	*ENSCAFG00000046035*	

Chr	bp	Z_α	Dist	$Z_\alpha^{r^2/E[r^2]}$	Dist	$Z_\alpha^{BetaCDF}$	Dist	A1	A2	MAF	VEP consequences	VEP Genes	ChIPpeakAnno Genes	Overlap
1	77,423,529	0.31	0.9880	3.13	0.9997	0.74	0.9952	A	G	38%	intron_variant, non_coding_transcript_variant	*ENSCAFG00000046035*	*ENSCAFG00000046035*	
1	77,447,316	0.40	0.9961	3.84	1.0000	0.77	0.9973	A	G	20%	intergenic_variant	-	*ENSCAFG00000046035*	
1	77,451,390	0.36	0.9931	2.94	0.9994	0.77	0.9975	A	G	15%	intergenic_variant	-	*ENSCAFG00000046035*	
1	77,569,697	0.27	0.9765	2.85	0.9992	0.63	0.9563	C	G	27%	intron_variant	*TLE1*	*TLE1*	
1	96,115,461	0.33	0.9903	3.40	0.9999	0.87	0.9998	A	T	10%	intergenic_variant	-	*ENSCAFG00000047511*	
2	36,841,014	0.37	0.9947	3.25	0.9998	0.73	0.9947	A	G	26%	upstream_gene_variant	*PCDH1*	*PCDH1*	
2	36,852,293	0.32	0.9881	2.81	0.9991	0.73	0.9934	G	A	32%	intergenic_variant	-	*DELE1*	
2	44,796,266	0.36	0.9939	2.42	0.9969	0.82	0.9992	A	G	10%	intron_variant, non_coding_transcript_variant	*ENSCAFG00000046333*	*ENSCAFG00000046333*	
2	44,806,665	0.34	0.9920	2.39	0.9963	0.81	0.9992	C	A	25%	intron_variant, non_coding_transcript_variant	*ENSCAFG00000046333*	*ENSCAFG00000046333, ENSCAFG00000049983*	
2	61,876,498	0.47	0.9989	2.62	0.9984	0.83	0.9994	C	A	45%	intron_variant	*FTO*	*FTO*	Akey
2	61,880,556	0.49	0.9993	2.65	0.9985	0.84	0.9996	A	G	40%	intron_variant	*FTO*	*FTO*	Akey
2	61,897,779	0.48	0.9991	2.69	0.9987	0.84	0.9997	G	A	40%	intron_variant	*FTO*	*FTO*	Akey
2	61,901,702	0.52	0.9997	2.89	0.9994	0.86	0.9998	G	A	37%	intron_variant	*FTO*	*FTO*	Akey
2	71,434,345	0.33	0.9904	2.57	0.9981	0.85	0.9997	A	G	24%	downstream_gene_variant, intron_variant	*SRSF4*	*SRSF4*	
2	71,457,825	0.46	0.9986	3.15	0.9997	0.88	0.9999	G	A	17%	intron_variant	*SRSF4*	*SRSF4*	
2	71,471,740	0.46	0.9986	3.14	0.9997	0.88	0.9999	C	A	27%	downstream_gene_variant, missense_variant	*EPB41, SRSF4*	*SRSF4*	

Appendix A

Chr	bp	Z_α	Dist	$Z_\alpha^{r^2/E[r^2]}$	Dist	$Z_\alpha^{BetaCDF}$	Dist	A1	A2	MAF	VEP consequences	VEP Genes	ChIPpeakAnno Genes	Overlap
2	71,476,526	0.41	0.9965	2.78	0.9990	0.85	0.9998	G	A	18%	downstream_gene_variant, intron_variant	*EPB41, SRSF4*	*SRSF4, EPB41*	
2	71,490,861	0.41	0.9970	2.79	0.9991	0.85	0.9997	G	A	23%	intron_variant, missense_variant	*EPB41*	*EPB41*	
2	71,528,478	0.29	0.9821	2.59	0.9982	0.82	0.9993	A	G	16%	intron_variant	*EPB41*	*EPB41*	
3	13,415,724	0.32	0.9882	2.81	0.9991	0.79	0.9985	G	A	16%	intergenic_variant	-	*U6*	
3	17,490,492	0.56	0.9998	2.47	0.9974	0.84	0.9997	G	A	23%	intergenic_variant	-	*ENSCAFG00000044202*	
3	17,501,276	0.55	0.9998	2.47	0.9974	0.85	0.9997	C	A	25%	downstream_gene_variant	*ENSCAFG00000044202*	*ENSCAFG00000044202*	
3	17,516,194	0.60	0.9999	2.62	0.9984	0.86	0.9998	A	G	16%	intergenic_variant	-	*ENSCAFG00000044202*	
3	72,708,942	0.59	0.9999	2.98	0.9995	0.88	0.9999	A	C	50%	intergenic_variant	-	*UBE2K*	
4	14,421,521	0.33	0.9905	3.36	0.9999	0.87	0.9998	G	A	14%	intergenic_variant	-	*RTKN2*	
4	17,518,453	0.38	0.9949	2.36	0.9959	0.82	0.9994	A	G	12%	intron_variant	*CTNNA3*	*ENSCAFG00000045900, CTNNA3*	
4	48,548,524	0.36	0.9936	2.30	0.9951	0.81	0.9991	G	A	38%	intergenic_variant	-	*U2*	
4	48,567,088	0.39	0.9956	2.60	0.9982	0.83	0.9996	C	G	9%	intergenic_variant	-	*U2*	
4	48,573,221	0.36	0.9938	2.46	0.9973	0.81	0.9992	G	A	16%	intergenic_variant	-	*U2*	
4	57,345,395	0.32	0.9888	2.65	0.9985	0.82	0.9994	A	C	23%	intergenic_variant	-	*GLRA1*	
4	57,366,377	0.31	0.9879	2.68	0.9986	0.82	0.9992	A	G	18%	intergenic_variant	-	*GLRA1*	Freedman
5	2,932,294	0.55	0.9998	2.49	0.9975	0.83	0.9995	G	A	35%	intron_variant	*NTM*	*ENSCAFG00000039069, NTM*	
5	2,951,769	0.53	0.9997	2.42	0.9968	0.81	0.9990	G	A	39%	intron_variant	*NTM*	*ENSCAFG00000039069,*	

Chr	bp	Z_α	Dist	$Z_\alpha^{r^2/E[r^2]}$	Dist	$Z_\alpha^{BetaCDF}$	Dist	A1	A2	MAF	VEP consequences	VEP Genes	ChIPpeakAnno Genes	Overlap
													NTM	
5	4,064,061	0.67	1.0000	3.62	0.9999	0.92	1.0000	A	G	22%	intron_variant	*SNX19*	*SNX19*	Wang
5	4,093,514	0.45	0.9982	2.88	0.9993	0.74	0.9950	G	A	13%	upstream_gene_variant	*ENSCAFG00000044958*	*ENSCAFG00000044958*	Wang
5	4,118,722	0.44	0.9980	2.83	0.9992	0.73	0.9935	C	A	11%	intergenic_variant	-	*ENSCAFG00000044958*	Wang
5	4,132,302	0.49	0.9994	3.02	0.9996	0.76	0.9969	G	A	16%	intergenic_variant	-	*ENSCAFG00000044958*	Wang
5	6,811,533	0.41	0.9967	2.88	0.9993	0.81	0.9990	A	G	30%	intron_variant, non_coding_transcript_variant	*ENSCAFG00000045669*	*ENSCAFG00000045669*	
5	6,838,932	0.45	0.9983	3.19	0.9998	0.88	0.9999	A	G	17%	intron_variant, non_coding_transcript_variant	*ENSCAFG00000045669*	*ENSCAFG00000045669*	
5	6,845,530	0.44	0.9978	3.26	0.9998	0.88	0.9999	G	A	26%	intron_variant, non_coding_transcript_variant	*ENSCAFG00000045669*	*ENSCAFG00000045669*	
5	6,859,691	0.43	0.9977	3.20	0.9998	0.88	0.9999	G	A	9%	intergenic_variant	-	*ENSCAFG00000045669*	
5	6,907,781	0.33	0.9909	2.42	0.9968	0.81	0.9990	G	A	20%	intergenic_variant	-	*ENSCAFG00000045669*	
5	6,919,588	0.35	0.9925	2.49	0.9976	0.82	0.9994	A	G	12%	intergenic_variant	-	*ENSCAFG00000045669*	
5	6,991,724	0.40	0.9962	2.80	0.9991	0.78	0.9979	C	A	25%	intergenic_variant	-	*ENSCAFG00000045669*	
5	40,202,215	0.43	0.9977	3.04	0.9996	0.85	0.9998	A	G	40%	intergenic_variant	-	*AKAP10*	
5	47,359,645	0.41	0.9966	2.84	0.9992	0.79	0.9983	C	A	32%	upstream_gene_variant	*ATG4C*	*ATG4C*	
6	33,510,473	0.51	0.9995	2.57	0.9981	0.85	0.9997	G	C	48%	intron_variant	*METTL22*	*METTL22*	Akey, Freedman
7	11,956,306	0.25	0.9697	2.22	0.9937	0.83	0.9995	G	A	27%	intergenic_variant	-	*ENSCAFG00000048595*	
7	11,957,885	0.23	0.9560	2.21	0.9936	0.83	0.9995	C	A	16%	intergenic_variant	-	*ENSCAFG00000048595*	

Appendix A

Chr	bp	Z_α	Dist	$Z_\alpha^{r^2/E[r^2]}$	Dist	$Z_\alpha^{BetaCDF}$	Dist	A1	A2	MAF	VEP consequences	VEP Genes	ChIPpeakAnno Genes	Overlap
7	13,448,116	0.31	0.9880	3.86	1.0000	0.64	0.9621	A	G	10%	upstream_gene_variant	*TOR1AIP1*	*TOR1AIP1*	Wang
7	24,652,821	0.49	0.9993	2.84	0.9992	0.81	0.9991	G	A	5%	intron_variant	*RABGAP1L*	*RABGAP1L, ENSCAFG00000047576*	Axelsson
7	24,664,438	0.49	0.9994	2.82	0.9992	0.77	0.9974	A	G	8%	intron_variant, upstream_gene_variant	*RABGAP1L*	*RABGAP1L, ENSCAFG00000047576*	Axelsson
8	7,735,497	0.36	0.9932	2.33	0.9955	0.81	0.9991	A	G	45%	intergenic_variant	-	*ENSCAFG00000046540*	
8	21,330,931	0.29	0.9821	2.87	0.9993	0.79	0.9986	A	G	29%	intron_variant, non_coding_transcript_variant	*ENSCAFG00000047946*	*ENSCAFG00000047946, U6*	Wang
8	21,345,264	0.32	0.9884	3.16	0.9997	0.81	0.9991	C	G	15%	intron_variant, non_coding_transcript_variant	*ENSCAFG00000047946*	*ENSCAFG00000047946, U6*	
9	29,654,139	0.36	0.9932	3.29	0.9998	0.76	0.9971	G	A	11%	intergenic_variant	-	*U6*	
9	29,669,984	0.31	0.9877	3.09	0.9997	0.74	0.9950	A	G	7%	intergenic_variant	-	*U6*	
9	29,671,758	0.32	0.9886	3.22	0.9998	0.75	0.9964	A	G	8%	intergenic_variant	-	*U6*	
9	29,710,986	0.33	0.9905	3.16	0.9997	0.74	0.9951	A	T	10%	intergenic_variant	-	*ENSCAFG00000049515*	
9	29,718,219	0.36	0.9938	3.29	0.9998	0.74	0.9955	A	G	9%	intergenic_variant	-	*ENSCAFG00000049515*	
9	29,731,101	0.36	0.9933	3.16	0.9997	0.75	0.9959	A	G	9%	intergenic_variant	-	*ENSCAFG00000049515*	
9	29,742,489	0.37	0.9946	2.99	0.9995	0.73	0.9942	A	C	20%	intergenic_variant	-	*ENSCAFG00000049515*	
9	29,752,455	0.39	0.9955	3.19	0.9998	0.77	0.9977	G	A	30%	intergenic_variant	-	*ENSCAFG00000049515*	
9	29,779,751	0.32	0.9886	2.87	0.9993	0.71	0.9910	G	A	16%	intergenic_variant	-	*ENSCAFG00000049515*	
9	29,799,057	0.30	0.9852	2.84	0.9992	0.73	0.9941	A	G	7%	intergenic_variant	-	*ENSCAFG00000049515*	
9	29,814,161	0.32	0.9888	3.07	0.9996	0.75	0.9966	A	G	7%	intron_variant,	*ENSCAFG00000049515*	*ENSCAFG00000049515*	

Chr	bp	Z_α	Dist	$Z_\alpha^{r^2/E[r^2]}$	Dist	$Z_\alpha^{BetaCDF}$	Dist	A1	A2	MAF	VEP consequences	VEP Genes	ChIPpeakAnno Genes	Overlap
											non_coding_transcript_variant			
9	34,984,408	0.47	0.9989	2.41	0.9967	0.81	0.9991	A	G	9%	downstream_gene_variant	*BRIP1*	*ENSCAFG00000017738*	
9	57,439,074	0.35	0.9930	2.90	0.9994	0.82	0.9993	A	G	30%	intergenic_variant	-	*MAPKAP1*	
10	44,372,549	0.36	0.9938	3.12	0.9997	0.81	0.9990	A	G	39%	intron_variant	*VWA3B*	*CNGA3, VWA3B*	Wang
10	44,388,924	0.43	0.9974	3.40	0.9999	0.84	0.9996	A	G	12%	intron_variant	*VWA3B*	*VWA3B*	Wang
10	44,534,551	0.29	0.9834	2.78	0.9990	0.70	0.9901	G	A	21%	intergenic_variant	-	*VWA3B*	Wang
10	44,543,279	0.34	0.9919	2.92	0.9994	0.73	0.9946	A	G	6%	intergenic_variant	-	*VWA3B*	Wang
10	46,053,118	0.50	0.9994	3.29	0.9998	0.85	0.9998	A	C	7%	missense_variant	*THADA*	*THADA, ENSCAFG00000047957*	
11	9,844,519	0.45	0.9982	2.80	0.9991	0.76	0.9968	G	A	5%	intergenic_variant	-	*ENSCAFG00000048218*	Freedman
11	54,017,181	0.30	0.9861	2.84	0.9992	0.72	0.9920	A	C	6%	intron_variant	*FRMPD1*	*FRMPD1, TRMT10B*	Axelsson, Wang
11	54,049,858	0.33	0.9909	3.65	1.0000	0.77	0.9978	G	A	6%	missense_variant	*FRMPD1*	*FRMPD1, TRMT10B*	Axelsson, Wang
11	54,049,870	0.33	0.9906	3.78	1.0000	0.75	0.9964	G	A	10%	missense_variant	*FRMPD1*	*FRMPD1, TRMT10B*	Axelsson, Wang
11	54,156,304	0.34	0.9914	3.30	0.9998	0.72	0.9927	A	G	9%	downstream_gene_variant	*ENSCAFG00000002395*	*ENSCAFG00000002395*	Akey
11	54,324,689	0.38	0.9949	3.58	0.9999	0.83	0.9994	A	G	8%	intron_variant, upstream_gene_variant	*ENSCAFG00000043130, SHB*	*SHB*	Akey
11	54,347,903	0.40	0.9963	4.00	1.0000	0.89	1.0000	A	G	10%	intron_variant, non_coding_transcript_variant	*ENSCAFG00000043130*	*ENSCAFG00000043130*	Akey
11	54,368,623	0.40	0.9962	4.02	1.0000	0.90	1.0000	G	A	18%	intron_variant,	*ENSCAFG00000043130*	*ENSCAFG00000043130*	Akey

Chr	bp	Z_α	Dist	$Z_\alpha^{r^2/E[r^2]}$	Dist	$Z_\alpha^{BetaCDF}$	Dist	A1	A2	MAF	VEP consequences	VEP Genes	ChIPpeakAnno Genes	Overlap
											non_coding_transcript_variant			
11	54,391,443	0.37	0.9943	3.28	0.9998	0.89	1.0000	G	A	9%	intron_variant, non_coding_transcript_variant	*ENSCAFG00000043130*	*ENSCAFG00000043130*	Akey
12	26,284,264	0.33	0.9906	2.67	0.9986	0.85	0.9997	A	G	7%	intergenic_variant	-	*ENSCAFG00000002473*	
12	30,314,914	0.55	0.9998	2.71	0.9988	0.84	0.9996	A	G	33%	intergenic_variant	-	*ENSCAFG00000042430*	
12	31,691,990	0.47	0.9990	3.17	0.9997	0.79	0.9984	A	G	12%	intron_variant	*ADGRB3*	*ADGRB3, U6*	
12	31,745,290	0.51	0.9996	3.63	1.0000	0.90	1.0000	A	G	17%	intron_variant	*ADGRB3*	*ADGRB3, U6*	
12	31,761,177	0.38	0.9952	2.73	0.9989	0.84	0.9996	G	A	9%	intron_variant	*ADGRB3*	*ADGRB3, U6*	
12	31,805,128	0.33	0.9908	2.50	0.9976	0.84	0.9996	G	C	8%	intron_variant	*ADGRB3*	*ADGRB3, U6*	
12	31,820,134	0.36	0.9937	2.65	0.9985	0.85	0.9997	A	G	9%	intron_variant	*ADGRB3*	*ADGRB3, U6*	
12	31,835,704	0.45	0.9984	2.97	0.9995	0.87	0.9998	C	A	8%	intron_variant, upstream_gene_variant	*ADGRB3, U6*	*ADGRB3, U6*	
12	39,245,810	0.49	0.9993	2.97	0.9995	0.81	0.9990	G	A	41%	intergenic_variant	-	*U6*	
12	47,393,616	0.27	0.9745	3.06	0.9996	0.60	0.9204	A	G	10%	intergenic_variant	-	*SPACA1*	
12	47,562,731	0.26	0.9723	3.03	0.9996	0.69	0.9873	A	G	26%	upstream_gene_variant	*CNR1*	*CNR1*	
13	11,012,218	0.30	0.9849	2.78	0.9990	0.75	0.9966	A	G	17%	intergenic_variant	-	*ENSCAFG00000043478*	
13	14,702,870	0.34	0.9911	2.35	0.9958	0.81	0.9992	A	G	36%	intergenic_variant	-	*U6*	
14	8,117,811	0.38	0.9951	2.78	0.9990	0.86	0.9998	A	G	5%	3_prime_UTR_variant	*LEP*	*LEP*	
14	17,850,921	0.36	0.9932	2.90	0.9994	0.79	0.9983	G	A	17%	intron_variant	*ANKIB1, KRIT1*	*KRIT1, ANKIB1*	
14	32,503,168	0.44	0.9981	2.24	0.9941	0.82	0.9994	G	A	30%	intergenic_variant	-	*ENSCAFG00000026754*	

Chr	bp	Z_α	Dist	$Z_\alpha^{r^2/E[r^2]}$	Dist	$Z_\alpha^{BetaCDF}$	Dist	A1	A2	MAF	VEP consequences	VEP Genes	ChIPpeakAnno Genes	Overlap
14	32,522,229	0.41	0.9966	2.14	0.9921	0.81	0.9991	A	C	35%	intergenic_variant	-	*ENSCAFG00000026754*	
14	32,529,441	0.41	0.9968	2.15	0.9923	0.81	0.9991	A	C	29%	intergenic_variant	-	*ENSCAFG00000026754*	
14	32,540,148	0.45	0.9984	2.32	0.9954	0.83	0.9995	G	A	41%	intergenic_variant	-	*ENSCAFG00000026754*	
14	32,568,553	0.42	0.9971	2.22	0.9937	0.81	0.9990	A	G	24%	intergenic_variant	-	*HDAC9*	
15	20,317,533	0.59	0.9999	2.56	0.9980	0.82	0.9993	A	C	27%	intergenic_variant	-	*ENSCAFG00000005722*	Akey
15	26,751,372	0.27	0.9753	2.81	0.9991	0.73	0.9935	C	A	9%	intron_variant	*LRRIQ1*	*LRRIQ1, U6*	
15	26,965,818	0.35	0.9929	2.96	0.9995	0.66	0.9744	G	A	37%	upstream_gene_variant	*ENSCAFG00000042655*	*ENSCAFG00000042655*	
16	7,462,818	0.61	1.0000	3.13	0.9997	0.81	0.9992	A	G	37%	downstream_gene_variant, upstream_gene_variant	*SSBP1, WEE2*	*SSBP1*	Cagan
16	13,634,700	0.48	0.9991	2.81	0.9991	0.83	0.9995	T	A	40%	3_prime_UTR_variant	*LYPD8*	*LYPD8*	
16	13,634,890	0.42	0.9973	2.69	0.9987	0.83	0.9995	A	C	14%	missense_variant	*LYPD8*	*LYPD8*	
16	13,670,264	0.34	0.9919	2.51	0.9977	0.82	0.9993	G	A	15%	intergenic_variant	-	*LYPD8*	
17	3,753,156	0.47	0.9987	2.68	0.9986	0.85	0.9997	A	C	14%	intergenic_variant	-	*ENSCAFG00000041862*	
18	9,493,237	0.46	0.9986	3.00	0.9995	0.72	0.9927	A	G	17%	intron_variant	*SUGCT*	*SUGCT, ENSCAFG00000028433*	
18	9,655,138	0.42	0.9973	2.82	0.9992	0.68	0.9822	A	C	15%	intron_variant	*SUGCT*	*SUGCT, ENSCAFG00000028433*	
18	19,746,195	0.41	0.9969	3.30	0.9999	0.91	1.0000	A	G	14%	intergenic_variant	-	*ENSCAFG00000040428*	
18	24,164,381	0.28	0.9800	3.25	0.9998	0.86	0.9998	G	A	25%	downstream_gene_variant	*ENSCAFG00000044560*	*ENSCAFG00000044560*	Akey
18	24,196,399	0.28	0.9791	3.41	0.9999	0.86	0.9998	A	G	16%	intergenic_variant	-	*ENSCAFG00000044560*	Akey

Appendix A

Chr	bp	Z_α	Dist	$Z_\alpha^{r^2/E[r^2]}$	Dist	$Z_\alpha^{BetaCDF}$	Dist	A1	A2	MAF	VEP consequences	VEP Genes	ChIPpeakAnno Genes	Overlap
18	24,209,031	0.26	0.9702	2.99	0.9995	0.80	0.9988	A	G	10%	intergenic_variant	-	*ENSCAFG00000046418*	Akey
18	24,261,759	0.29	0.9819	3.03	0.9996	0.80	0.9987	G	A	13%	downstream_gene_variant	*SEMA3D*	*ENSCAFG00000046418*	Akey, Freedman, Wang
18	24,286,833	0.27	0.9778	3.04	0.9996	0.82	0.9994	A	G	17%	intron_variant	*SEMA3D*	*ENSCAFG00000046418, SEMA3D*	Akey, Freedman, Wang
18	24,292,509	0.29	0.9829	3.33	0.9999	0.88	1.0000	G	A	13%	intron_variant	*SEMA3D*	*ENSCAFG00000046418, SEMA3D*	Akey, Freedman, Wang
18	24,303,383	0.29	0.9831	3.27	0.9998	0.88	1.0000	A	G	14%	intron_variant	*SEMA3D*	*SEMA3D, U6*	Akey, Freedman, Wang
18	24,312,302	0.33	0.9908	3.15	0.9997	0.87	0.9998	A	G	13%	intron_variant	*SEMA3D*	*SEMA3D, U6*	Akey, Freedman, Wang
18	29,130,730	0.23	0.9543	3.13	0.9997	0.76	0.9970	G	A	12%	intergenic_variant	-	*ENSCAFG00000044495*	
18	29,299,675	0.23	0.9580	2.98	0.9995	0.69	0.9868	A	G	16%	intergenic_variant	-	*ENSCAFG00000046942*	
18	29,324,878	0.23	0.9584	2.83	0.9992	0.66	0.9771	A	G	14%	intergenic_variant	-	*ENSCAFG00000046942*	
18	29,376,574	0.28	0.9790	2.79	0.9991	0.70	0.9890	A	G	27%	intergenic_variant	-	*ENSCAFG00000046942*	
18	29,595,073	0.27	0.9761	2.79	0.9991	0.80	0.9989	A	G	23%	upstream_gene_variant	*ENSCAFG00000046942*	*ENSCAFG00000046942*	
18	41,422,687	0.39	0.9955	2.91	0.9994	0.71	0.9908	G	A	22%	upstream_gene_variant	*ENSCAFG00000043513*	*ENSCAFG00000043513*	
18	41,445,354	0.32	0.9895	2.78	0.9991	0.65	0.9709	G	A	17%	missense_variant	*ENSCAFG00000008258*	*ENSCAFG00000008258*	
19	4,767,099	0.51	0.9995	2.51	0.9977	0.81	0.9992	G	A	31%	intergenic_variant	-	*ENSCAFG00000003755*	Akey

Chr	bp	Z_α	Dist	$Z_\alpha^{r^2/E[r^2]}$	Dist	$Z_\alpha^{BetaCDF}$	Dist	A1	A2	MAF	VEP consequences	VEP Genes	ChIPpeakAnno Genes	Overlap
19	4,771,876	0.54	0.9998	2.68	0.9986	0.82	0.9993	A	G	32%	intergenic_variant	-	*ENSCAFG00000003755*	Akey
19	4,798,748	0.60	0.9999	2.92	0.9994	0.84	0.9996	A	G	31%	intergenic_variant	-	*ENSCAFG00000003755*	Akey
19	4,813,917	0.60	1.0000	2.94	0.9994	0.87	0.9999	A	G	15%	intergenic_variant	-	*ENSCAFG00000003755*	Akey
19	6,162,402	0.55	0.9998	3.08	0.9996	0.88	0.9999	G	A	19%	intergenic_variant	-	*ENSCAFG00000038491*	Akey
19	6,178,251	0.53	0.9997	2.77	0.9990	0.84	0.9996	G	A	26%	intergenic_variant	-	*ENSCAFG00000038491*	Akey
19	6,201,219	0.44	0.9978	2.78	0.9990	0.84	0.9996	A	G	20%	intergenic_variant	-	*ENSCAFG00000038491*	Akey
19	6,216,997	0.41	0.9968	3.02	0.9996	0.85	0.9998	G	A	24%	intergenic_variant	-	*ENSCAFG00000038491*	Akey
19	6,553,427	0.57	0.9999	2.85	0.9992	0.88	0.9999	G	A	44%	intergenic_variant	-	*ENSCAFG00000038491*	Akey
19	6,560,183	0.60	1.0000	2.98	0.9995	0.88	0.9999	A	G	32%	intergenic_variant	-	*ENSCAFG00000038491*	Akey
19	6,590,666	0.61	1.0000	2.94	0.9995	0.88	0.9999	A	T	22%	intergenic_variant	-	*ENSCAFG00000038491*	Akey
19	6,629,569	0.48	0.9991	2.27	0.9948	0.82	0.9993	A	G	41%	intergenic_variant	-	*ENSCAFG00000041840*	Akey
19	6,648,984	0.47	0.9989	2.25	0.9944	0.82	0.9993	A	C	23%	intergenic_variant	-	*ENSCAFG00000041840*	
19	6,663,845	0.45	0.9983	2.17	0.9927	0.81	0.9991	G	A	24%	intergenic_variant	-	*ENSCAFG00000041840*	
19	6,672,903	0.45	0.9983	2.16	0.9924	0.81	0.9990	A	C	46%	intergenic_variant	-	*ENSCAFG00000041840*	
19	6,689,197	0.46	0.9986	2.26	0.9946	0.82	0.9993	G	A	23%	intergenic_variant	-	*ENSCAFG00000041840*	
19	6,728,854	0.50	0.9995	2.39	0.9965	0.83	0.9995	A	C	39%	intergenic_variant	-	*ENSCAFG00000041840*	
19	6,741,733	0.50	0.9995	2.35	0.9959	0.83	0.9995	A	C	27%	intergenic_variant	-	*ENSCAFG00000041840*	
19	6,926,648	0.55	0.9998	2.68	0.9987	0.81	0.9991	A	C	26%	intergenic_variant	-	*ENSCAFG00000041840*	
19	6,970,430	0.47	0.9990	2.87	0.9993	0.76	0.9967	A	G	26%	intergenic_variant	-	*ENSCAFG00000041840*	

Appendix A

Chr	bp	Z_α	Dist	$Z_\alpha^{r^2/E[r^2]}$	Dist	$Z_\alpha^{BetaCDF}$	Dist	A1	A2	MAF	VEP consequences	VEP Genes	ChIPpeakAnno Genes	Overlap
19	7,095,253	0.55	0.9998	3.16	0.9997	0.84	0.9996	C	A	23%	intergenic_variant	-	*ENSCAFG00000041840*	
19	7,097,389	0.52	0.9997	3.33	0.9999	0.84	0.9997	C	A	25%	intergenic_variant	-	*ENSCAFG00000041840*	
19	7,117,822	0.51	0.9995	3.35	0.9999	0.83	0.9995	G	A	25%	intergenic_variant	-	*ENSCAFG00000041840*	
19	7,122,489	0.49	0.9994	3.24	0.9998	0.82	0.9993	A	G	45%	intergenic_variant	-	*ENSCAFG00000041840*	
19	12,407,112	0.40	0.9963	3.02	0.9996	0.77	0.9973	A	G	9%	intron_variant, non_coding_transcript_variant	*ENSCAFG00000046515*	*ENSCAFG00000046515*	
20	2,971,861	0.41	0.9970	2.41	0.9967	0.81	0.9991	G	A	10%	3_prime_UTR_variant	*ISY1*	*ISY1, RAB43*	
20	8,744,328	0.36	0.9936	3.04	0.9996	0.78	0.9980	G	A	44%	intergenic_variant	-	*SETD5*	
20	8,894,743	0.16	0.8584	2.95	0.9995	0.71	0.9907	C	A	15%	5_prime_UTR_variant	*SRGAP3*	*SRGAP3*	
20	12,119,654	0.35	0.9922	2.79	0.9991	0.80	0.9987	G	A	9%	intergenic_variant	-	*ENSCAFG00000019297*	
20	13,387,022	0.40	0.9962	2.78	0.9990	0.83	0.9995	A	G	11%	intergenic_variant	-	*ENSCAFG00000005959*	
20	18,037,927	0.18	0.8859	2.87	0.9993	0.75	0.9959	C	A	43%	intergenic_variant	-	*CNTN3*	
20	18,060,817	0.17	0.8689	2.87	0.9993	0.78	0.9979	A	G	26%	intergenic_variant	-	*CNTN3*	
20	18,066,749	0.17	0.8806	2.99	0.9995	0.82	0.9992	A	C	39%	intergenic_variant	-	*CNTN3*	
20	18,076,728	0.17	0.8850	3.01	0.9996	0.79	0.9984	A	G	35%	intergenic_variant	-	*CNTN3*	
20	18,090,687	0.16	0.8613	2.87	0.9993	0.78	0.9983	A	G	16%	intergenic_variant	-	*ENSCAFG00000036778*	
20	18,099,570	0.16	0.8584	2.88	0.9993	0.79	0.9983	A	G	48%	intergenic_variant	-	*ENSCAFG00000036778*	
20	23,922,281	0.24	0.9642	2.93	0.9994	0.69	0.9864	G	A	26%	intron_variant	*SUCLG2*	*SUCLG2*	
20	35,581,314	0.28	0.9798	2.82	0.9992	0.78	0.9981	G	A	23%	intron_variant	*CACNA2D3*	*CACNA2D3, LRTM1*	

Chr	bp	Z_α	Dist	$Z_\alpha^{r^2/E[r^2]}$	Dist	$Z_\alpha^{BetaCDF}$	Dist	A1	A2	MAF	VEP consequences	VEP Genes	ChIPpeakAnno Genes	Overlap
20	35,587,808	0.28	0.9784	2.97	0.9995	0.78	0.9981	A	G	17%	intron_variant	*CACNA2D3*	*CACNA2D3, LRTM1*	
20	35,732,329	0.29	0.9816	3.17	0.9997	0.80	0.9987	A	G	40%	intron_variant	*CACNA2D3*	*CACNA2D3*	
20	35,738,272	0.29	0.9833	3.43	0.9999	0.79	0.9985	A	G	26%	intron_variant	*CACNA2D3*	*CACNA2D3*	
20	36,834,508	0.22	0.9494	2.85	0.9992	0.51	0.6238	C	A	49%	intergenic_variant	-	*ENSCAFG00000041071*	
22	11,073,667	0.29	0.9833	2.79	0.9991	0.82	0.9994	G	A	42%	intergenic_variant	-	*ENSCAFG00000045070*	
22	12,027,888	0.59	0.9999	2.57	0.9981	0.81	0.9990	A	G	35%	intron_variant, non_coding_transcript_variant	*ENSCAFG00000042444*	*ENSCAFG00000042444, U6*	
22	12,039,716	0.61	1.0000	2.88	0.9993	0.81	0.9990	C	A	38%	intron_variant, non_coding_transcript_variant	*ENSCAFG00000042444*	*ENSCAFG00000042444, U6*	
22	12,064,068	0.60	0.9999	2.88	0.9993	0.80	0.9987	A	G	23%	intron_variant, non_coding_transcript_variant	*ENSCAFG00000042444*	*ENSCAFG00000042444, U6*	
22	17,089,718	0.43	0.9976	2.15	0.9922	0.82	0.9992	A	G	46%	intergenic_variant	-	*ENSCAFG00000048196*	
22	17,102,316	0.41	0.9970	2.15	0.9923	0.82	0.9992	A	G	50%	intergenic_variant	-	*PCDH20*	
22	17,113,959	0.42	0.9973	2.24	0.9941	0.82	0.9994	C	G	31%	intergenic_variant	-	*PCDH20*	
22	17,129,084	0.42	0.9973	2.25	0.9944	0.82	0.9994	G	A	46%	intergenic_variant	-	*PCDH20*	
22	17,133,195	0.47	0.9987	2.48	0.9974	0.84	0.9997	G	A	49%	intergenic_variant	-	*PCDH20*	
22	17,153,150	0.53	0.9997	2.66	0.9986	0.85	0.9998	T	A	15%	missense_variant, upstream_gene_variant	*ENSCAFG00000038721, PCDH20*	*PCDH20*	
22	17,154,006	0.60	0.9999	2.94	0.9994	0.88	0.9999	A	G	38%	synonymous_variant, upstream_gene_variant	*ENSCAFG00000038721, PCDH20*	*PCDH20*	
22	17,166,191	0.56	0.9999	2.79	0.9991	0.87	0.9999	A	C	49%	intron_variant, non_coding_transcript_variant	*ENSCAFG00000038721*	*ENSCAFG00000038721*	

Appendix A

Chr	bp	Z_α	Dist	$Z_\alpha^{r^2/E[r^2]}$	Dist	$Z_\alpha^{BetaCDF}$	Dist	A1	A2	MAF	VEP consequences	VEP Genes	ChIPpeakAnno Genes	Overlap
22	18,774,821	0.45	0.9984	2.49	0.9976	0.83	0.9996	G	A	7%	intron_variant, non_coding_transcript_variant	*ENSCAFG00000047507, ENSCAFG00000049065*	*ENSCAFG00000047507, ENSCAFG00000049065*	
22	18,960,901	0.43	0.9978	2.40	0.9966	0.81	0.9992	A	G	30%	intergenic_variant	-	*U6*	
22	18,962,347	0.43	0.9974	2.45	0.9972	0.83	0.9996	A	G	10%	intergenic_variant	-	*U6*	
22	19,870,809	0.40	0.9964	2.25	0.9944	0.82	0.9993	G	A	38%	intergenic_variant	-	*ENSCAFG00000046482*	
22	19,925,395	0.43	0.9975	2.31	0.9953	0.81	0.9991	G	A	14%	intergenic_variant	-	*ENSCAFG00000046482*	
22	19,975,468	0.44	0.9980	2.38	0.9963	0.82	0.9994	G	A	5%	intergenic_variant	-	*ENSCAFG00000046482*	
22	23,003,825	0.47	0.9989	2.33	0.9955	0.83	0.9995	G	C	12%	intergenic_variant	-	*ENSCAFG00000026838*	
22	29,093,614	0.38	0.9953	2.77	0.9990	0.84	0.9996	A	G	47%	upstream_gene_variant	*U1*	*U1*	
22	31,194,138	0.31	0.9876	3.36	0.9999	0.66	0.9768	G	C	38%	intergenic_variant	-	*SLAIN1*	
22	31,334,345	0.22	0.9491	2.79	0.9991	0.69	0.9878	A	C	7%	intergenic_variant	-	*EDNRB*	
22	31,347,124	0.26	0.9730	3.31	0.9999	0.68	0.9837	G	A	30%	intergenic_variant	-	*EDNRB*	
22	35,859,272	0.26	0.9726	2.78	0.9990	0.68	0.9828	G	A	16%	intergenic_variant	-	*ENSCAFG00000005282*	
22	35,875,929	0.30	0.9848	3.04	0.9996	0.69	0.9862	C	A	6%	intergenic_variant	-	*ENSCAFG00000005282*	
22	36,027,691	0.30	0.9849	2.95	0.9995	0.65	0.9715	G	A	7%	intergenic_variant	-	*ENSCAFG00000005282*	
22	36,040,150	0.31	0.9876	3.41	0.9999	0.66	0.9752	A	C	23%	intergenic_variant	-	*ENSCAFG00000005282*	
22	42,724,459	0.46	0.9987	2.34	0.9956	0.82	0.9993	G	A	42%	intron_variant	*GPC5*	*GPC5, U4*	
22	44,550,605	0.35	0.9930	2.90	0.9994	0.68	0.9848	G	A	15%	intron_variant	*GPC6*	*GPC6*	
24	291,964	0.48	0.9991	2.31	0.9953	0.83	0.9995	G	A	36%	downstream_gene_variant, intron_variant	*GZF1*	*GZF1*	

Chr	bp	Z_α	Dist	$Z_\alpha^{r^2/E[r^2]}$	Dist	$Z_\alpha^{BetaCDF}$	Dist	A1	A2	MAF	VEP consequences	VEP Genes	ChIPpeakAnno Genes	Overlap
24	454,092	0.50	0.9995	2.17	0.9927	0.82	0.9992	G	A	29%	intergenic_variant	-	*ENSCAFG00000041583*	
26	21,573,616	0.51	0.9996	2.36	0.9959	0.82	0.9992	A	G	41%	intron_variant	*TTC28*	*PITPNB, TTC28*	
26	22,151,015	0.35	0.9928	3.99	1.0000	0.68	0.9846	A	G	29%	downstream_gene_variant, intron_variant	*CCDC117, TTC28, U6*	*TTC28, CCDC117, U6*	
26	22,156,289	0.25	0.9662	2.83	0.9992	0.62	0.9506	T	A	29%	3_prime_UTR_variant, intron_variant	*CCDC117, TTC28*	*TTC28, CCDC117, U6*	
27	44,328,723	0.48	0.9991	2.47	0.9974	0.83	0.9994	A	G	18%	intron_variant	*CACNA1C*	*CACNA1C, ENSCAFG00000016073*	
28	8,210,333	0.47	0.9987	3.06	0.9996	0.70	0.9897	A	T	9%	missense_variant	*PLCE1*	*PLCE1*	Freedman
28	8,210,550	0.45	0.9984	2.92	0.9994	0.67	0.9801	C	A	11%	missense_variant	*PLCE1*	*PLCE1*	Freedman
28	8,216,688	0.47	0.9988	2.88	0.9993	0.68	0.9829	G	A	8%	intron_variant	*PLCE1*	*PLCE1*	Freedman
28	8,230,318	0.49	0.9994	2.91	0.9994	0.69	0.9858	G	A	10%	intron_variant	*PLCE1*	*PLCE1*	Freedman
28	10,677,902	0.33	0.9900	2.13	0.9917	0.81	0.9991	G	A	32%	intron_variant	*RRP12*	*FRAT2, RRP12*	
30	1,552,291	0.52	0.9997	2.30	0.9951	0.82	0.9994	G	A	9%	intergenic_variant	-	*RYR3*	vonHoldt
30	1,558,195	0.59	0.9999	2.56	0.9980	0.85	0.9997	A	G	8%	intergenic_variant	-	*RYR3*	vonHoldt
30	1,732,646	0.53	0.9997	2.44	0.9971	0.82	0.9994	T	A	9%	intergenic_variant	-	*ENSCAFG00000008172*	vonHoldt
30	1,744,087	0.51	0.9995	2.36	0.9960	0.83	0.9995	A	G	7%	intron_variant	*ENSCAFG00000008172*	*ENSCAFG00000008172*	vonHoldt
30	4,822,803	0.49	0.9994	2.72	0.9988	0.85	0.9997	G	A	17%	intergenic_variant	-	*ENSCAFG00000000808*	
30	4,880,566	0.45	0.9983	2.46	0.9972	0.82	0.9992	A	C	15%	intergenic_variant	-	*ENSCAFG00000000808*	
31	15,063,496	0.35	0.9925	2.44	0.9970	0.82	0.9993	A	G	30%	intergenic_variant	-	*U6*	

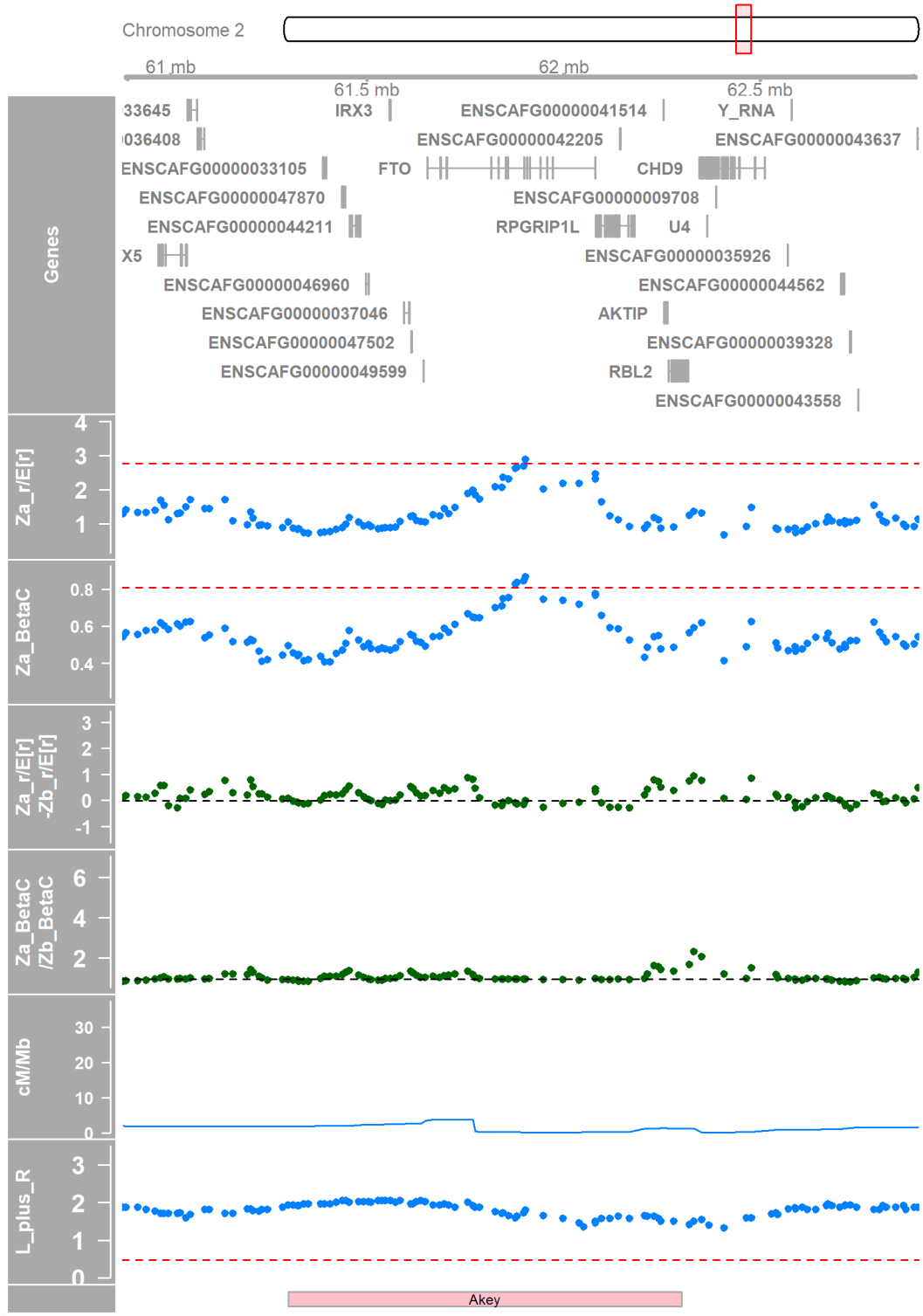
Appendix A

Chr	bp	Z_α	Dist	$Z_\alpha^{r^2/E[r^2]}$	Dist	$Z_\alpha^{BetaCDF}$	Dist	A1	A2	MAF	VEP consequences	VEP Genes	ChIPpeakAnno Genes	Overlap
31	15,074,189	0.36	0.9931	2.42	0.9969	0.82	0.9993	A	G	11%	intergenic_variant	-	*U6*	
32	24,657,487	0.35	0.9923	2.44	0.9971	0.81	0.9991	A	C	46%	intergenic_variant	-	*ENSCAFG00000045443*	
32	25,070,561	0.29	0.9836	2.47	0.9974	0.81	0.9991	A	G	45%	downstream_gene_variant	*ENSCAFG00000010861*	*ENSCAFG00000010861*	
34	1,427,518	0.55	0.9998	2.43	0.9969	0.83	0.9995	G	A	25%	upstream_gene_variant	*ENSCAFG00000048743*	*ENSCAFG00000048743*	
37	4,766,480	0.36	0.9936	2.29	0.9950	0.83	0.9995	G	A	11%	downstream_gene_variant, intron_variant, non_coding_transcript_variant	*ENSCAFG00000044056, ENSCAFG00000050002*	*ENSCAFG00000050002*	
37	14,949,623	0.37	0.9946	2.86	0.9993	0.82	0.9993	G	A	12%	intron_variant, non_coding_transcript_variant	*ENSCAFG00000049414*	*ENSCAFG00000049414*	

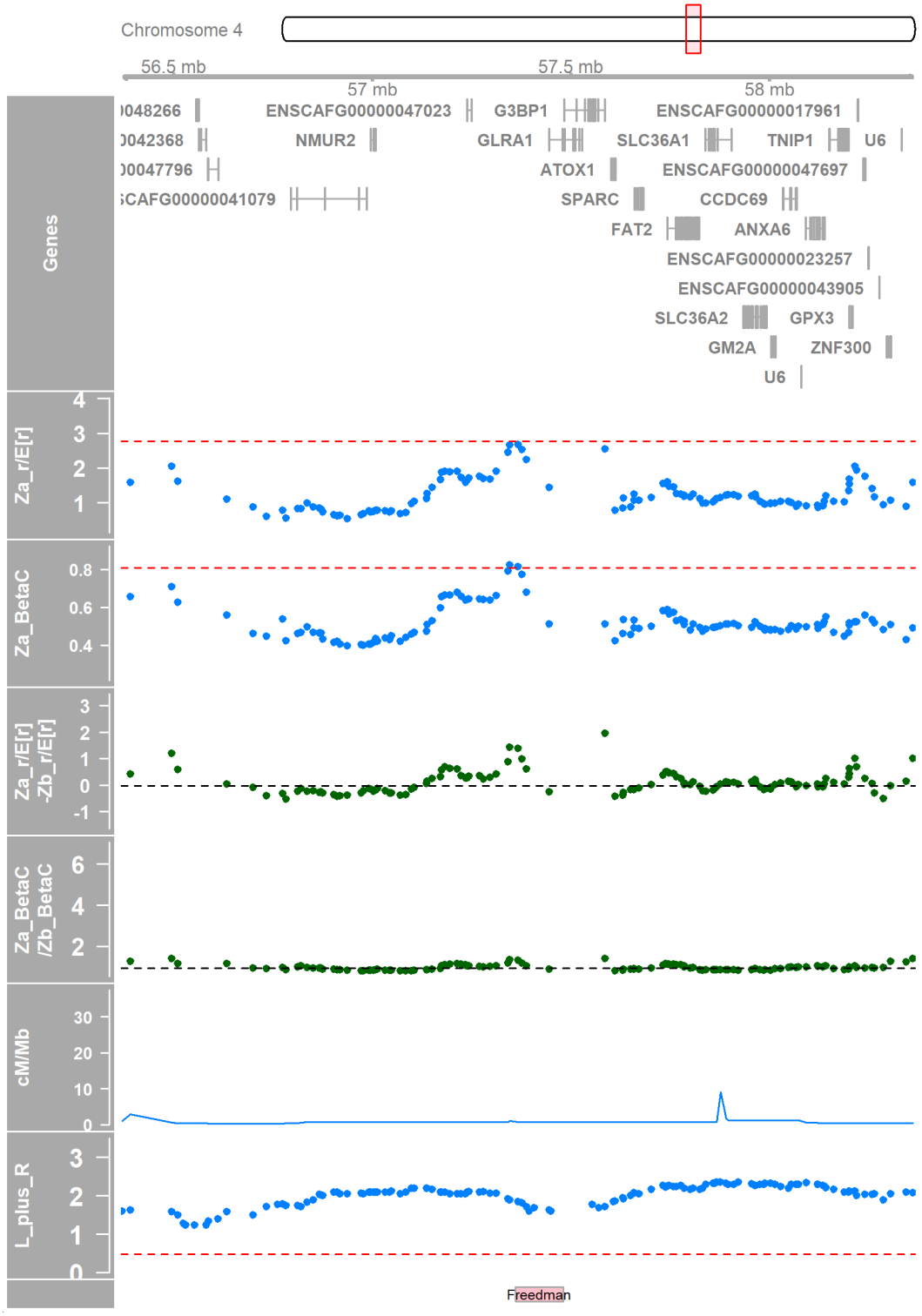
A.4.4 Candidate regions: previously published

These figures show the candidate regions from this study. A candidate region is a region within which at least one SNP is in the top 0.1% of either of the statistics $Z_{\alpha}^{r^2/E[r^2]}$ or $Z_{\alpha}^{BetaCDF}$. They also need to have been outliers in the second cluster of dogs generated in the PCA for one of these statistics. If the region was also overlapped by a region published in one of the studies listed in Table 8-4 then it will be in this section, otherwise it is considered novel for this study and is in section A.4.5. An extra 1 Mb either side of the region is included in the figure to show the genes in the vicinity. Figures were generated as described in section 8.2.8.

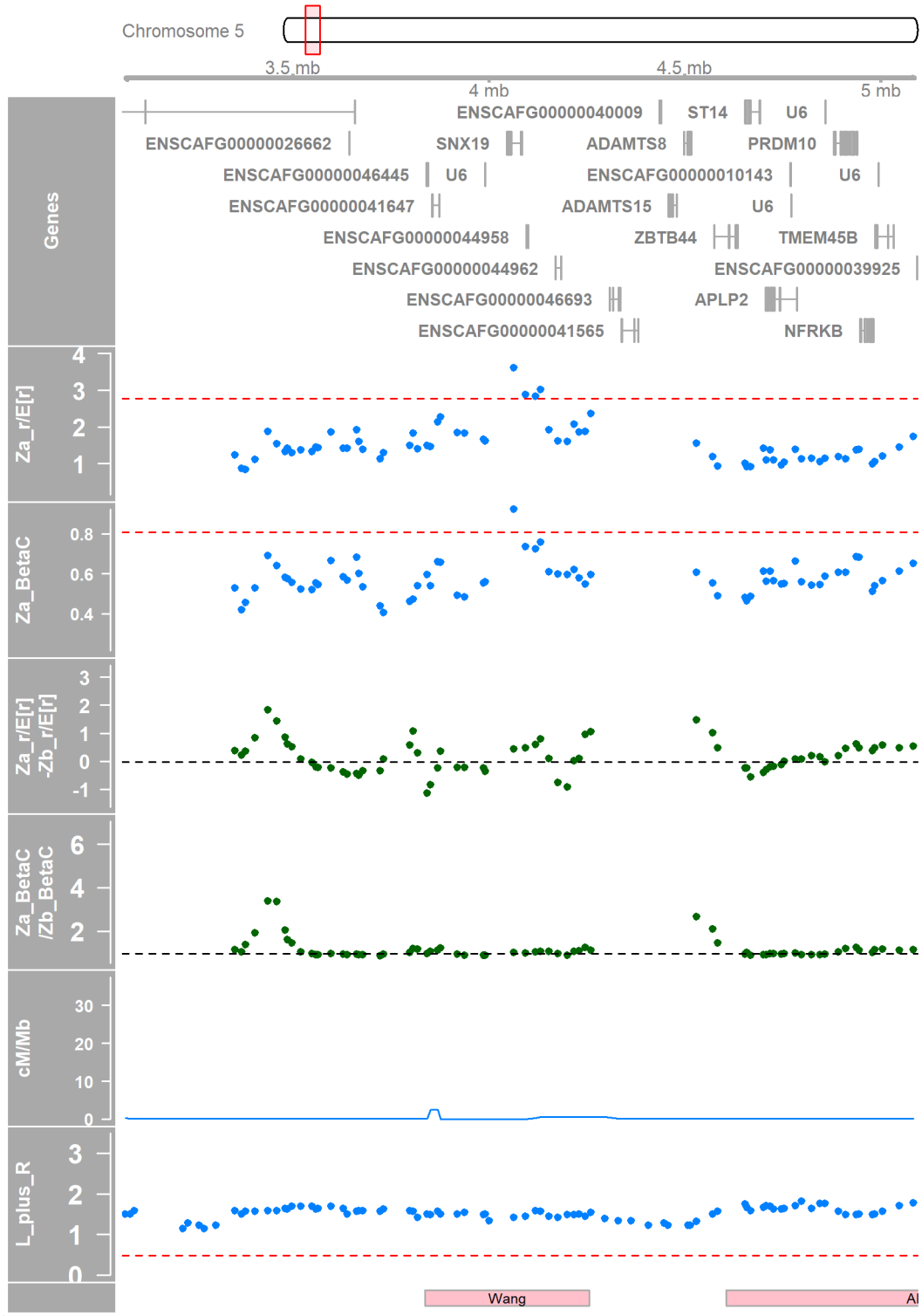
From the top: An ideogram of the chromosome highlighting the region; the genome axis track; the genes in the region from Ensembl; plot of $Z_{\alpha}^{r^2/E[r^2]}$ with a red line indicating the top 0.1% of values; plot of $Z_{\alpha}^{BetaCDF}$ with a red line indicating the top 0.1% of values; plot of $Z_{\alpha}^{r^2/E[r^2]} - Z_{\beta}^{r^2/E[r^2]}$ with a black line indicating the median; plot of $\frac{Z_{\alpha}^{BetaCDF}}{Z_{\beta}^{BetaCDF}}$ with a black line indicating the median; the recombination rate in cM/Mb; plot of $\log_{10}\left(\binom{|L|}{2} + \binom{|R|}{2}\right)$ with a red line showing the bottom 0.1%; previously published candidate regions.



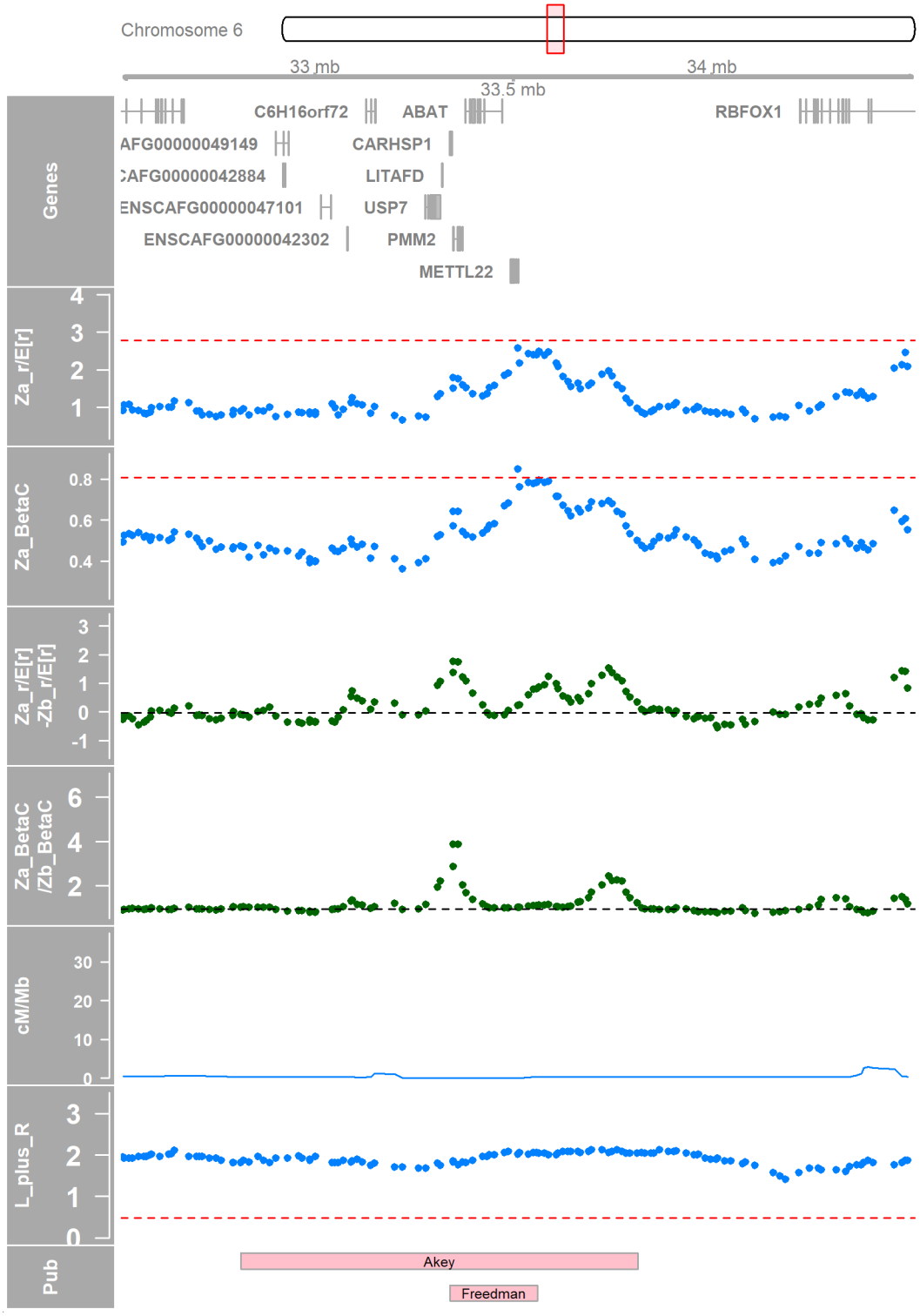
Appendix Figure 11 Candidate region 2:61876498-61901702



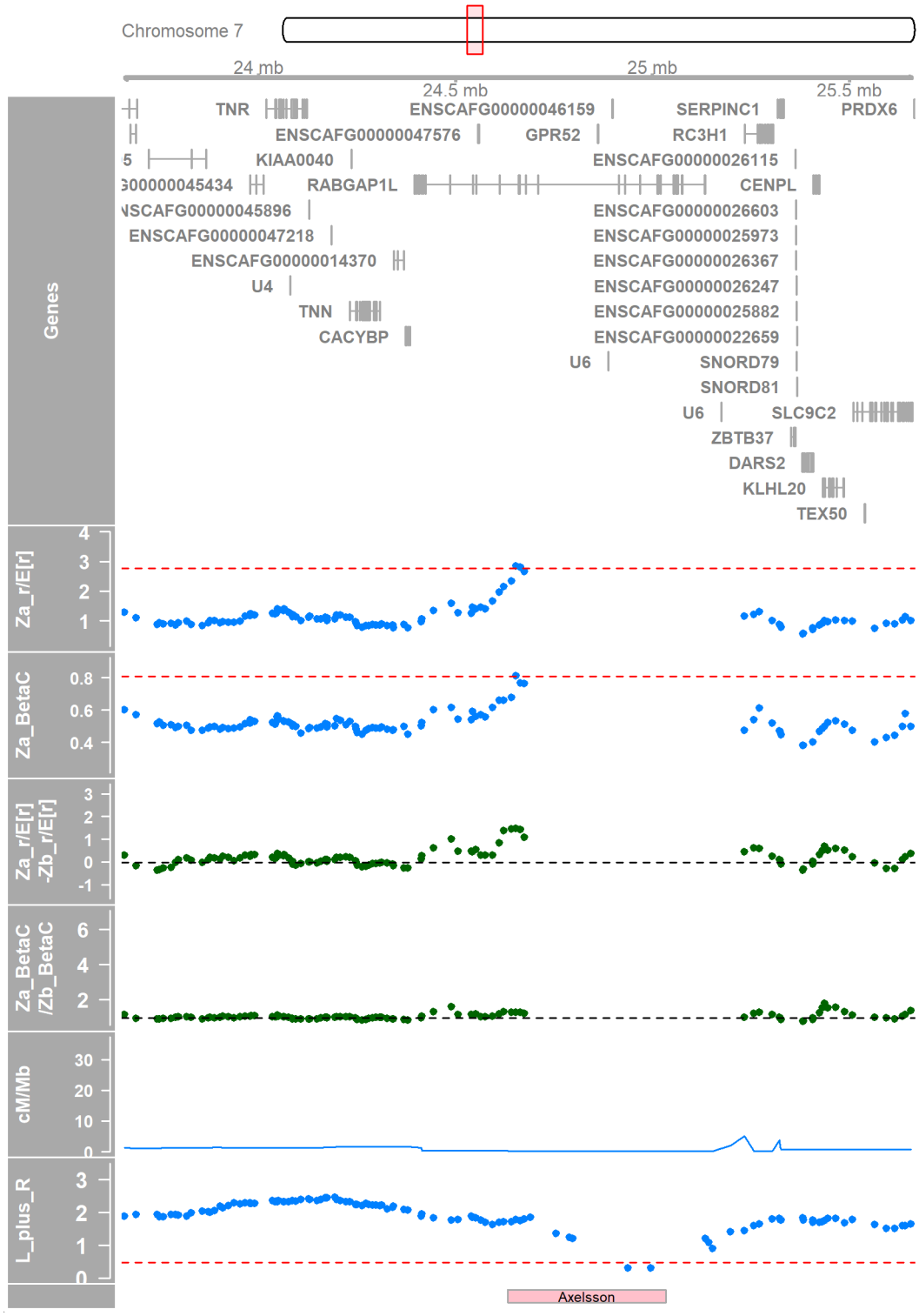
Appendix Figure 12 Candidate region 4:57366377



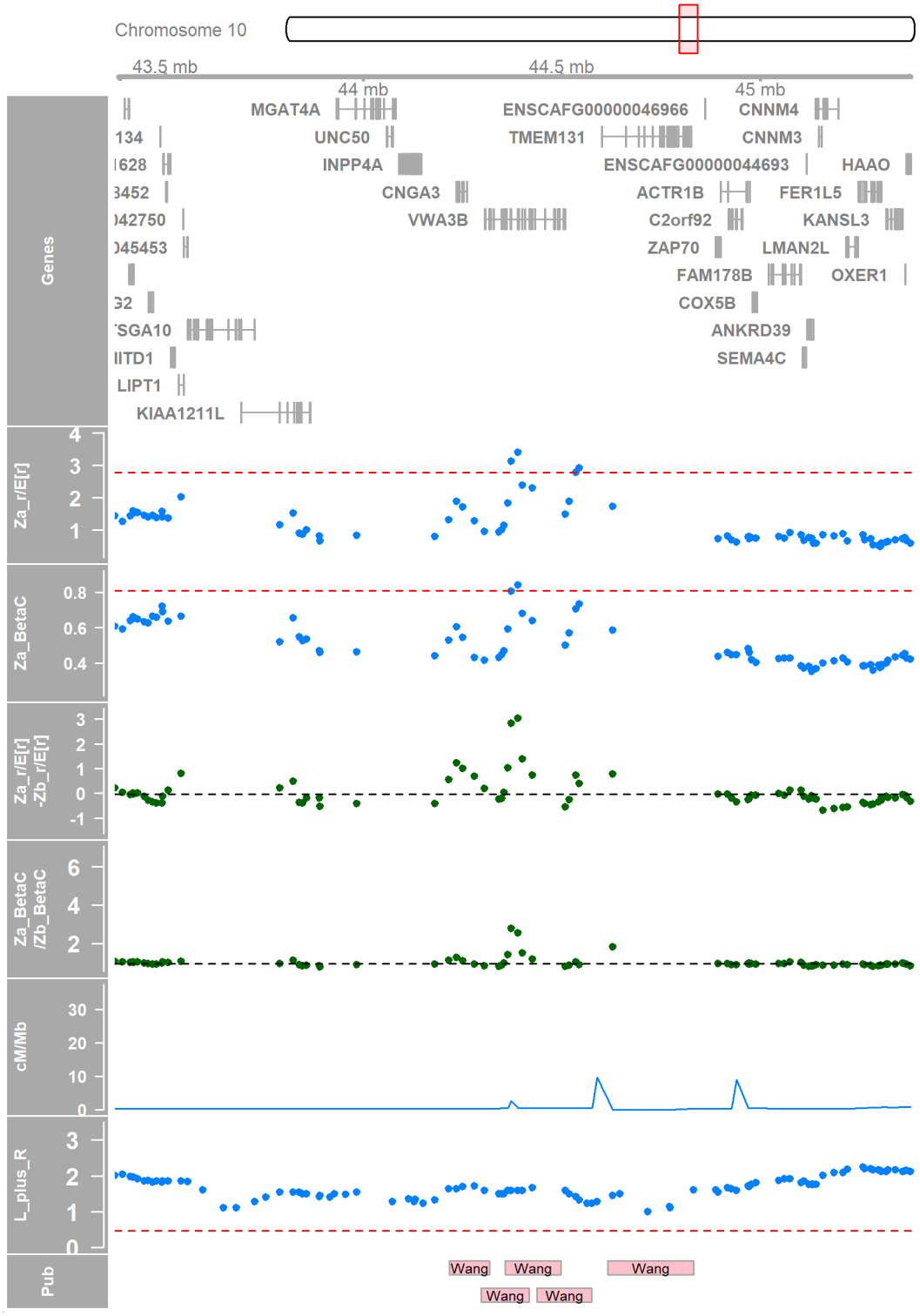
Appendix Figure 13 Candidate region 5:4064061-4093514



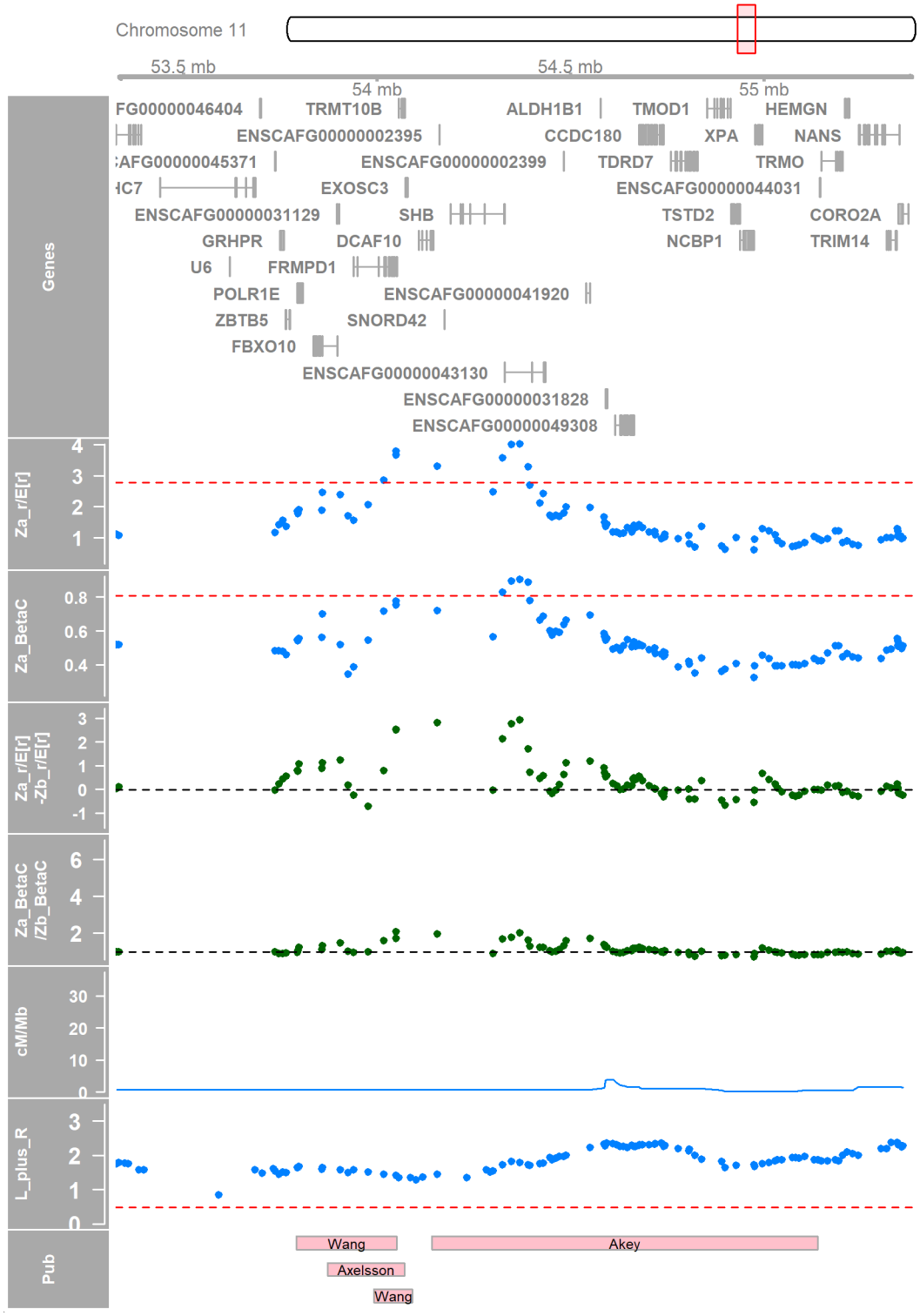
Appendix Figure 14 Candidate region 6:33510473



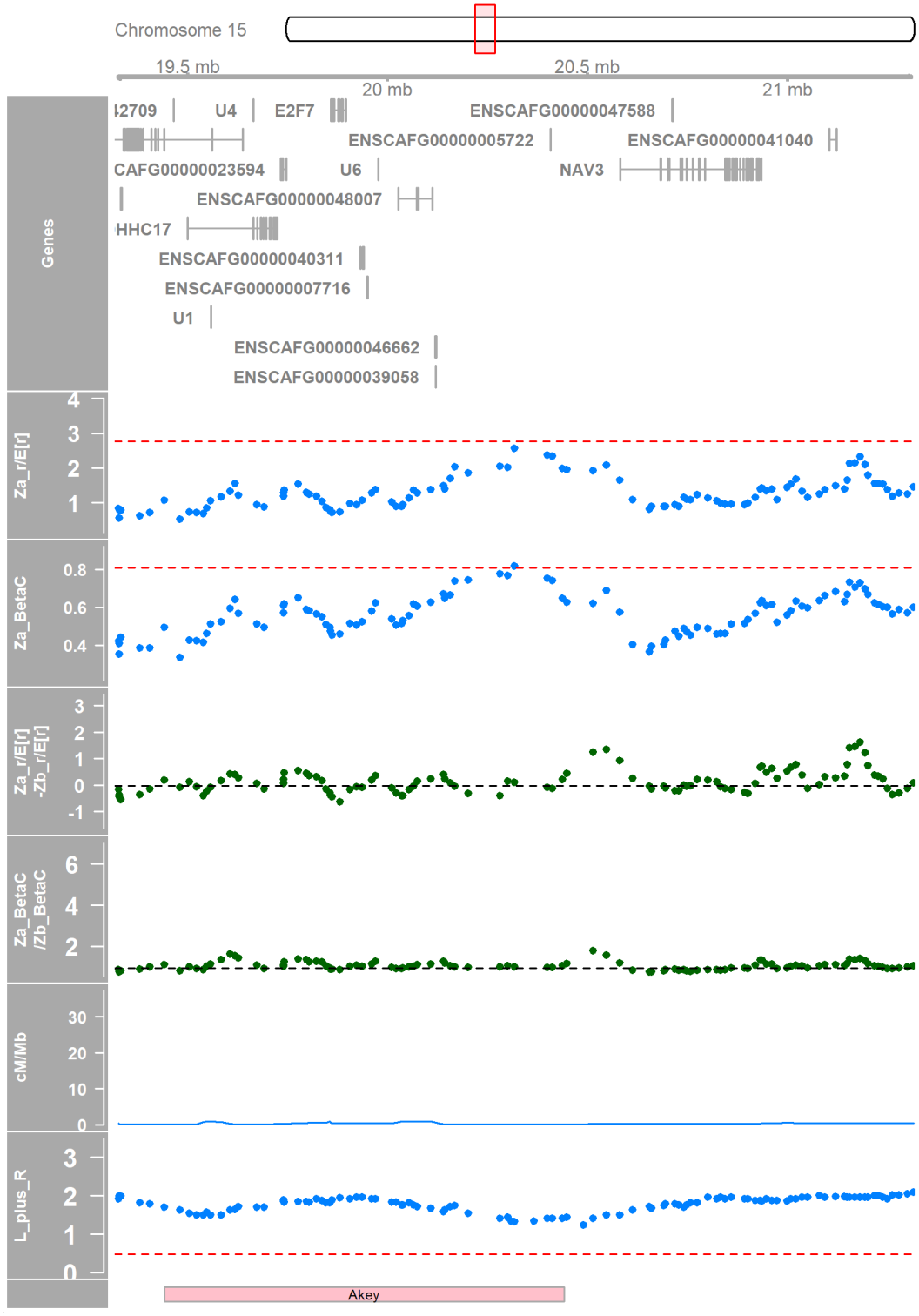
Appendix Figure 15 Candidate region 7:24652821-24664438



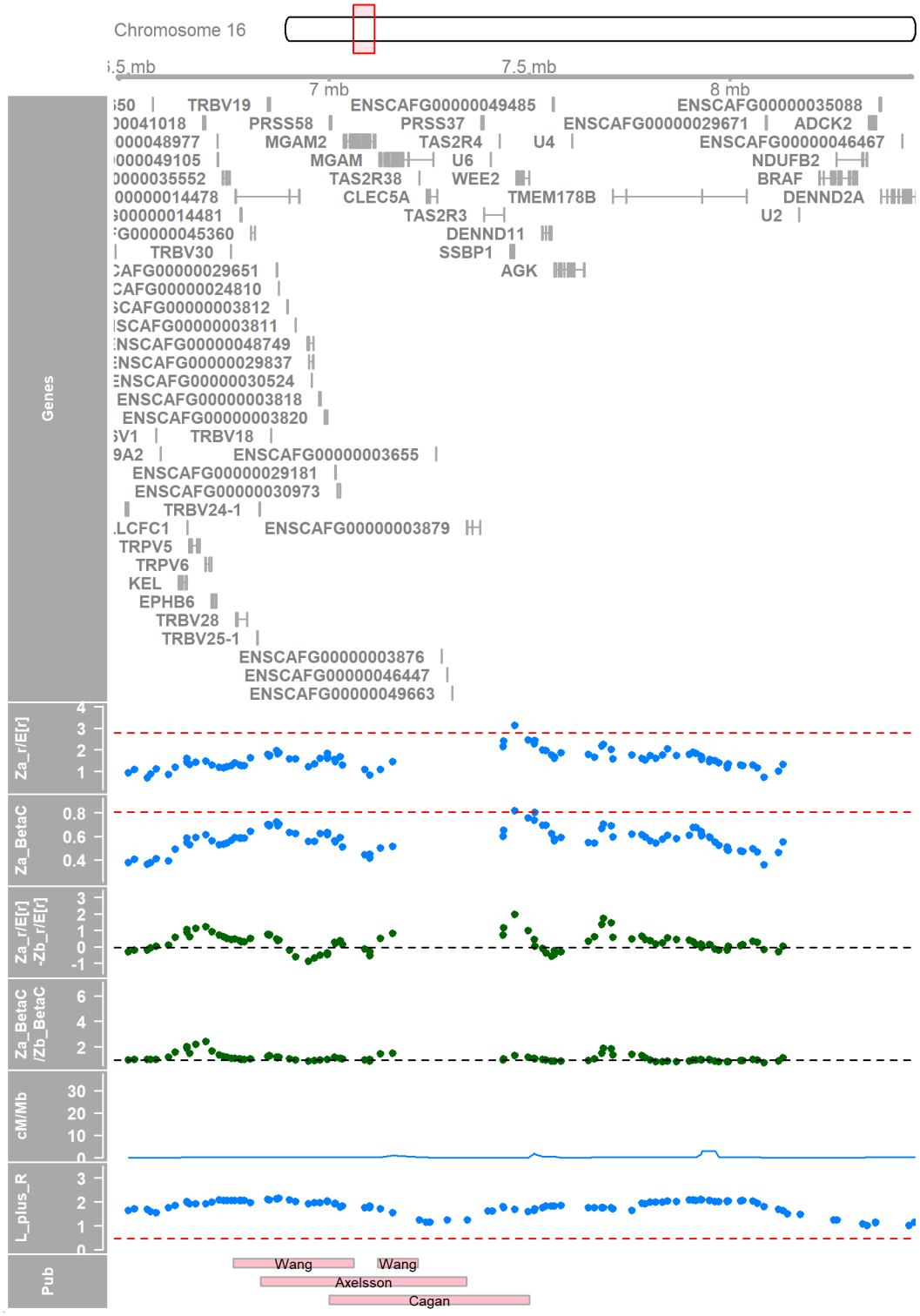
Appendix Figure 16 Candidate region 10:44372549-44388924



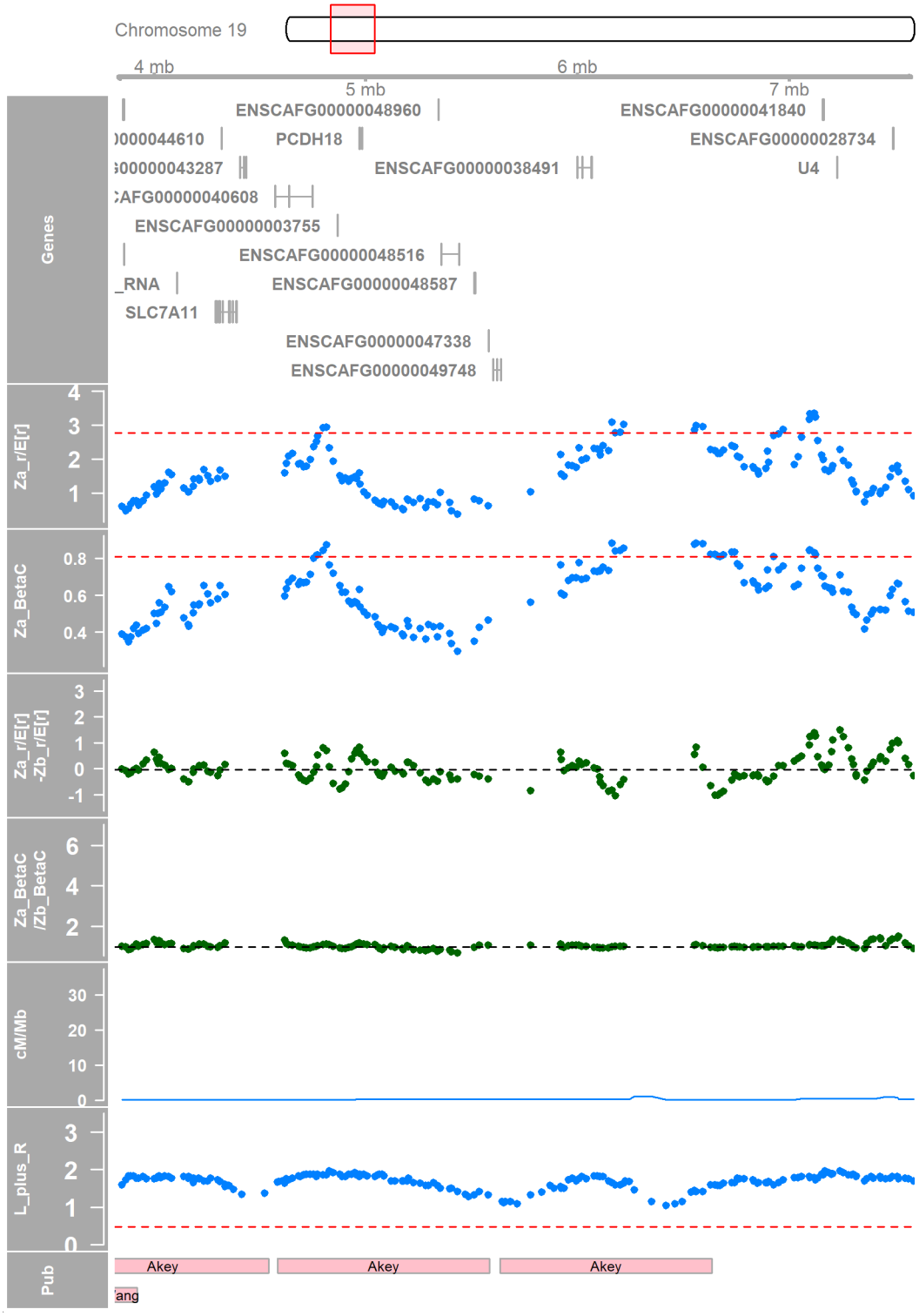
Appendix Figure 17 Candidate region 11:54324689-54391443



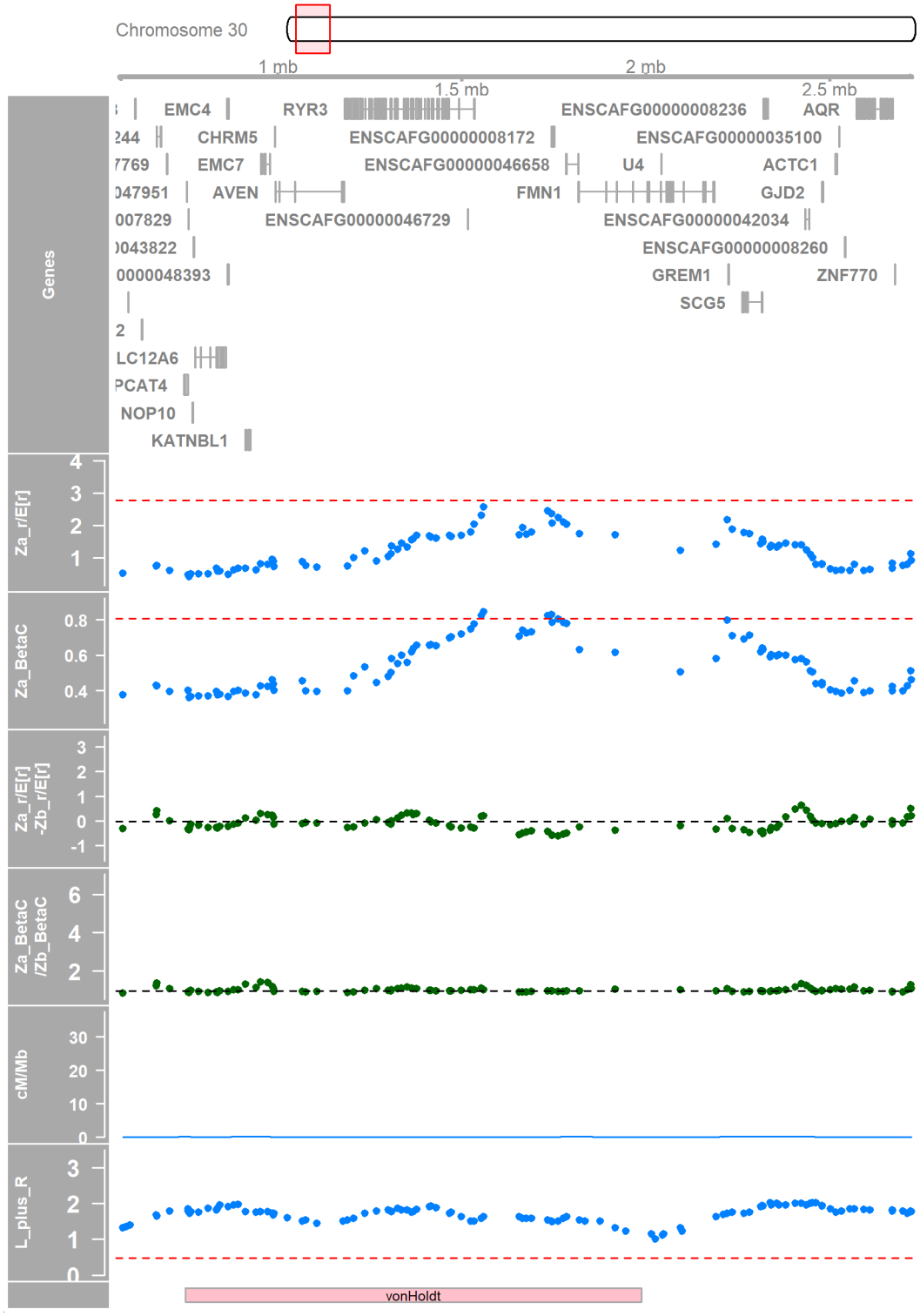
Appendix Figure 18 Candidate region 15:20317533



Appendix Figure 19 Candidate region 16:7462818



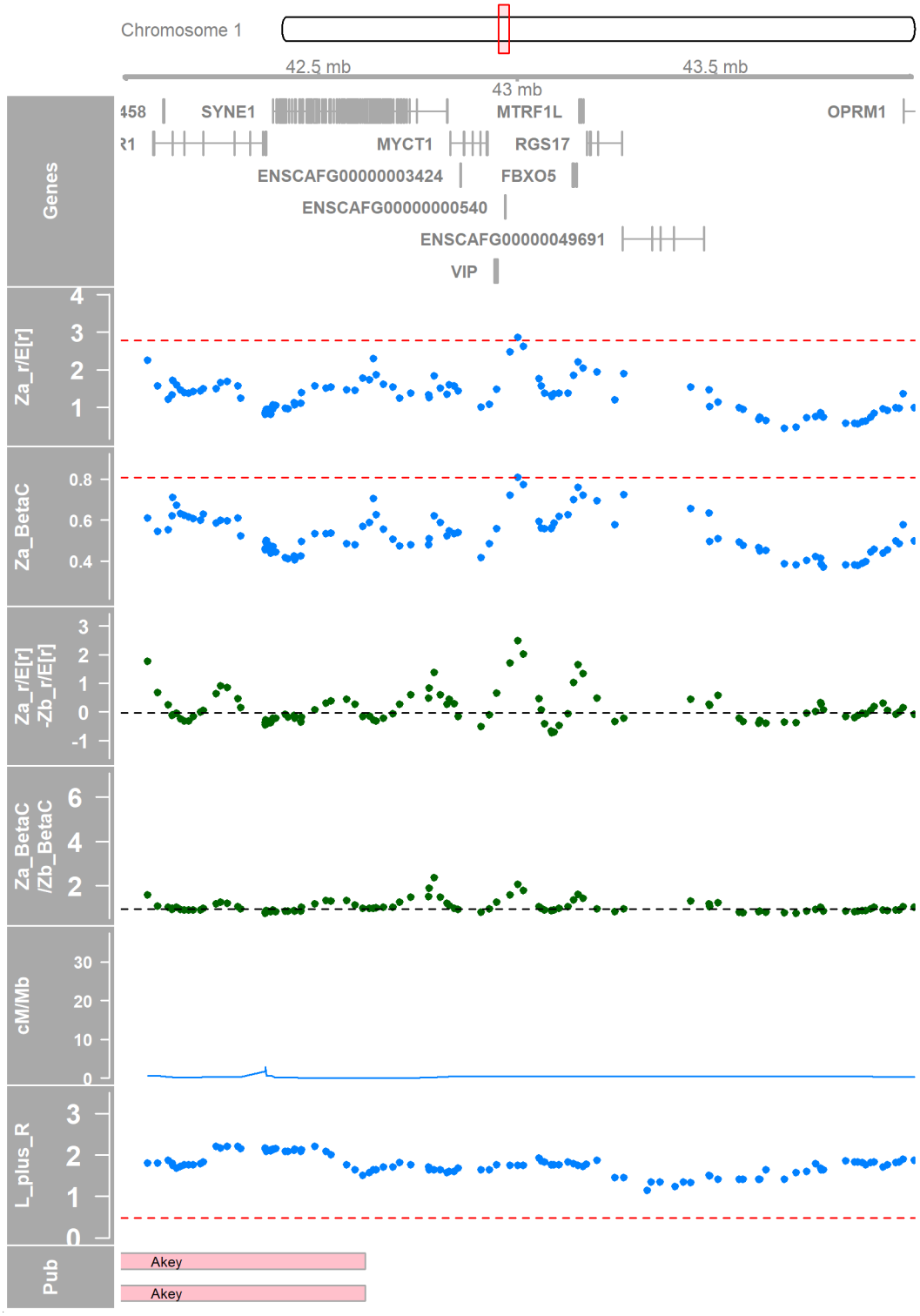
Appendix Figure 20 Candidate region 19:4813917-6590666



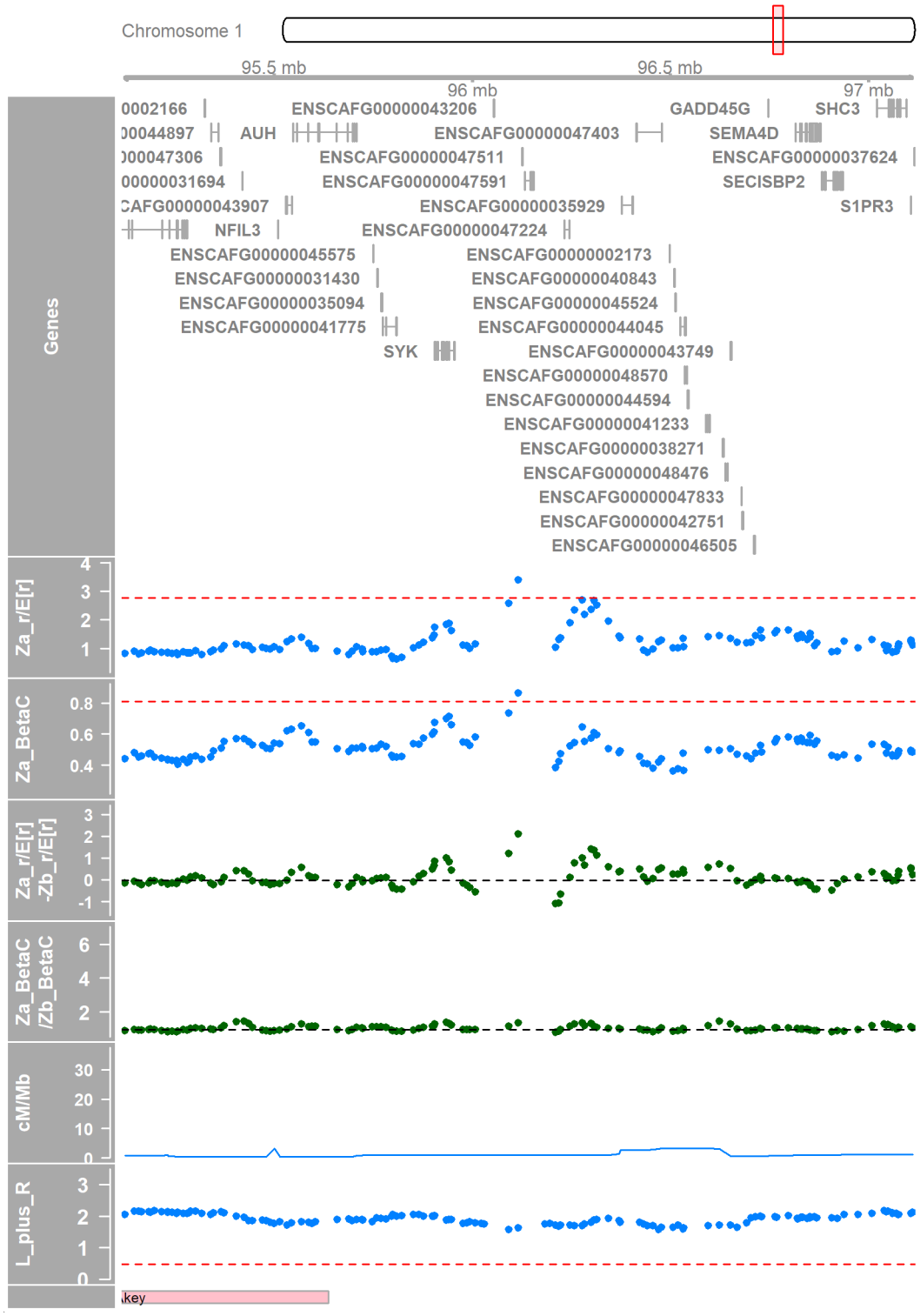
Appendix Figure 21 Candidate region 30:1558195-1732646

A.4.5 Candidate regions: novel

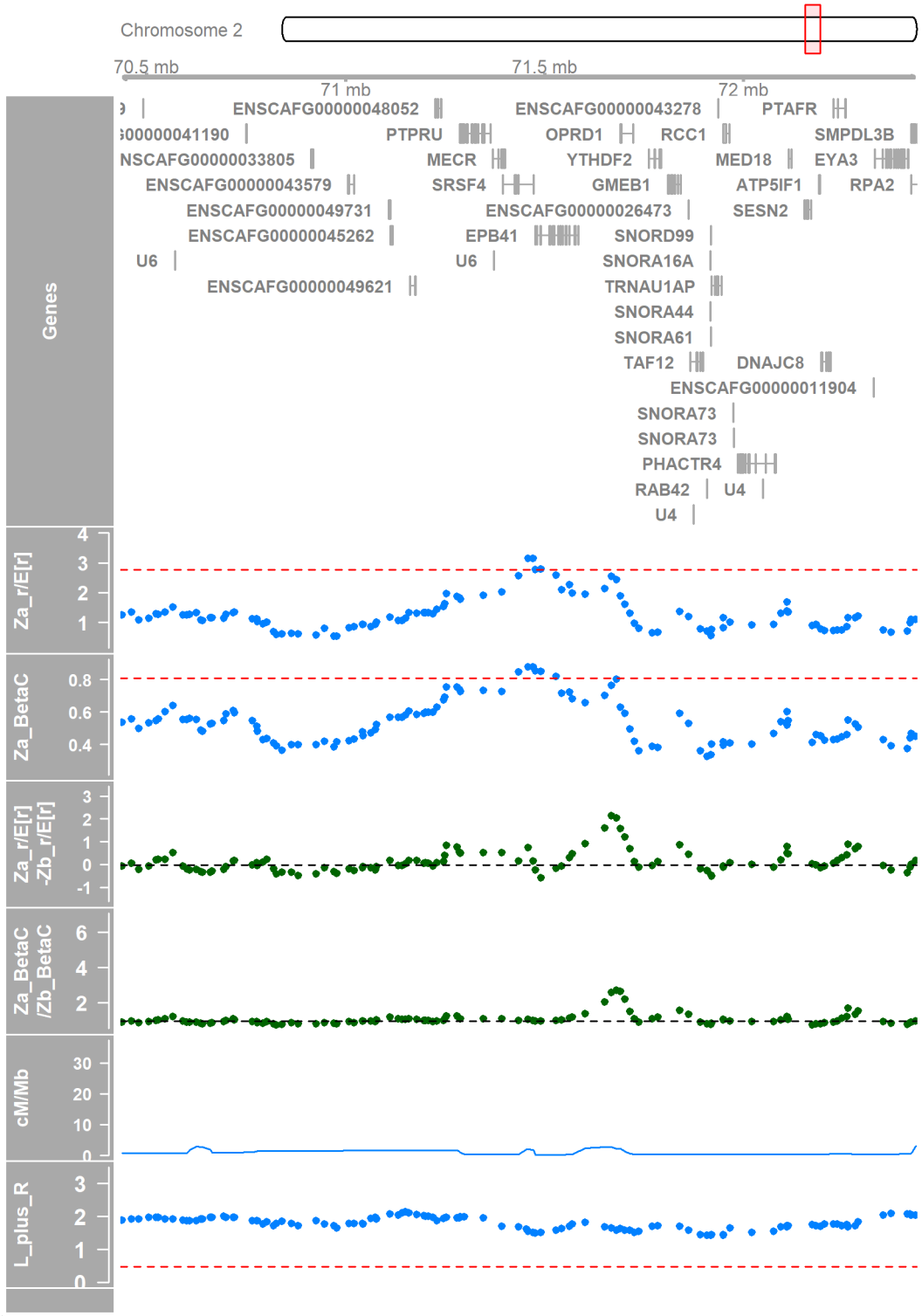
These graphs are the same as described in the previous section A.4.4, but where the candidate region does not overlap a previously published region.



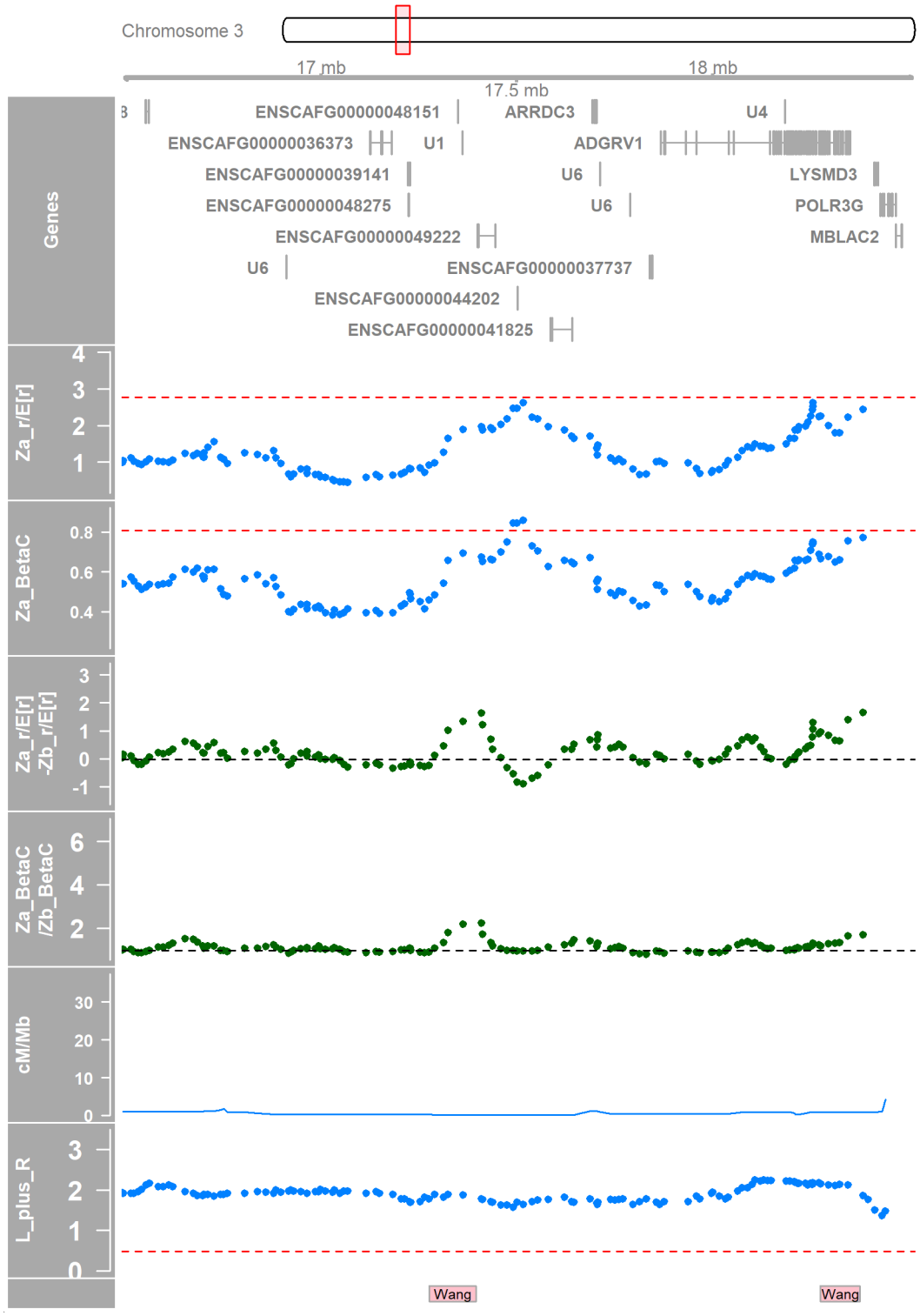
Appendix Figure 22 Candidate region 1:43001368



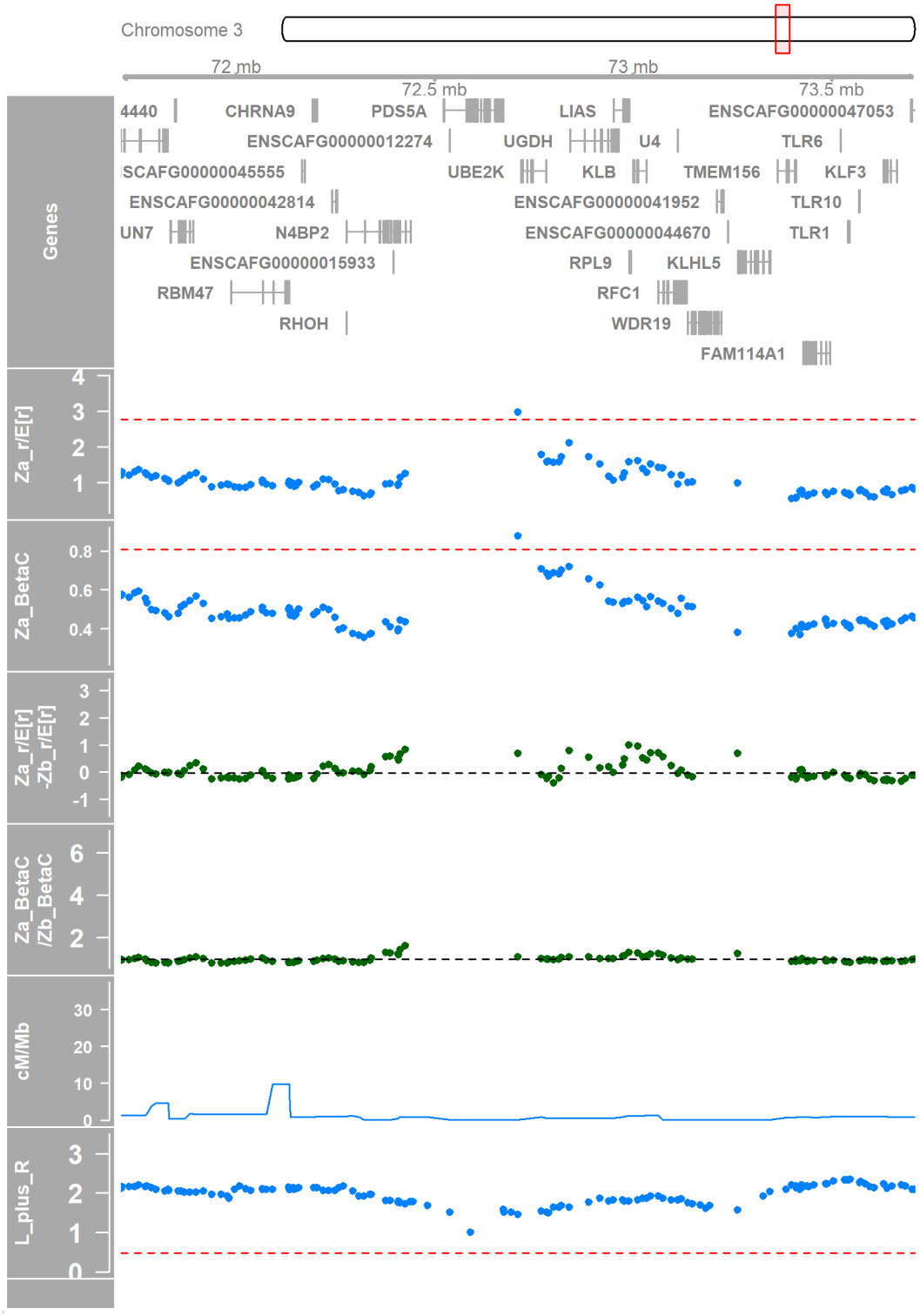
Appendix Figure 23 Candidate region 1:96115461



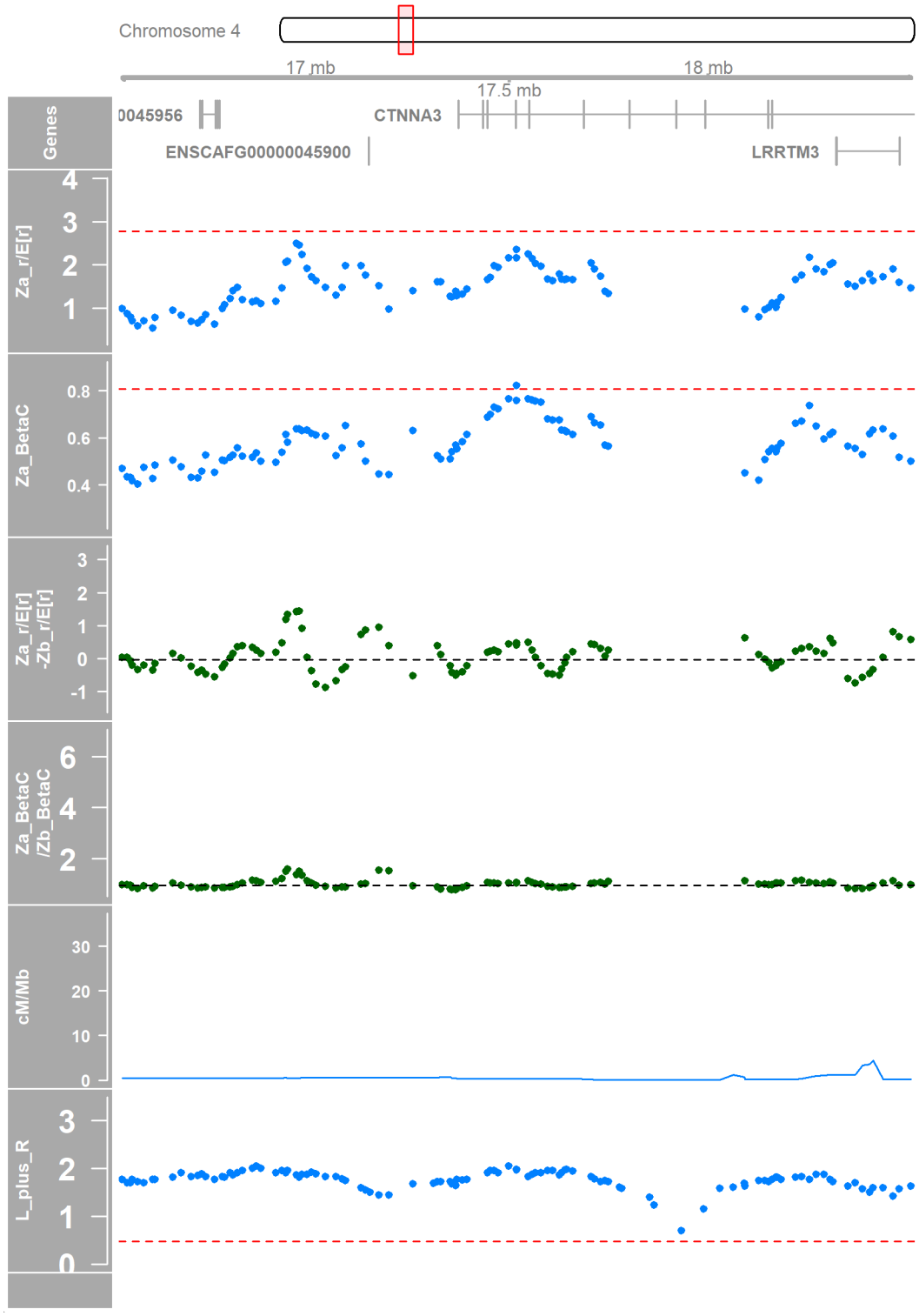
Appendix Figure 24 Candidate region 2:71434345



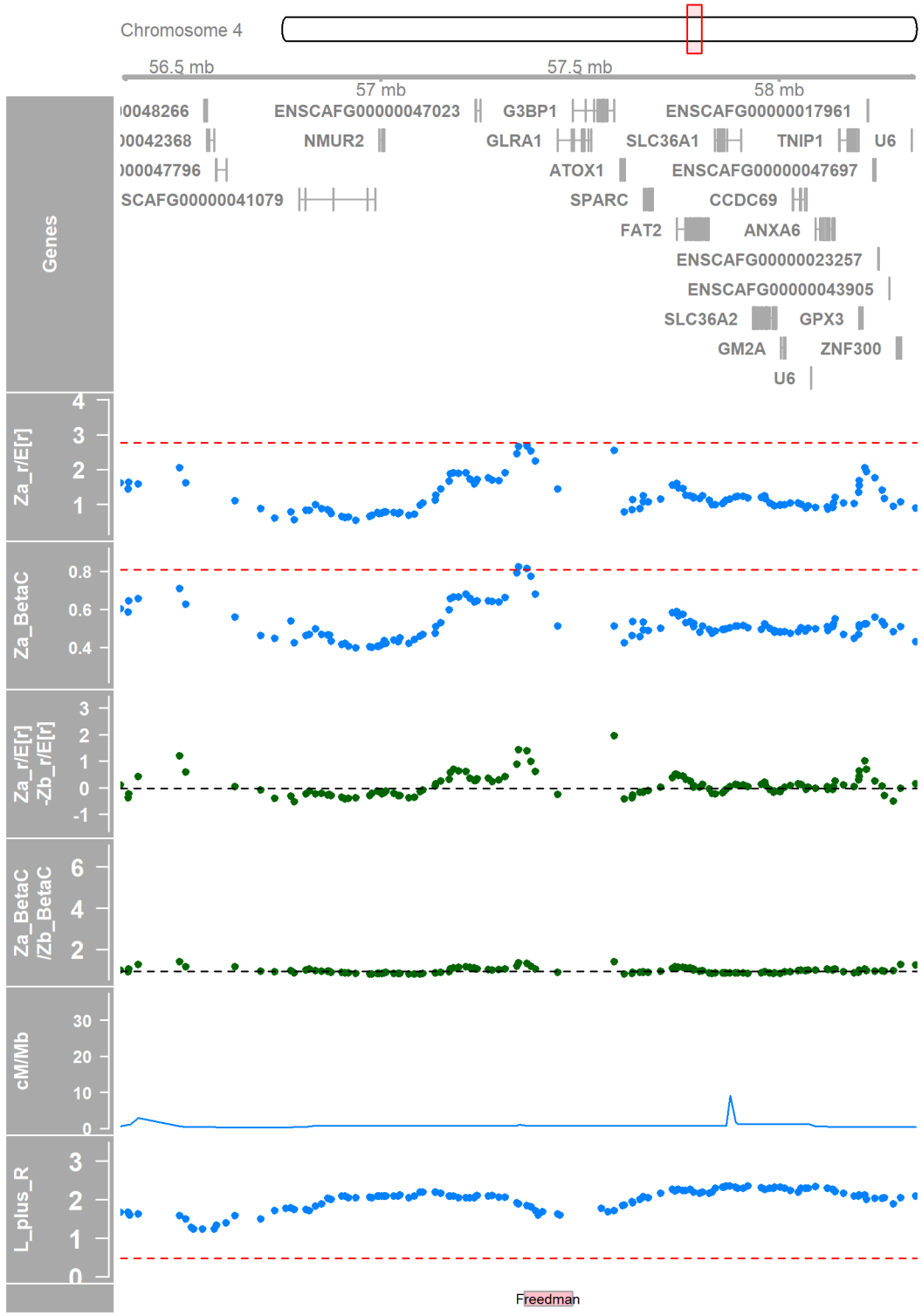
Appendix Figure 25 Candidate region 3:17490492-17516194



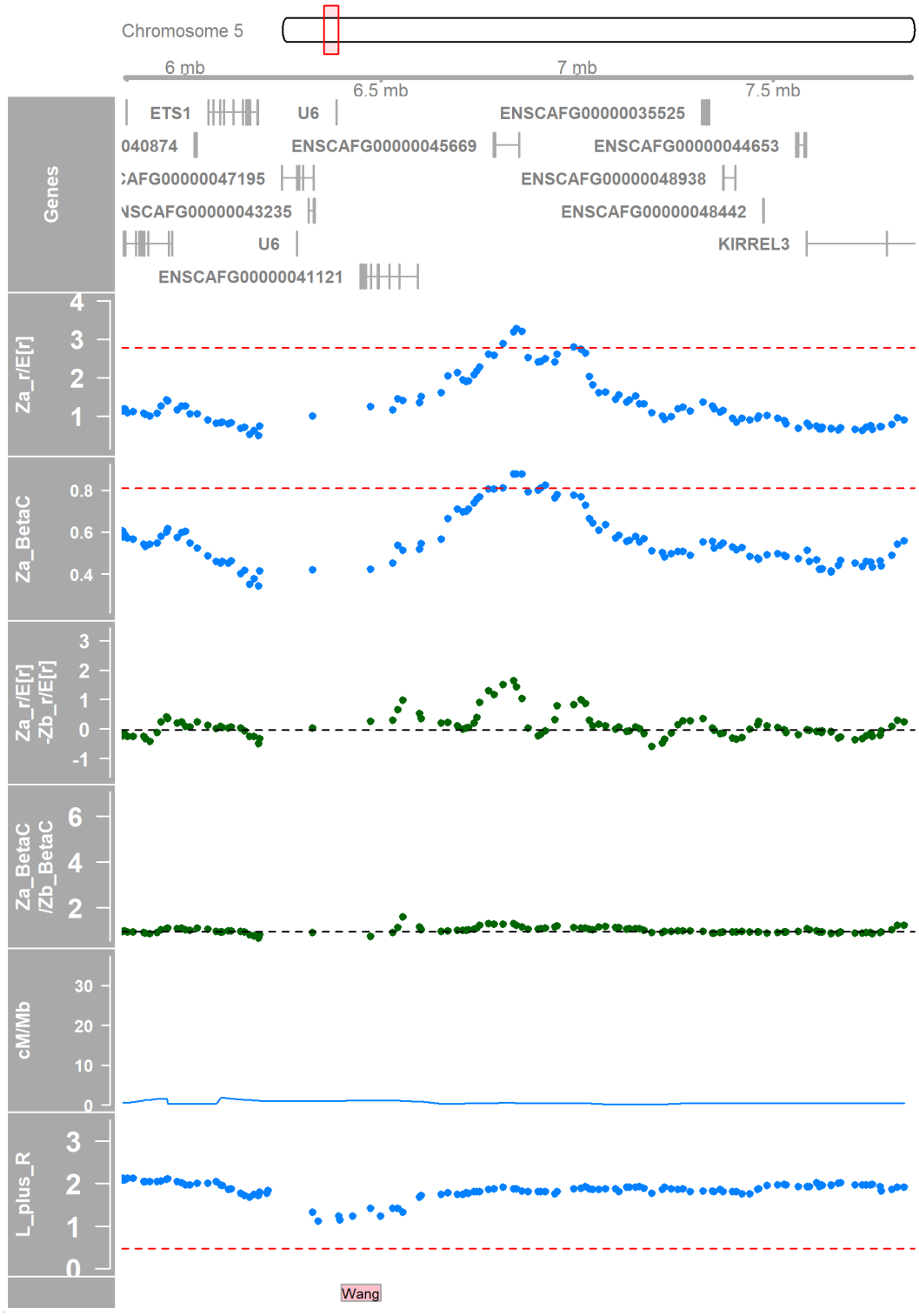
Appendix Figure 26 Candidate region 3:72708942



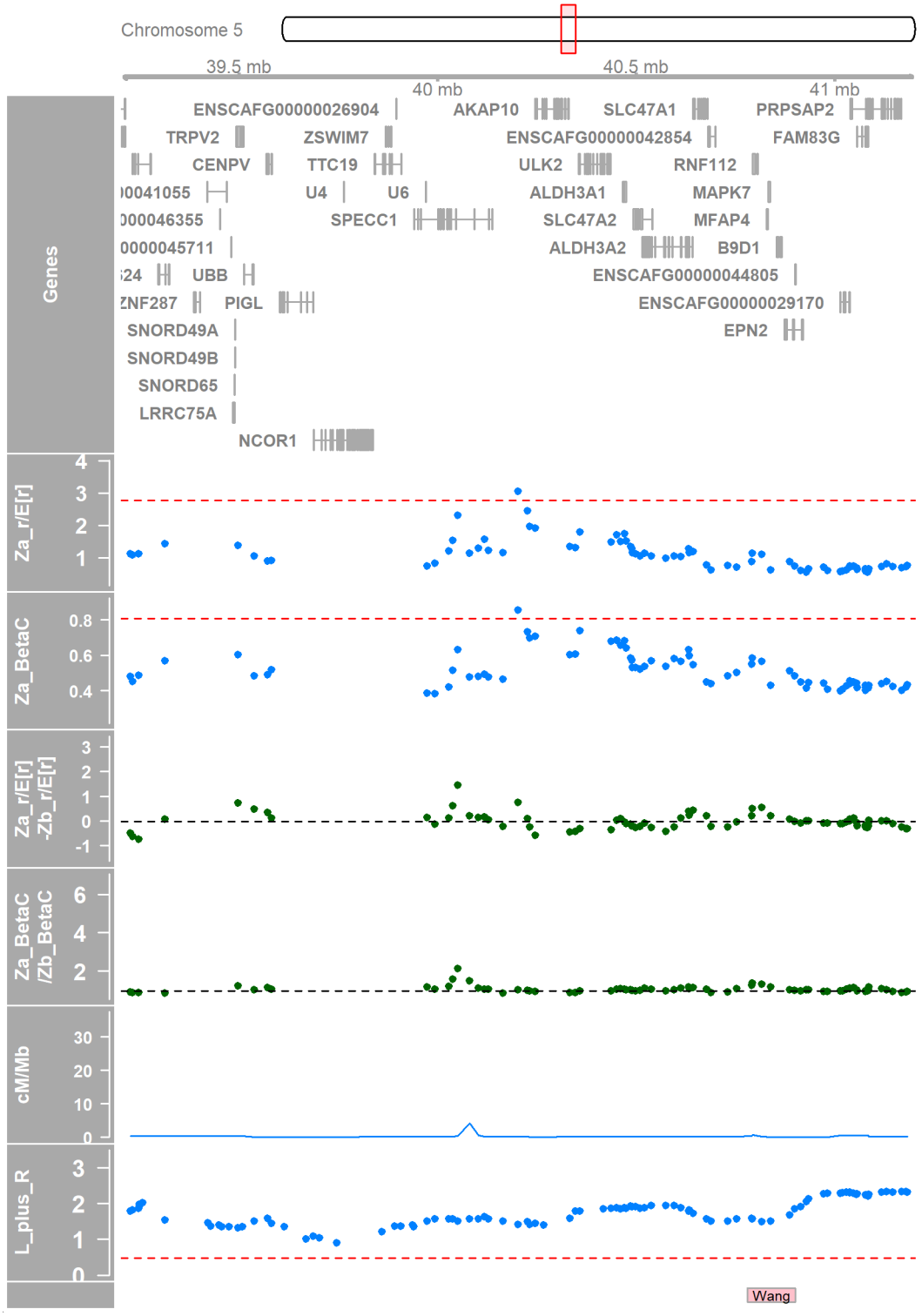
Appendix Figure 27 Candidate region 4:17518453



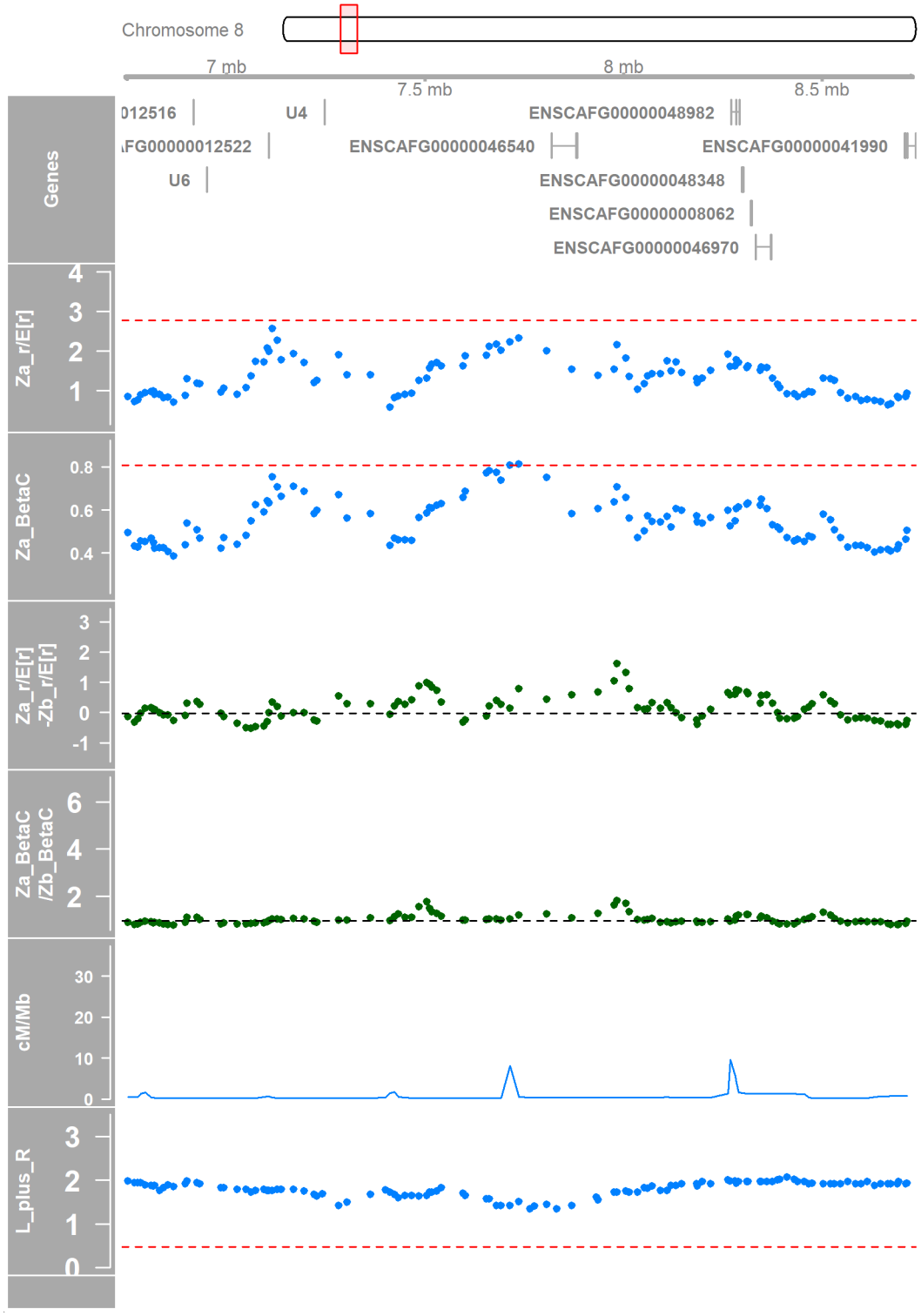
Appendix Figure 28 Candidate region 4:57345395



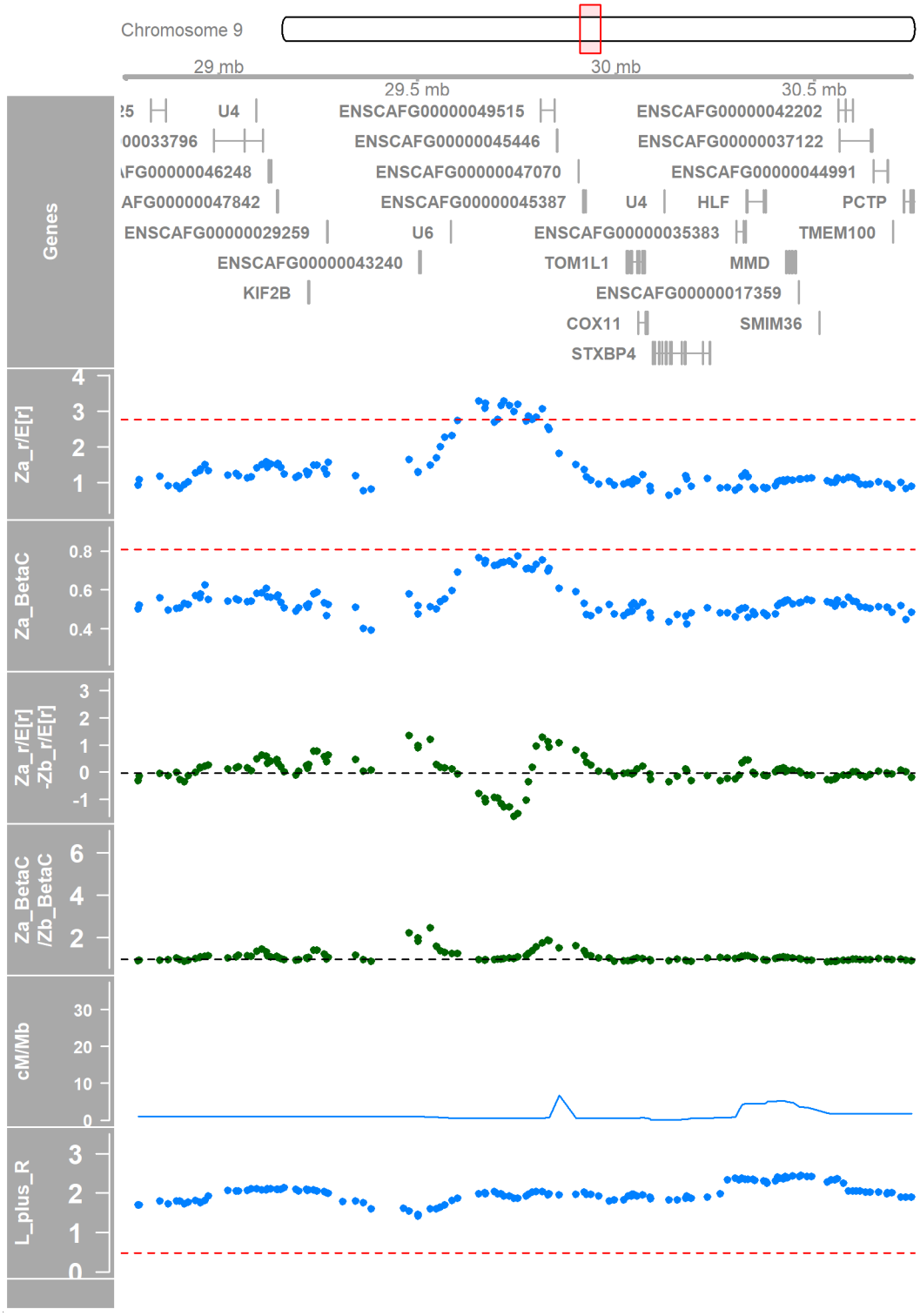
Appendix Figure 29 Candidate region 5:6838932-6859691



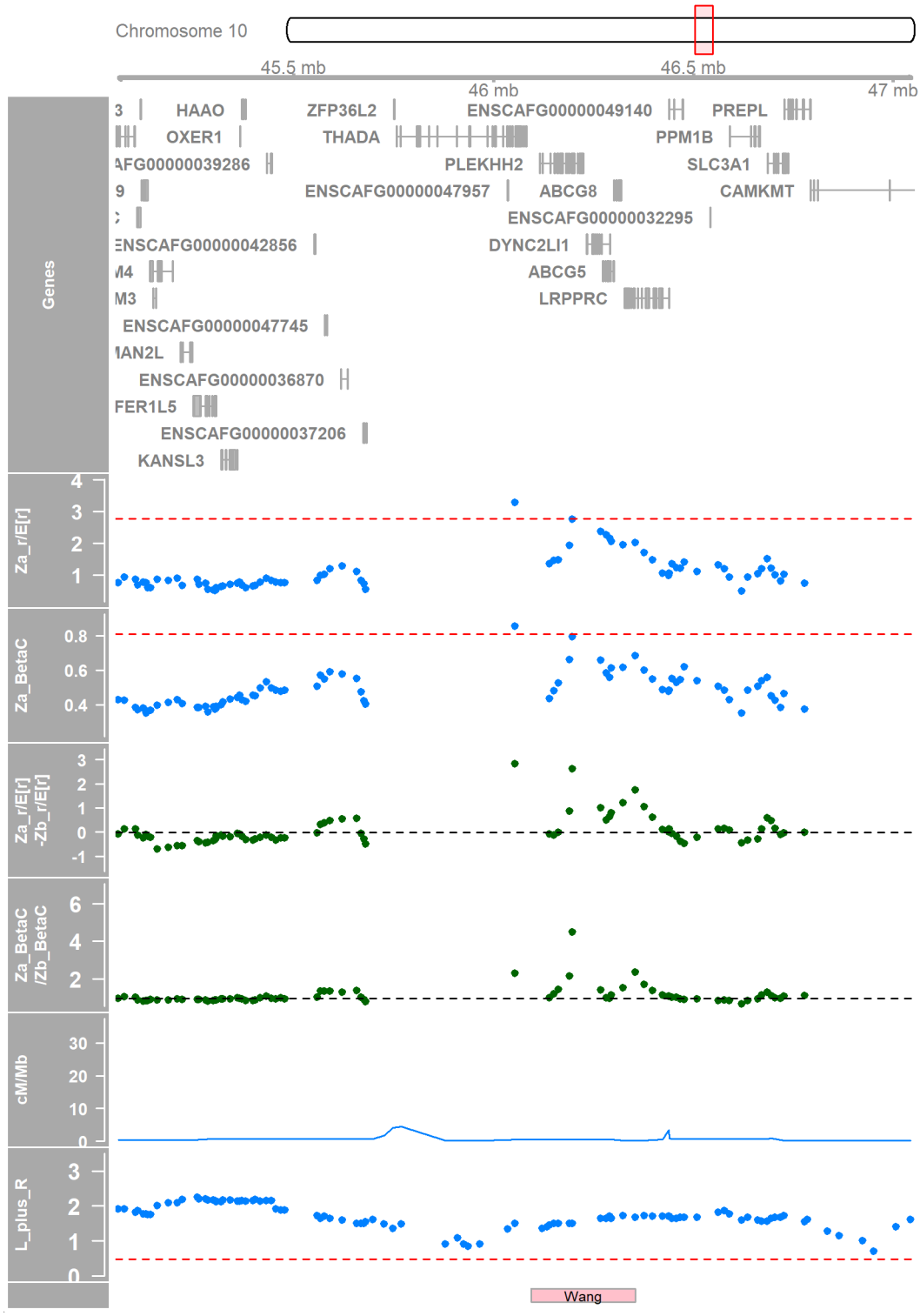
Appendix Figure 30 Candidate region 5:40202215



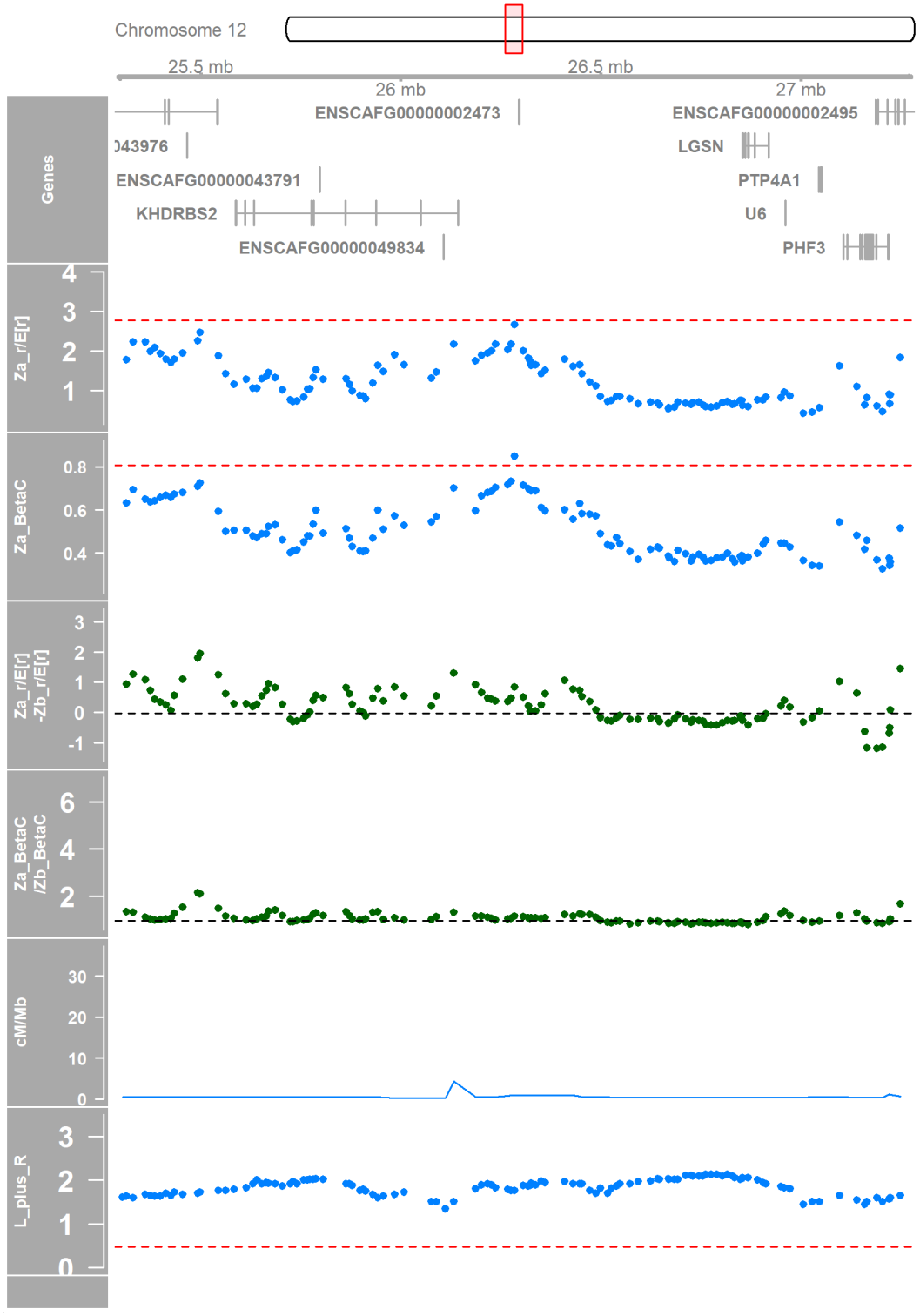
Appendix Figure 31 Candidate region 8:7735497



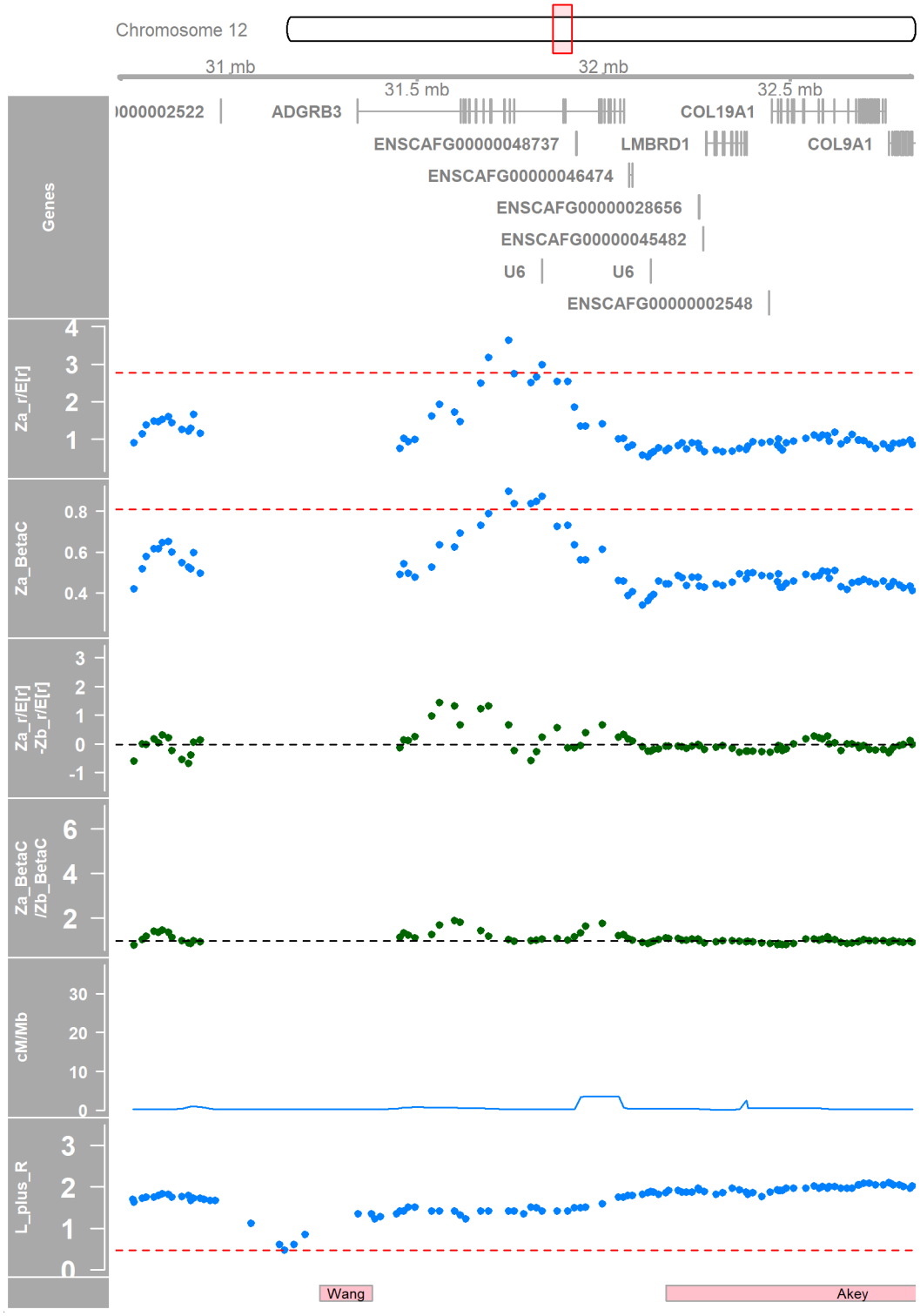
Appendix Figure 32 Candidate region 9:29752455



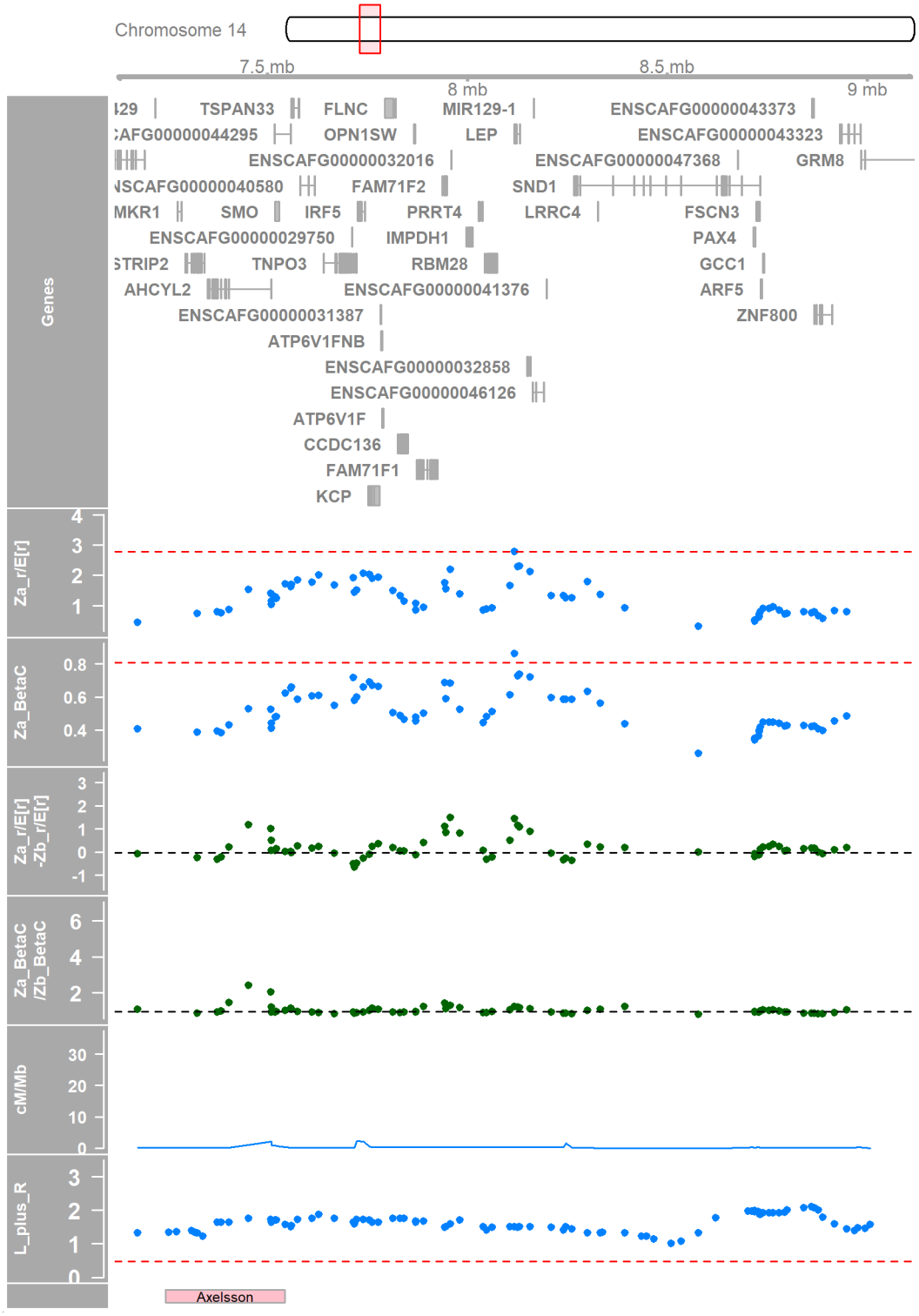
Appendix Figure 33 Candidate region 10:46053118



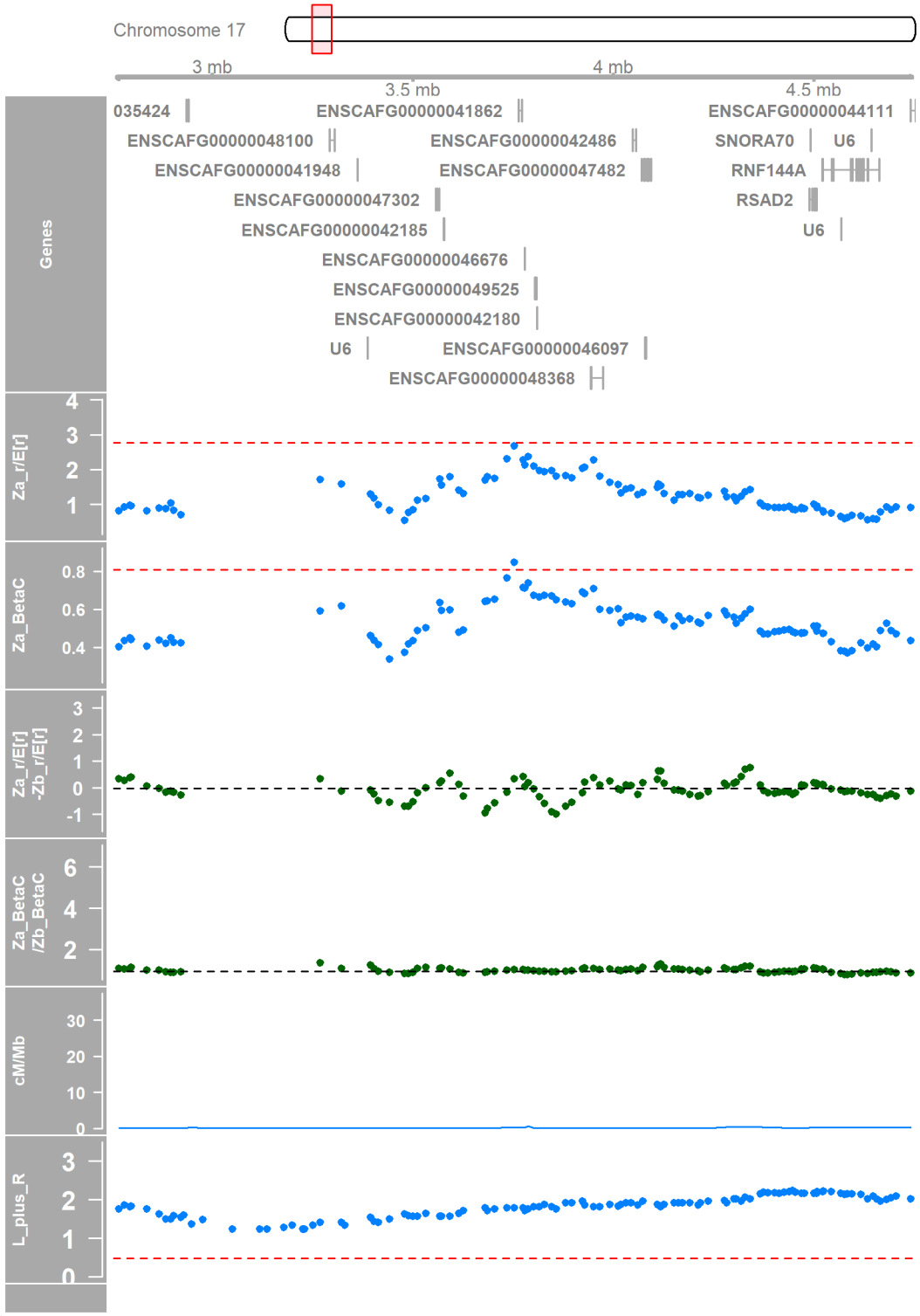
Appendix Figure 34 Candidate region 12:26284264



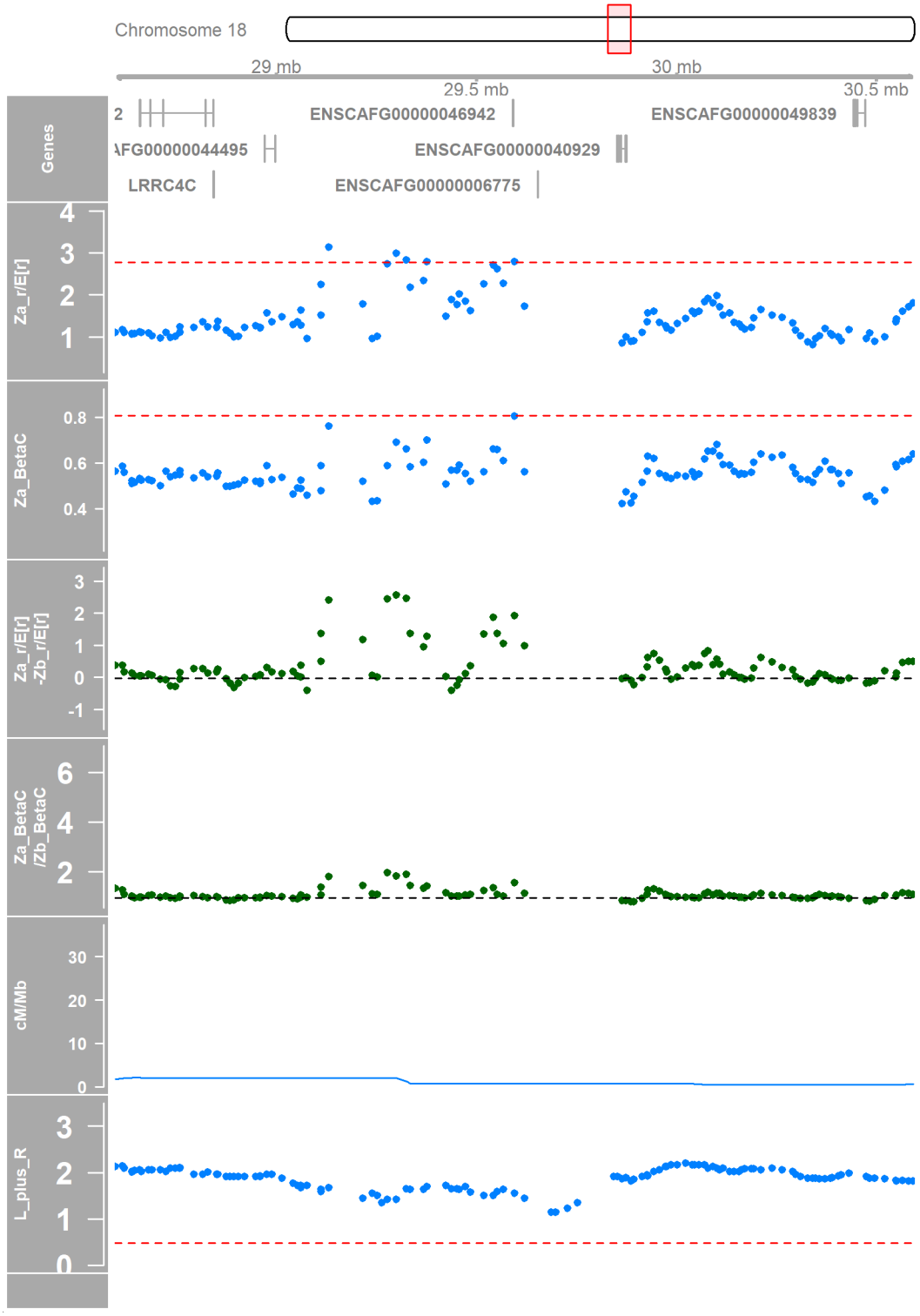
Appendix Figure 35 Candidate region 12:31691990-31835704



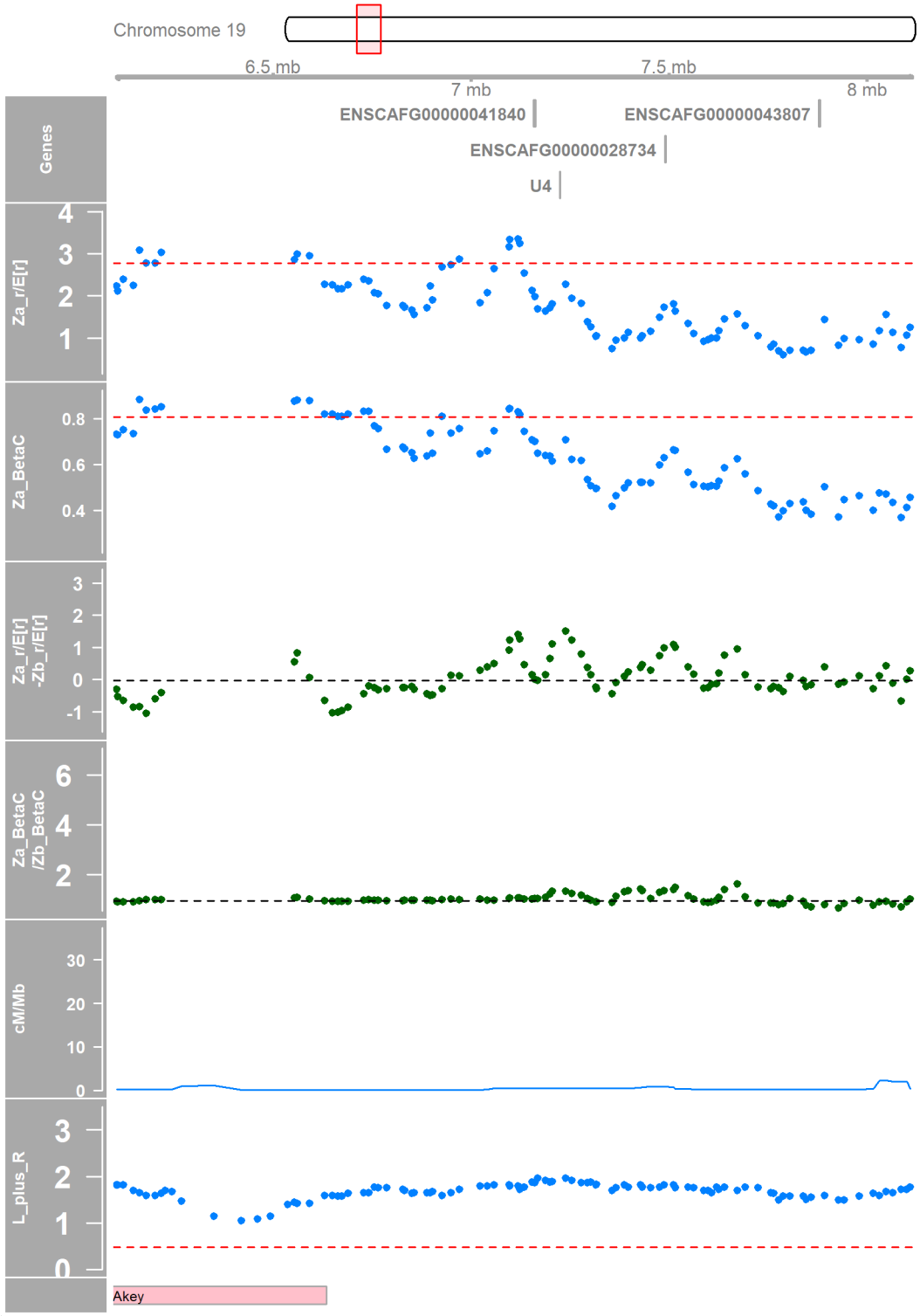
Appendix Figure 36 Candidate region 14:8117811



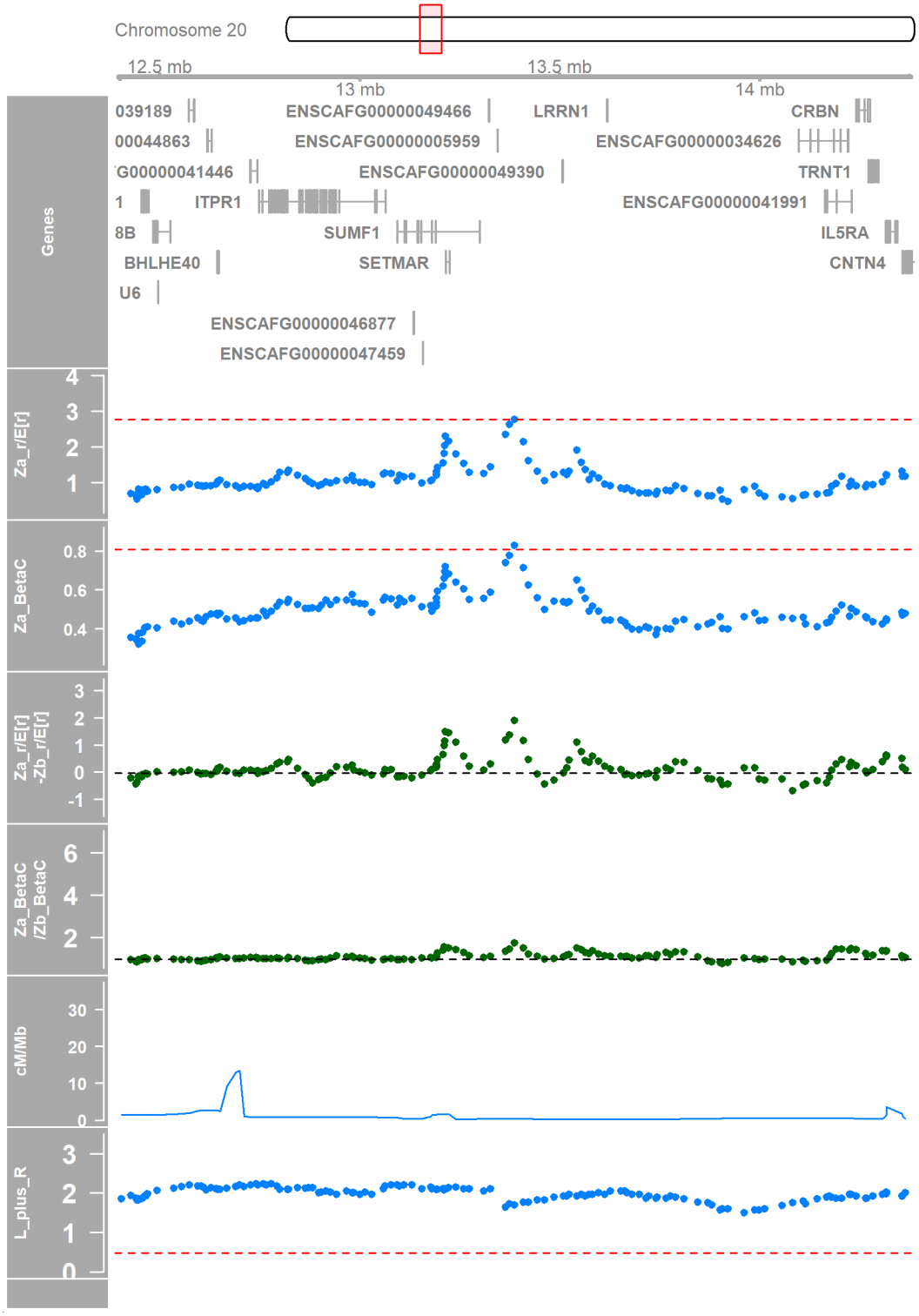
Appendix Figure 37 Candidate region 17:3753156



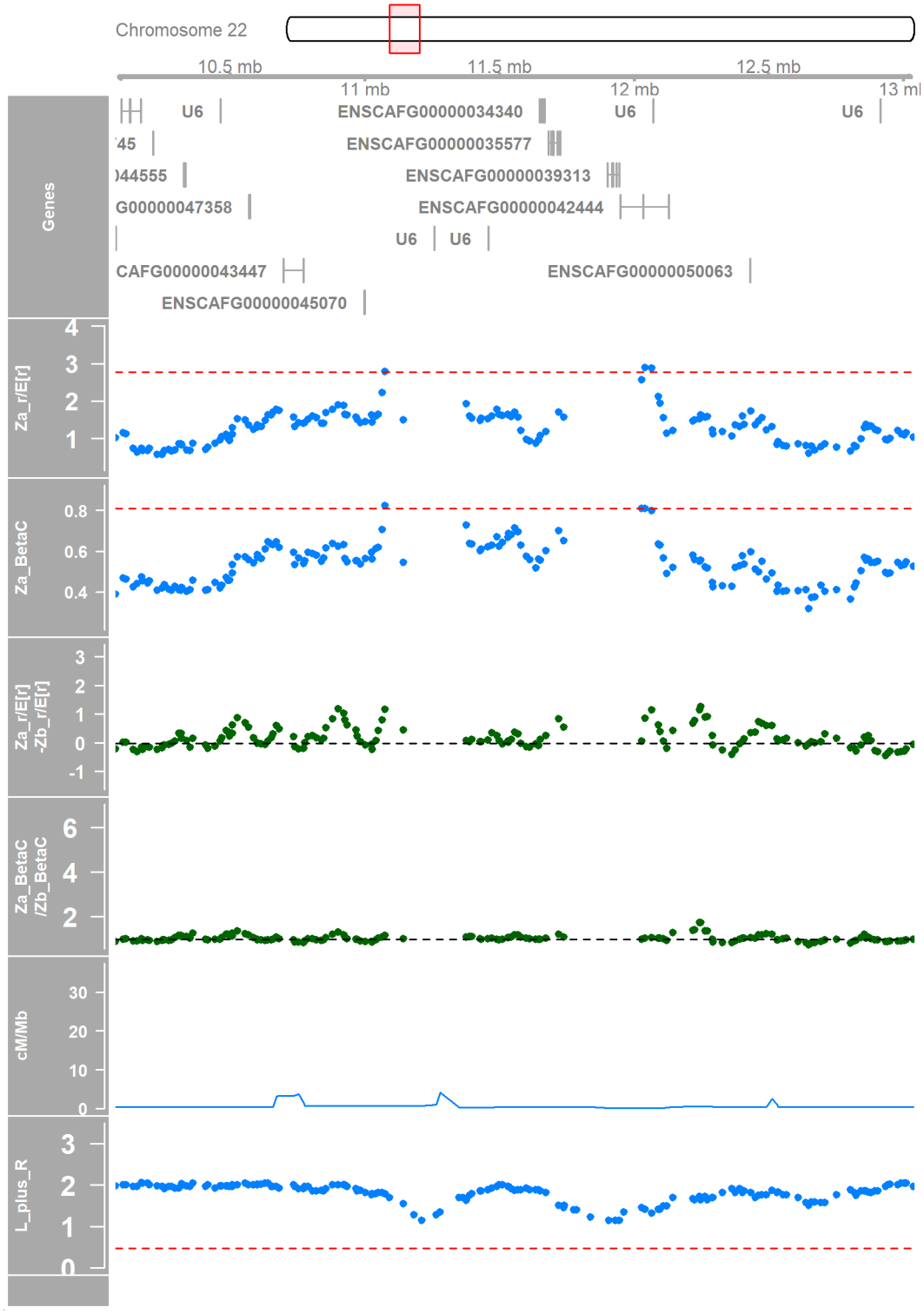
Appendix Figure 38 Candidate region 18:29595073



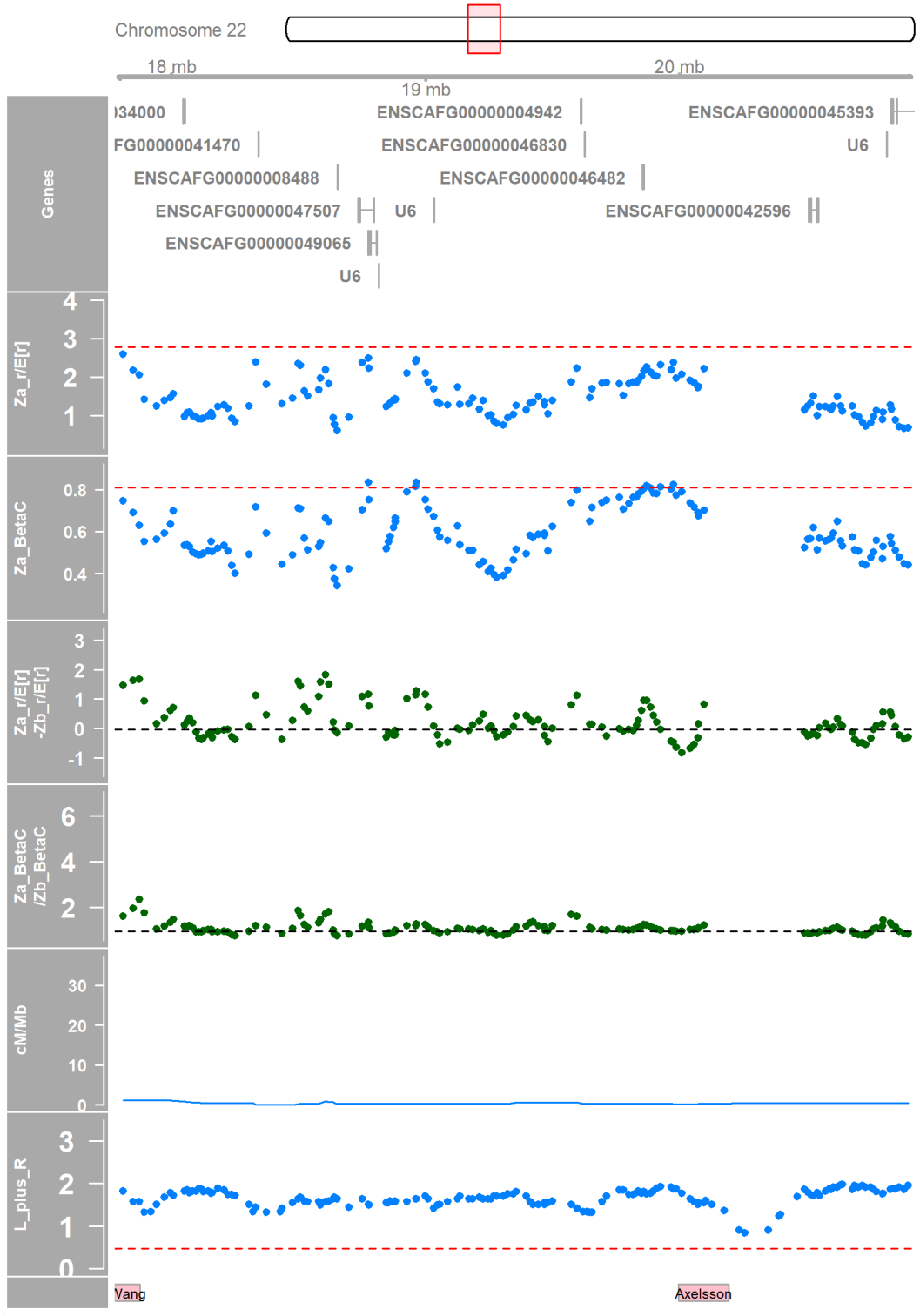
Appendix Figure 39 Candidate region 19:7095253-7122489



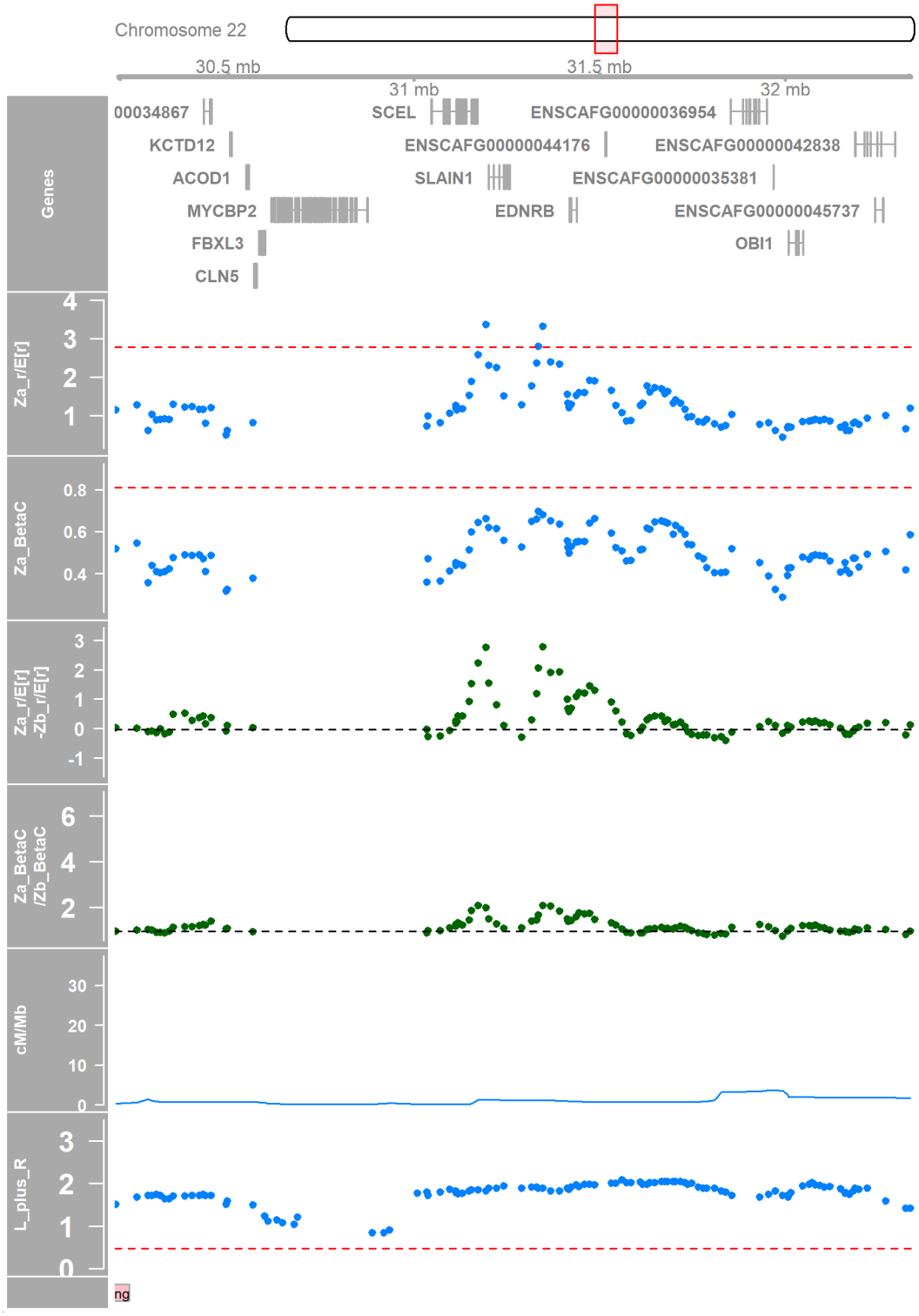
Appendix Figure 40 Candidate region 20:13387022



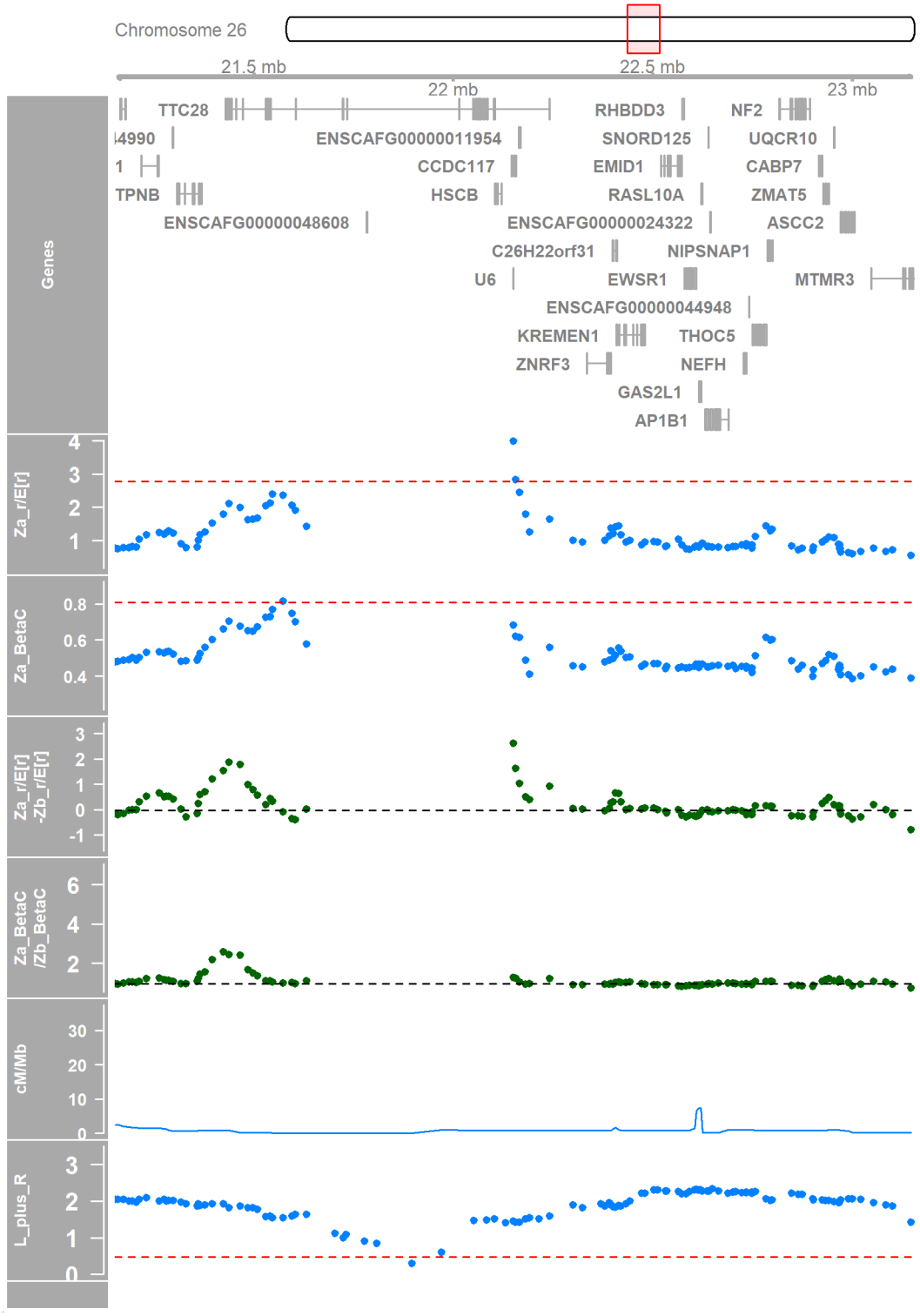
Appendix Figure 41 Candidate region 22:11073667-12039716



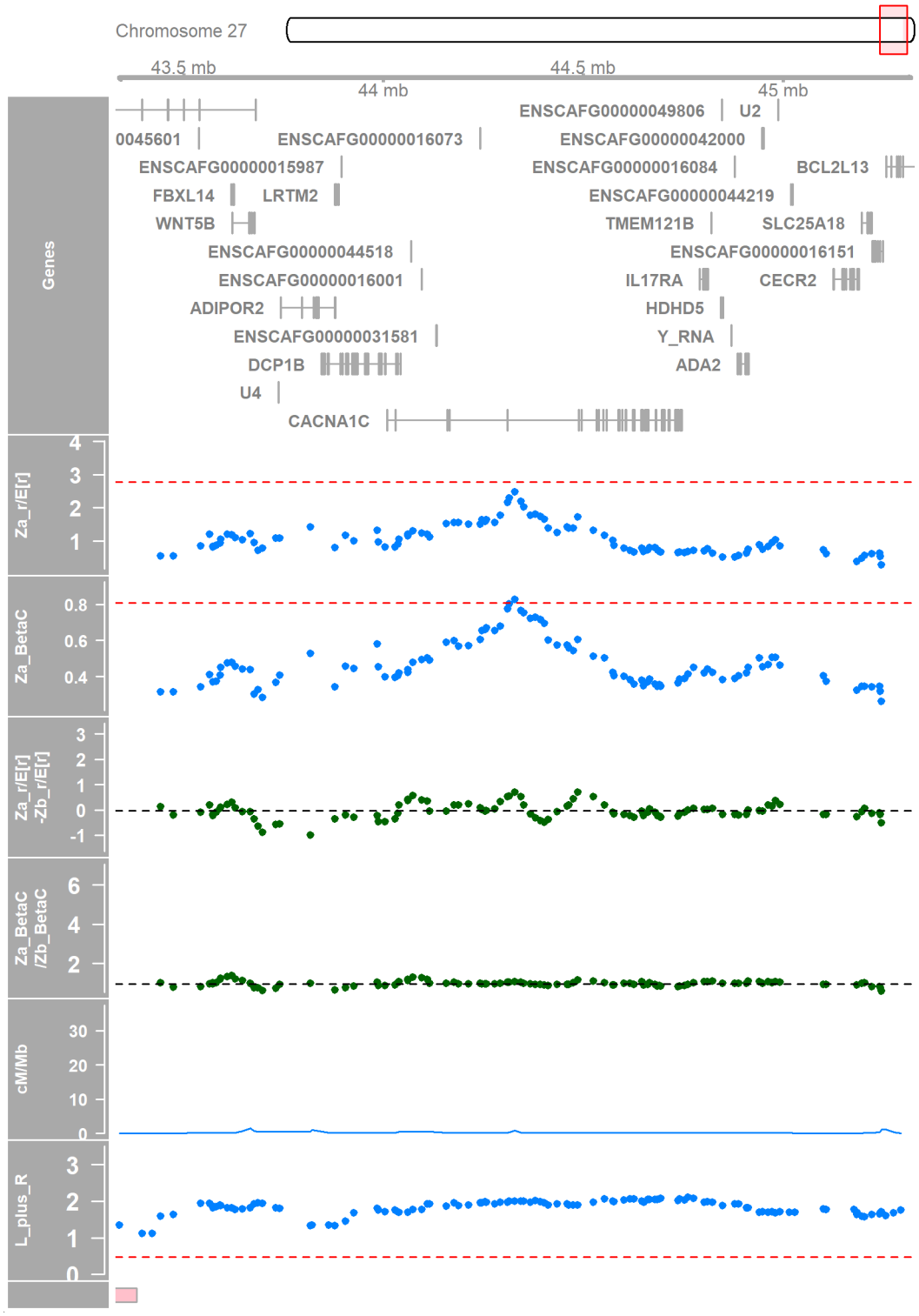
Appendix Figure 42 Candidate region 22:18774821-19925395



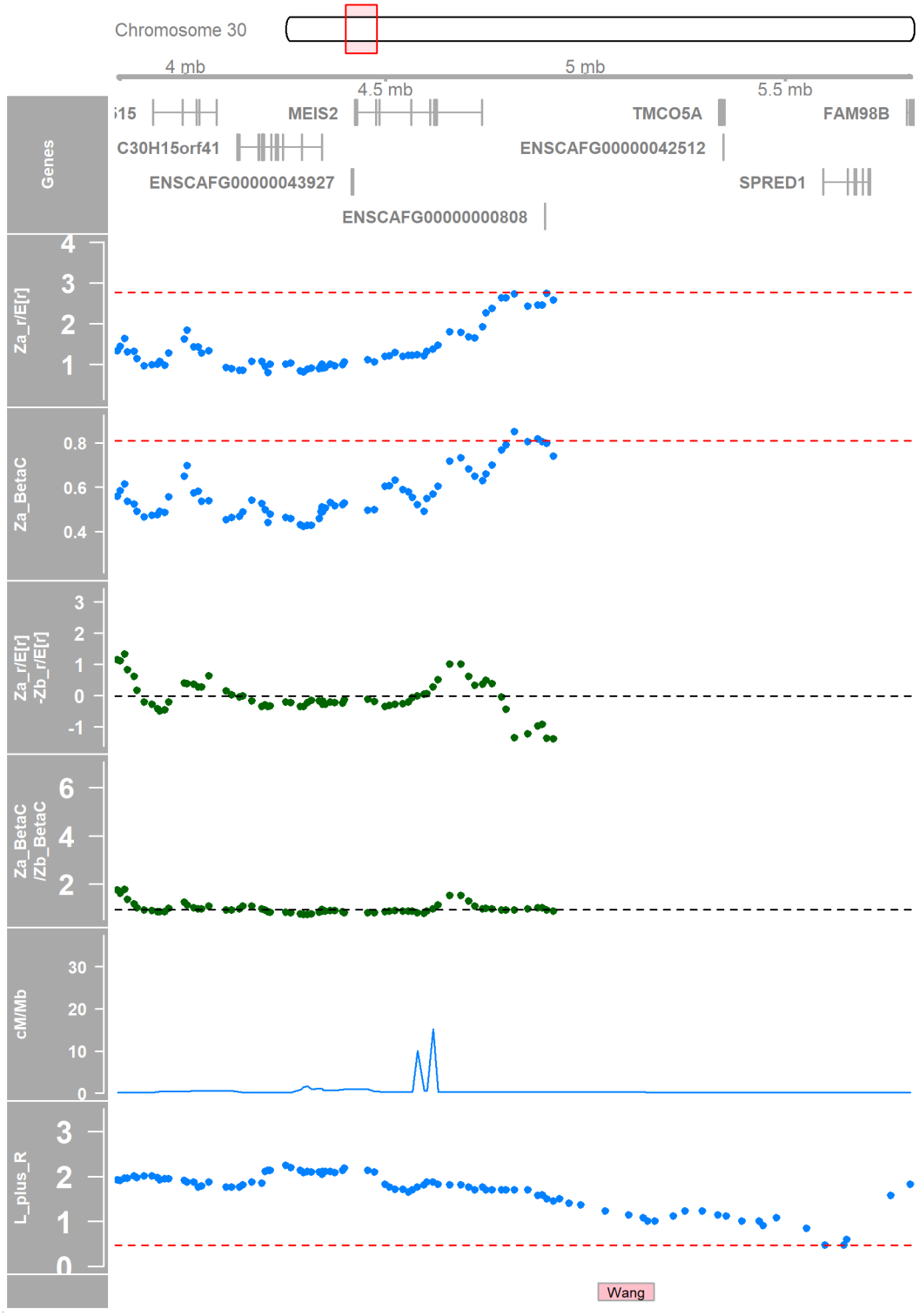
Appendix Figure 43 Candidate region 22:31194138-31347124



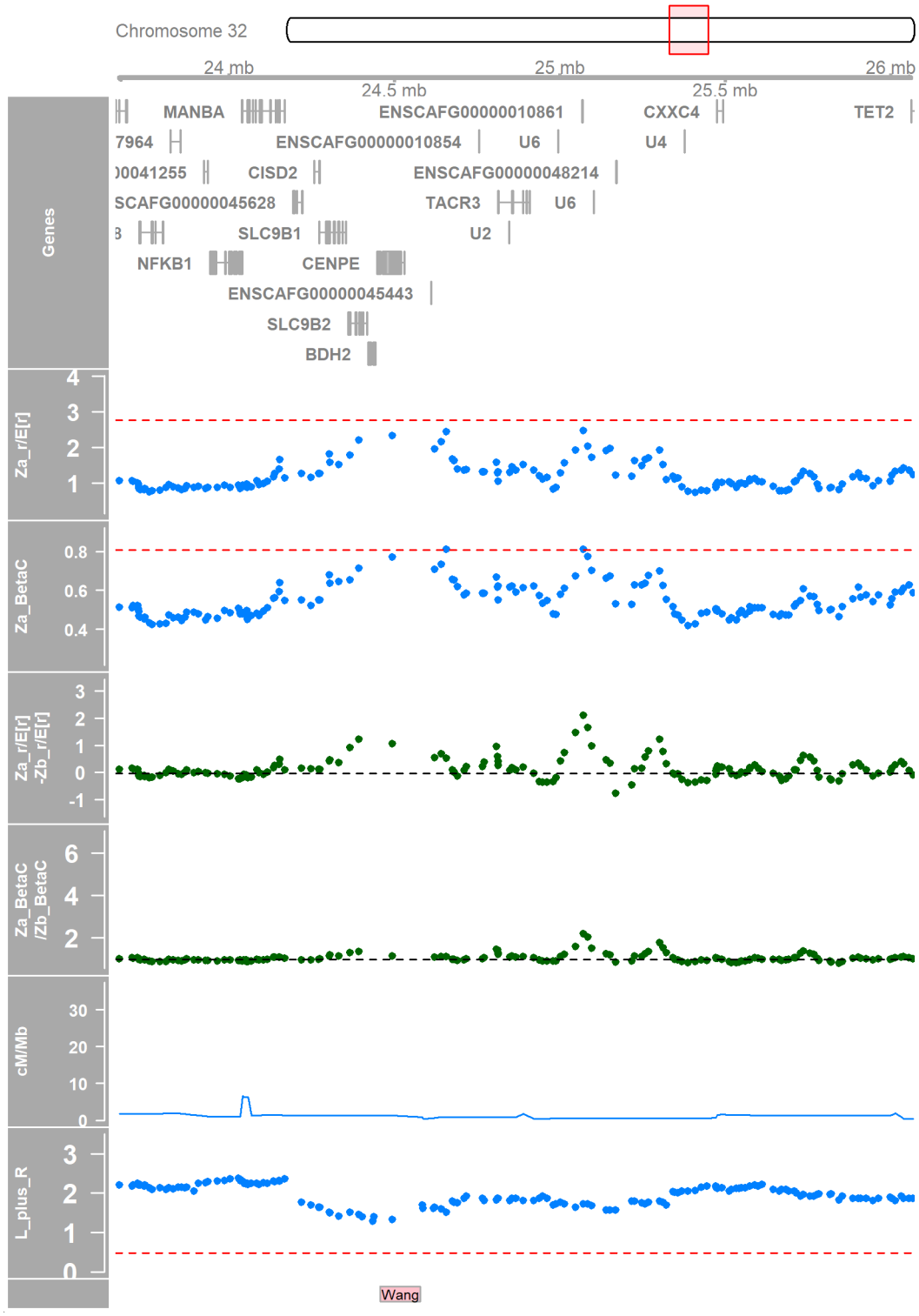
Appendix Figure 44 Candidate region 26:22151015-22156289



Appendix Figure 45 Candidate region 27:44328723



Appendix Figure 46 Candidate region 30:4822803



Appendix Figure 47 Candidate region 32:24657487-25070561

List of References

1. Horscroft, C., et al., *Sequencing era methods for identifying signatures of selection in the genome.* Briefings in Bioinformatics, 2018. **20**(6): p. 1997-2008.

2. Mora, C., et al., *How Many Species Are There on Earth and in the Ocean?* PLOS Biology, 2011. **9**(8): p. e1001127.

3. Betts, H.C., et al., *Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin.* Nature Ecology & Evolution, 2018. **2**(10): p. 1556-1562.

4. Allen, J.A., *The influence of Physical conditions in the genesis of species.* Radical Review, 1877. **1**: p. 108-140.

5. Wedel, M.J., *A monument of inefficiency: The presumed course of the recurrent laryngeal nerve in sauropod dinosaurs.* Acta Palaeontologica Polonica, 2012. **57**(2): p. 251-256.

6. Ballesteros, J.A. and P.P. Sharma, *A Critical Appraisal of the Placement of Xiphosura (Chelicerata) with Account of Known Sources of Phylogenetic Error.* Systematic Biology, 2019. **68**(6): p. 896-917.

7. Baldauf, S.L. and J.D. Palmer, *Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins.* Proceedings of the National Academy of Sciences, 1993. **90**(24): p. 11558-11562.

8. Pace, N.R., *Time for a change.* Nature, 2006. **441**(7091): p. 289-289.

9. Gheyas, A.A., et al., *Functional classification of 15 million SNPs detected from diverse chicken populations.* DNA Research, 2015. **22**(3): p. 205-217.

10. Cortés, A.J., F. López-Hernández, and D. Osorio-Rodriguez, *Predicting Thermal Adaptation by Looking Into Populations' Genomic Past.* Frontiers in Genetics, 2020. **11**(1093).

11. Juliana, P., et al., *Integrating genomic-enabled prediction and high-throughput phenotyping in breeding for climate-resilient bread wheat.* Theoretical and Applied Genetics, 2019. **132**(1): p. 177-194.

12. Garner, J.B., et al., *Genomic Selection Improves Heat Tolerance in Dairy Cattle.* Scientific Reports, 2016. **6**(1): p. 34114.

13. Rovelli, G., et al., *The genetics of phenotypic plasticity in livestock in the era of climate change: a review.* Italian Journal of Animal Science, 2020. **19**(1): p. 997-1014.

14. Elliott, D.E. and J.V. Weinstock, *Helminth-host immunological interactions: prevention and control of immune-mediated diseases.* Annals of the New York Academy of Sciences, 2012. **1247**: p. 83-96.

15. Merlo, L.M.F., et al., *Cancer as an evolutionary and ecological process.* Nature Reviews Cancer, 2006. **6**(12): p. 924-935.

16. Gluckman, P., et al., *Principles of evolutionary medicine*. 2016: Oxford University Press.

17. Little, T.J., et al., *Harnessing evolutionary biology to combat infectious disease.* Nature Medicine, 2012. **18**(2): p. 217-220.

18. Cassini, A., et al., *Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis.* The Lancet Infectious Diseases, 2019. **19**(1): p. 56-66.

19. Holmes, A.H., et al., *Understanding the mechanisms and drivers of antimicrobial resistance.* The Lancet, 2016. **387**(10014): p. 176-187.

20. Trevisanato, S.I., *RECONSTRUCTING ANAXIMANDER'S BIOLOGICAL MODEL UNVEILS A THEORY OF EVOLUTION AKIN TO DARWIN'S, THOUGH CENTURIES BEFORE THE BIRTH OF SCIENCE.* Acta Medico-Historica Adriatica, 2016. **14**(1): p. 63-72.

21. Papazian, M., *Gods and fossils: Inference and scientific method in Xenophanes's philosophy*, in *Philosopher Kings and Tragic Heroes*, H.L. Reid and D. Tanasi, Editors. 2016, Parnassos Press – Fonte Aretusa. p. 61-76.

22. Wright, M.R., *Empedocles, the extant fragments*. 1981: Yale University Press.

23. Barnes, J., *Complete works of Aristotle, volume 1: The revised Oxford translation*. Vol. 192. 1984: Princeton University Press.

24. Hort, A., *Theophrastus: enquiry into plants.* Cambridge (Mass.), 1916.

25. Tzu, C., *The complete works of Chuang Tzu.* Trans. Burton Watson. New York, 1968.

26. Campbell, G.L., *Lucretius on creation and evolution: a commentary on De rerum natura, book five, lines 772-1104*. 2003: Oxford University Press on Demand.

27. Zirkle, C., *Natural Selection before the "Origin of Species".* Proceedings of the American Philosophical Society, 1941. **84**(1): p. 71-123.

28. Hamidullah, M., *The Emergence of Islam: Lectures on the Development of Islamic World-view, Intellectual Tradition and Polity*. 1999: Adam Publishers.

29. Khaldun, I., *Muqaddimah*. 2003, Trans. Franz Rosenthal. Lubnan: Dar al-Fikr.

30. Linné, C.v. and L. Salvius, *Caroli Linnaei...Systema naturae per regna tria naturae :secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis*. Vol. v.1. 1758, Holmiae :: Impensis Direct. Laurentii Salvii.

31. Stearn, W.T., *The Background of Linnaeus's Contributions to the Nomenclature and Methods of Systematic Biology.* Systematic Biology, 1959. **8**(1): p. 4-22.

32. Reid, G.M., *Carolus Linnaeus (1707-1778): his life, philosophy and science and its relationship to modern biology and medicine.* Taxon, 2009. **58**(1): p. 18-31.

33. Diogo, R., *Links between the discovery of primates and anatomical comparisons with humans, the chain of being, our place in nature, and racism.* Journal of Morphology, 2018. **279**(4): p. 472-493.

34. Glass, B., *Maupertuis, Pioneer of Genetics and Evolution (1959)*, in *The Essential Naturalist*. 2019, University of Chicago Press. p. 424-439.

35. Buffon, G.L.L., *Natural history.* 1791.

36. Harris, M., *The rise of anthropological theory : a history of theories of culture*. 1968, London: Routledge & K. Paul.

37. Israel, J.I., *Radical enlightenment: philosophy and the making of modernity, 1650-1750*. 2001: Oxford University Press, USA.

38. Lamarck, J.-B., *Zoological Philosophy, translated by Hugh Elliot*. 1914, Macmillan and Co., London.

39. Jablonka, E., M.J. Lamb, and E. Avital, *'Lamarckian' mechanisms in darwinian evolution.* Trends in Ecology & Evolution, 1998. **13**(5): p. 206-210.

40. Lumey, L.H., A.D. Stein, and A.C.J. Ravelli, *Timing of prenatal starvation in women and birth weight in their first and second born offspring: the Dutch famine birth cohort study.* European Journal of Obstetrics & Gynecology and Reproductive Biology, 1995. **61**(1): p. 23-30.

41. Darwin, E., *Zoonomia; Or, The Laws of Organic Life: In Three Parts: Complete in Two Volumes*. Vol. 1. 1809: Thomas & Andrews.

42. Darwin, C. and A. Wallace, *On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection.* Journal of the proceedings of the Linnean Society of London. Zoology, 1858. **3**(9): p. 45-62.

43. Darwin, C., *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. 1st edition ed. 1859, London: John Murray.

44. Mendel, J.G., *Experiments in plant hybridization.* Trans. William Batesman. Journal of the Royal Horticultural Society, 1901. **26**: p. 1-32.

45. Fisher, R.A., *The Genetical Theory of Natural Selection*. 1930, Oxford: The Clarendon Press.

46. Haldane, J.B., *The causes of evolution*. Vol. 5. 1990: Princeton University Press.

47. Huxley, J., *Evolution: The modern synthesis*. 1942, London: Allen & Unwin.

48. Wright, S., *Statistical genetics and evolution.* Bull. Amer. Math. Soc., 1942. **48**: p. 223-246.

49. Wright, S., *The genetical structure of populations.* Ann Eugen, 1951. **15**(4): p. 323-54.

50. Dobzhansky, T., *Genetics and the Origin of Species*. 1937, New York: Columbia university press.

51. Morgan, T.H., et al., *The Mechanism of Mendelian Heredity*. 1915, New York: Henry Holt.

52. Mayr, E., *Systematics and the Origin of Species, from the Viewpoint of a Zoologist*. 1942, Cambridge: Harvard University Press.

53. Mayr, E., *The objects of selection.* Proceedings of the National Academy of Sciences, 1997. **94**(6): p. 2091-2094.

54. Huneman, P., *Special Issue Editor's Introduction: "Revisiting the Modern Synthesis".* Journal of the History of Biology, 2019. **52**(4): p. 509-518.

55. Rose, M.R. and T.H. Oakley, *The new biology: beyond the Modern Synthesis.* Biology Direct, 2007. **2**(1): p. 30.

56. Blainville, H.d., *Analyse des principaux travaux dans les sciences physiques, publiés dans l'année 1821.* Journal de physique, 1822. **94**.

57. Cuvier, G., *Recherches sur les ossements fossiles de quadrupèdes, tome II ossements fossiles de quadrupèdes pachydermes et d'éléphants, déterrés dans les terrains meubles ou d'alluvion*. 1812: Déterville.

58. Simpson, G.G., *Tempo and Mode in Evolution*. 1944, New York: Columbia University Press.

59. Fuss, J., et al., *Potential hominin affinities of Graecopithecus from the Late Miocene of Europe.* PloS one, 2017. **12**(5): p. e0177127-e0177127.

60. Klein, R.G., *ANTHROPOLOGY: What Do We Know About Neanderthals and Cro-Magnon Man?* The American Scholar, 1983. **52**(3): p. 386-392.

61. Krause, J., et al., *The complete mitochondrial DNA genome of an unknown hominin from southern Siberia.* Nature, 2010. **464**(7290): p. 894-897.

62. Green, R.E., et al., *A Draft Sequence of the Neandertal Genome.* Science, 2010. **328**(5979): p. 710-722.

63. Sawyer, S., et al., *Nuclear and mitochondrial DNA sequences from two Denisovan individuals.* Proceedings of the National Academy of Sciences of the United States of America, 2015. **112**(51): p. 15696-15700.

64. Petraglia, M., *Hominins on the move: An assessment of anthropogenic shaping of environments in the Palaeolithic*, in *Human Dispersal and Species Movement: From Prehistory to the Present*, M. Petraglia, N. Boivin, and R. Crassard, Editors. 2017, Cambridge University Press: Cambridge. p. 90-118.

65. Dahm, R., *Discovering DNA: Friedrich Miescher and the early years of nucleic acid research.* Human Genetics, 2008. **122**(6): p. 565-581.

66. Avery, O.T., C.M. MacLeod, and M. McCarty, *Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III.* The Journal of experimental medicine, 1944. **79**(2): p. 137-158.

67. Hershey, A.D. and M. Chase, *Independent functions of viral protein and nucleic acid in growth of bacteriophage.* Journal of general physiology, 1952. **36**(1): p. 39-56.

68. Watson, J.D. and F.H.C. Crick, *Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid.* Nature, 1953. **171**(4356): p. 737-738.

69. Franklin, R.E. and R.G. Gosling, *Molecular Configuration in Sodium Thymonucleate.* Nature, 1953. **171**(4356): p. 740-741.

70. Nägeli, C., *Zur entwickelungsgeschichte des pollens bei den phanerogamen*. 1842, Zurich: Orell, Füssli & Comp.

71. TJIO, J.H. and A. LEVAN, *THE CHROMOSOME NUMBER OF MAN.* Hereditas, 1956. **42**(1-2): p. 1-6.

72. Lejeune, J., *Etude des chromosomes somatiques de neuf enfants mongoliens.* CR Acad Sci (Paris), 1959. **248**: p. 1721-1722.

73. Patau, K., et al., *MULTIPLE CONGENITAL ANOMALY CAUSED BY AN EXTRA AUTOSOME.* The Lancet, 1960. **275**(7128): p. 790-793.

74. Ford, C., et al., *A sex-chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome).* 1959.

75. Jacobs, P.A. and J.A. Strong, *A case of human intersexuality having a possible XXY sex-determining mechanism.* Nature, 1959. **183**(4657): p. 302-303.

76. Langer-Safer, P.R., M. Levine, and D.C. Ward, *Immunological method for mapping genes on Drosophila polytene chromosomes.* Proceedings of the National Academy of Sciences, 1982. **79**(14): p. 4381-4385.

77. Rowley, J.D., *A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining.* Nature, 1973. **243**(5405): p. 290-293.

78. Miller, D., *Evolution of primate chromosomes.* Science, 1977. **198**(4322): p. 1116-1124.

79. Ijdo, J., et al., *Origin of human chromosome 2: an ancestral telomere-telomere fusion.* Proceedings of the National Academy of Sciences, 1991. **88**(20): p. 9051-9055.

80. Meyer, M., et al., *A High-Coverage Genome Sequence from an Archaic Denisovan Individual.* Science, 2012. **338**(6104): p. 222-226.

81. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proceedings of the national academy of sciences, 1977. **74**(12): p. 5463-5467.

82. Sanger, F., et al., *Nucleotide sequence of bacteriophage φX174 DNA.* Nature, 1977. **265**(5596): p. 687-695.

83. Maxam, A.M. and W. Gilbert, *A new method for sequencing DNA.* Proceedings of the National Academy of Sciences, 1977. **74**(2): p. 560-564.

84. Staden, R., *A strategy of DNA sequencing employing computer programs.* Nucleic Acids Research, 1979. **6**(7): p. 2601-2610.

85. Anderson, S., *Shotgun DNA sequencing using cloned DNase I-generated fragments.* Nucleic Acids Research, 1981. **9**(13): p. 3015-3027.

86. Saiki, R., et al., *Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia.* Science, 1985. **230**(4732): p. 1350-1354.

87. Gansauge, M.-T. and M. Meyer, *Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA.* Nature Protocols, 2013. **8**(4): p. 737-748.

88. Nyrén, P., *Enzymatic method for continuous monitoring of DNA polymerase activity.* Analytical Biochemistry, 1987. **167**(2): p. 235-238.

89. Nyren, P., B. Pettersson, and M. Uhlen, *Solid Phase DNA Minisequencing by an Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay.* Analytical Biochemistry, 1993. **208**(1): p. 171-175.

90. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-380.

91. Heather, J.M. and B. Chain, *The sequence of sequencers: The history of sequencing DNA.* Genomics, 2016. **107**(1): p. 1-8.

92. McKernan, K.J., et al., *Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding.* Genome Research, 2009. **19**(9): p. 1527-1541.

93. Drmanac, R., et al., *Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays.* Science, 2010. **327**(5961): p. 78-81.

94. Rothberg, J.M., et al., *An integrated semiconductor device enabling non-optical genome sequencing.* Nature, 2011. **475**(7356): p. 348-352.

95. Braslavsky, I., et al., *Sequence information can be obtained from single DNA molecules.* Proceedings of the National Academy of Sciences, 2003. **100**(7): p. 3960-3964.

96. Harris, T.D., et al., *Single-molecule DNA sequencing of a viral genome.* Science, 2008. **320**(5872): p. 106-109.

97. Eid, J., et al., *Real-Time DNA Sequencing from Single Polymerase Molecules.* Science, 2009. **323**(5910): p. 133-138.

98. Varongchayakul, N., et al., *Single-molecule protein sensing in a nanopore: a tutorial.* Chemical Society Reviews, 2018. **47**(23): p. 8512-8524.

99. Mikheyev, A.S. and M.M.Y. Tin, *A first look at the Oxford Nanopore MinION sequencer.* Molecular Ecology Resources, 2014. **14**(6): p. 1097-1102.

100. The International HapMap Consortium, et al., *The International HapMap Project.* Nature, 2003. **426**: p. 789.

101. The International HapMap Consortium, et al., *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. **449**(7164): p. 851-U3.

102. The 1000 Genomes Project Consortium, et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**: p. 1061.

103. Gurdasani, D., et al., *The African Genome Variation Project shapes medical genetics in Africa.* Nature, 2014. **517**: p. 327.

104. Mark, C., et al., *The 100,000 Genomes Project Protocol*. 2017.

105. Sudlow, C., et al., *UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age.* PLOS Medicine, 2015. **12**(3): p. e1001779.

106. Al Kuwari, H., et al., *The Qatar Biobank: background and methods.* BMC Public Health, 2015. **15**(1): p. 1208.

107. Chen, Z., et al., *China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up.* International Journal of Epidemiology, 2011. **40**(6): p. 1652-1666.

108. FinnGen, *Documentation of R4 release*. 2020. [cited 22/12/2020] Available from: https://finngen.gitbook.io/documentation/.

109. Mullard, A., *23andMe sets sights on UK/Canada, signs up Genentech.* Nature Biotechnology, 2015. **33**(2): p. 119-119.

110. Kossel, A. and A. Neumann, *Ueber das thymin, ein spaltungsproduct der nucleinsäure.* Berichte der deutschen chemischen Gesellschaft, 1893. **26**(3): p. 2753-2756.

111. Freese, E., *THE DIFFERENCE BETWEEN SPONTANEOUS AND BASE-ANALOGUE INDUCED MUTATIONS OF PHAGE T4.* Proceedings of the National Academy of Sciences of the United States of America, 1959. **45**(4): p. 622-633.

112. Clamp, M., et al., *Distinguishing protein-coding and noncoding genes in the human genome.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(49): p. 19428-19433.

113. Crick, F.H., *On protein synthesis.* Symp Soc Exp Biol, 1958. **12**: p. 138-63.

114. Nirenberg, M.W. and J.H. Matthaei, *The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides.* Proceedings of the National Academy of Sciences, 1961. **47**(10): p. 1588-1602.

115. Crick, F.H.C., et al., *General Nature of the Genetic Code for Proteins.* Nature, 1961. **192**(4809): p. 1227-1232.

116. McClintock, B., *The origin and behavior of mutable loci in maize.* Proceedings of the National Academy of Sciences of the United States of America, 1950. **36**(6): p. 344-355.

117. Hof, A.E.v.t., et al., *The industrial melanism mutation in British peppered moths is a transposable element.* Nature, 2016. **534**: p. 102.

118. Cheng, Z., et al., *A genome-wide comparison of recent chimpanzee and human segmental duplications.* Nature, 2005. **437**(7055): p. 88-93.

119. Hoban, S., et al., *Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions.* The American Naturalist, 2016. **188**(4): p. 379-397.

120. Hodgkinson, A. and A. Eyre-Walker, *Human Triallelic Sites: Evidence for a New Mutational Mechanism?* Genetics, 2010. **184**(1): p. 233-241.

121. Morita, A., et al., *Genotyping of triallelic SNPs using TaqMan® PCR.* Molecular and Cellular Probes, 2007. **21**(3): p. 171-176.

122. Creighton, H.B. and B. McClintock, *A Correlation of Cytological and Genetical Crossing-Over in Zea Mays.* Proceedings of the National Academy of Sciences of the United States of America, 1931. **17**(8): p. 492-497.

123. Janssens, F.A., R. Koszul, and D. Zickler, *La Theorie de la Chiasmatypie.* Nouvelle interprétation des cinèses de maturation, 2012. **191**(2): p. 319-346.

124. MORGAN, T.H., *RANDOM SEGREGATION VERSUS COUPLING IN MENDELIAN INHERITANCE.* Science, 1911. **34**(873): p. 384-384.

125. Haldane, J., *The combination of linkage values and the calculation of distances between the loci of linked factors.* J Genet, 1919. **8**(29): p. 299-309.

126. Gerrish, P.J. and R.E. Lenski, *The fate of competing beneficial mutations in an asexual population*, in *Mutation and Evolution*, R.C. Woodruff and J.N. Thompson, Editors. 1998, Springer Netherlands: Dordrecht. p. 127-144.

127. Hill, W.G. and A. Robertson, *The effect of linkage on limits to artificial selection.* Genetical Research, 1966. **8**(3): p. 269-294.

128. Muller, H.J., *The relation of recombination to mutational advance.* Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 1964. **1**(1): p. 2-9.

129. Felsenstein, J., *The evolutionary advantage of recombination.* Genetics, 1974. **78**(2): p. 737-756.

130. Myers, S., et al., *The distribution and causes of meiotic recombination in the human genome.* Biochemical Society Transactions, 2006. **34**: p. 526-530.

131. Hassold, T. and P. Hunt, *To err (meiotically) is human: the genesis of human aneuploidy.* Nature Reviews Genetics, 2001. **2**(4): p. 280-291.

132. Pardo-Manuel de Villena, F. and C. Sapienza, *Recombination is proportional to the number of chromosome arms in mammals.* Mammalian Genome, 2001. **12**(4): p. 318-322.

133. Sturtevant, A.H., *The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association.* Journal of Experimental Zoology, 1913. **14**(1): p. 43-59.

134. Fan, H.C., et al., *Whole-genome molecular haplotyping of single cells.* Nature Biotechnology, 2011. **29**(1): p. 51.

135. Kong, A., et al., *A high-resolution recombination map of the human genome.* Nature Genetics, 2002. **31**: p. 241.

136. Myers, S., et al., *A fine-scale map of recombination rates and hotspots across the human genome.* Science, 2005. **310**(5746): p. 321-324.

137. Pengelly, R.J., et al., *Whole genome sequences are required to fully resolve the linkage disequilibrium structure of human populations.* BMC Genomics, 2015. **16**(1): p. 666.

138. Kong, A., et al., *Fine-scale recombination rate differences between sexes, populations and individuals.* Nature, 2010. **467**(7319): p. 1099-1103.

139. Clark, A.G., X. Wang, and T. Matise, *Contrasting Methods of Quantifying Fine Structure of Human Recombination*, in *Annual Review of Genomics and Human Genetics, Vol 11*, A. Chakravarti and E. Green, Editors. 2010, Annual Reviews: Palo Alto. p. 45-64.

140. Wu, R. and Z.-B. Zeng, *Joint Linkage and Linkage Disequilibrium Mapping in Natural Populations.* Genetics, 2001. **157**(2): p. 899-909.

141. Sun, L., et al., *A unifying experimental design for dissecting tree genomes.* Trends in Plant Science, 2015. **20**(8): p. 473-476.

142. Pritchard, J.K. and M. Przeworski, *Linkage disequilibrium in humans: Models and data.* American Journal of Human Genetics, 2001. **69**(1): p. 1-14.

143. Slatkin, M., *Linkage disequilibrium--understanding the evolutionary past and mapping the medical future.* Nature reviews. Genetics, 2008. **9**(6): p. 477-485.

144. Lewontin, R.C., *The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models.* Genetics, 1964. **49**(1): p. 49-67.

145. Hill, W.G. and A. Robertson, *Linkage disequilibrium in finite populations.* Theoretical and Applied Genetics, 1968. **38**(6): p. 226-231.

146. Devlin, B. and N. Risch, *A COMPARISON OF LINKAGE DISEQUILIBRIUM MEASURES FOR FINE-SCALE MAPPING.* Genomics, 1995. **29**(2): p. 311-322.

147. Morton, N.E., et al., *The optimal measure of allelic association.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(9): p. 5217-5221.

148. Mueller, J.C., *Linkage disequilibrium for different scales and applications.* Briefings in Bioinformatics, 2004. **5**(4): p. 355-364.

149. Qanbari, S., *On the Extent of Linkage Disequilibrium in the Genome of Farm Animals.* Frontiers in Genetics, 2020. **10**(1304).

150. Collins, A.R., *Linkage Disequilibrium and Association Mapping Analysis and Applications*. Methods in Molecular Biology, 376. 2007, Totowa, NJ: Humana Press.

151. Alachiotis, N. and P. Pavlidis, *Scalable linkage-disequilibrium-based selective sweep detection: a performance guide.* GigaScience, 2016. **5**(1): p. 1-22.

152. Nielsen, R., *Molecular signatures of natural selection*, in *Annual Review of Genetics*. 2005, Annual Reviews: Palo Alto. p. 197-218.

153. Comeron, J.M., *Background selection as null hypothesis in population genomics: insights and challenges from Drosophila studies.* Philosophical Transactions of the Royal Society B-Biological Sciences, 2017. **372**(1736): p. 13.

154. Muller, H.J., *Our load of mutations.* American journal of human genetics, 1950. **2**(2): p. 111-176.

155. Hernandez, R.D., et al., *Classic Selective Sweeps Were Rare in Recent Human Evolution.* Science, 2011. **331**(6019): p. 920-924.

156. Rees, D.C., T.N. Williams, and M.T. Gladwin, *Sickle-cell disease.* Lancet, 2010. **376**(9757): p. 2018-2031.

157. Akey, J.M., *Constructing genomic maps of positive selection in humans: Where do we go from here?* Genome Research, 2009. **19**(5): p. 711-722.

158. Maynard Smith, J. and J. Haigh, *The hitch-hiking effect of a favourable gene.* Genet Res, 1974. **23**(1): p. 23-35.

159. Haasl, R.J. and B.A. Payseur, *Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication.* Molecular Ecology, 2016. **25**(1): p. 5-23.

160. Hermisson, J. and P.S. Pennings, *Soft Sweeps.* Molecular Population Genetics of Adaptation From Standing Genetic Variation, 2005. **169**(4): p. 2335-2352.

161. Pennings, P.S. and J. Hermisson, *Soft Sweeps III: The Signature of Positive Selection from Recurrent Mutation.* PLOS Genetics, 2006. **2**(12): p. e186.

162. Bersaglieri, T., et al., *Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene.* American Journal of Human Genetics, 2004. **74**(6): p. 1111-1120.

163. Tishkoff, S.A., et al., *Convergent adaptation of human lactase persistence in Africa and Europe.* Nat Genet, 2007. **39**(1): p. 31-40.

164. Enattah, N.S., et al., *Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture.* American Journal of Human Genetics, 2008. **82**(1): p. 57-72.

165. Jensen, J.D., *On the unfounded enthusiasm for soft selective sweeps.* 2014. **5**: p. 5281.

166. Schrider, D.R., et al., *Soft Shoulders Ahead: Spurious Signatures of Soft and Partial Selective Sweeps Result from Linked Hard Sweeps.* Genetics, 2015. **200**(1): p. 267-284.

167. Williams, K.-A. and P. Pennings, *Drug Resistance Evolution in HIV in the Late 1990s: Hard Sweeps, Soft Sweeps, Clonal Interference and the Accumulation of Drug Resistance Mutations.* G3: Genes|Genomes|Genetics, 2020. **10**(4): p. 1213-1223.

168. Pritchard, J.K., J.K. Pickrell, and G. Coop, *The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation.* Current Biology, 2010. **20**(4): p. R208-R215.

169. Turchin, M.C., et al., *Evidence of widespread selection on standing variation in Europe at height-associated SNPs.* Nat Genet, 2012. **44**(9): p. 1015-1019.

170. Lettre, G., *Recent progress in the study of the genetics of height.* Human Genetics, 2011. **129**(5): p. 465-472.

171. Berg, J.J. and G. Coop, *A Population Genetic Signal of Polygenic Adaptation.* PLOS Genetics, 2014. **10**(8): p. e1004412.

172. Fan, S., et al., *Going global by adapting local: A review of recent human adaptation.* Science, 2016. **354**(6308): p. 54-59.

173. Hermisson, J. and P.S. Pennings, *Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation.* Methods in Ecology and Evolution, 2017. **8**(6): p. 700-716.

174. Crisci, J., et al., *The impact of equilibrium assumptions on tests of selection.* Frontiers in Genetics, 2013. **4**(235).

175. Wilson, B.A., D.A. Petrov, and P.W. Messer, *Soft Selective Sweeps in Complex Demographic Scenarios.* Genetics, 2014. **198**(2): p. 669-684.

176. Bank, C., et al., *Thinking too positive? Revisiting current methods of population genetic selection inference.* Trends in Genetics, 2014. **30**(12): p. 540-546.

177. Schrider, D.R., A.G. Shanku, and A.D. Kern, *Effects of Linked Selective Sweeps on Demographic Inference and Model Selection.* Genetics, 2016. **204**(3): p. 1207-1223.

178. Mathew, L.A. and J.D. Jensen, *Evaluating the ability of the pairwise joint site frequency spectrum to co-estimate selection and demography.* Frontiers in Genetics, 2015. **6**: p. 268.

179. Wright, S., *Evolution in Mendelian Populations.* Genetics, 1931. **16**(2): p. 97-159.

180. Karlin, S., *Rates of Approach to Homozygosity for Finite Stochastic Models with Variable Population Size.* The American Naturalist, 1968. **102**(927): p. 443-455.

181. Henn, B.M., L.L. Cavalli-Sforza, and M.W. Feldman, *The great human expansion.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(44): p. 17758-17764.

182. Cavalli-Sforza, L.L., P. Menozzi, and A. Piazza, *The History and Geography of Human Genes*. 1996: Princeton University Press.

183. Tishkoff, S.A., et al., *The Genetic Structure and History of Africans and African Americans.* Science, 2009. **324**(5930): p. 1035-1044.

184. Akey, J.M., et al., *Population History and Natural Selection Shape Patterns of Genetic Variation in 132 Genes.* PLOS Biology, 2004. **2**(10): p. e286.

185. Takahata, N., *Allelic genealogy and human evolution.* Molecular Biology and Evolution, 1993. **10**(1): p. 2-22.

186. Marth, G.T., et al., *The Allele Frequency Spectrum in Genome-Wide Human Variation Data Reveals Signals of Differential Demographic History in Three Large World Populations.* Genetics, 2004. **166**(1): p. 351-372.

187. Winckler, W., et al., *Comparison of fine-scale recombination rates in humans and chimpanzees.* Science, 2005. **308**(5718): p. 107-111.

188. Tenesa, A., et al., *Recent human effective population size estimated from linkage disequilibrium.* Genome research, 2007. **17**(4): p. 520-526.

189. Charlesworth, B., *In defence of doing sums in genetics.* Heredity, 2019. **123**(1): p. 44-49.

190. Watterson, G.A., *On the number of segregating sites in genetical models without recombination.* Theoretical Population Biology, 1975. **7**(2): p. 256-276.

191. Browning, S.R. and B.L. Browning, *Identity by Descent Between Distant Relatives: Detection and Applications.* Annual Review of Genetics, 2012. **46**(1): p. 617-633.

192. Gurgul, A., et al., *The application of genome-wide SNP genotyping methods in studies on livestock genomes.* Journal of Applied Genetics, 2014. **55**(2): p. 197-208.

193. Talebi, R., et al., *Runs of Homozygosity in Modern Chicken Revealed by Sequence Data.* G3: Genes|Genomes|Genetics, 2020. **10**(12): p. 4615-4623.

194. Ellingson, S.R. and D.W. Fardo, *Automated quality control for genome wide association studies.* F1000Research, 2016. **5**: p. 1889-1889.

195. Moorjani, P., et al., *Variation in the molecular clock of primates.* Proceedings of the National Academy of Sciences, 2016. **113**(38): p. 10607-10612.

196. Patterson, N., et al., *Genetic evidence for complex speciation of humans and chimpanzees.* Nature, 2006. **441**(7097): p. 1103-1108.

197. Schlebusch, C.M., et al., *Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago.* Science, 2017. **358**(6363): p. 652-655.

198. Gittelman, R.M., et al., *Archaic Hominin Admixture Facilitated Adaptation to Out-of-Africa Environments.* Curr Biol, 2016. **26**(24): p. 3375-3382.

199. Dannemann, M., Aida M. Andrés, and J. Kelso, *Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors.* The American Journal of Human Genetics, 2016. **98**(1): p. 22-33.

200. Vernot, B. and J.M. Akey, *Resurrecting Surviving Neandertal Lineages from Modern Human Genomes.* Science, 2014. **343**(6174): p. 1017-1021.

201. Bocquet-Appel, J.-P., *When the World's Population Took Off: The Springboard of the Neolithic Demographic Transition.* Science, 2011. **333**(6042): p. 560-561.

202. Diamond, J., *Guns, Germs, and Steel: The Fates of Human Societies (20th Anniversary Edition)*. 2017: W. W. Norton.

203. Armelagos, G.J. and K.N. Harper, *Genomics at the origins of agriculture, part two.* Evolutionary Anthropology: Issues, News, and Reviews, 2005. **14**(3): p. 109-121.

204. Barrera-Redondo, J., D. Piñero, and L.E. Eguiarte, *Genomic, Transcriptomic and Epigenomic Tools to Study the Domestication of Plants and Animals: A Field Guide for Beginners.* Frontiers in Genetics, 2020. **11**(742).

205. Enattah, N.S., et al., *Identification of a variant associated with adult-type hypolactasia.* Nature Genetics, 2002. **30**(2): p. 233-237.

206. Sverrisdóttir, O.Ó., et al., *Direct Estimates of Natural Selection in Iberia Indicate Calcium Absorption Was Not the Only Driver of Lactase Persistence in Europe.* Molecular Biology and Evolution, 2014. **31**(4): p. 975-983.

207. Gerbault, P., et al., *Evolution of lactase persistence: an example of human niche construction.* Philosophical Transactions of the Royal Society B: Biological Sciences, 2011. **366**(1566): p. 863-877.

208. Burger, J., et al., *Low Prevalence of Lactase Persistence in Bronze Age Europe Indicates Ongoing Strong Selection over the Last 3,000 Years.* Current Biology, 2020. **30**(21): p. 4307-4315.e13.

209. Liebert, A., et al., *World-wide distributions of lactase persistence alleles and the complex effects of recombination and selection.* Human Genetics, 2017. **136**(11-12): p. 1445-1453.

210. Harding, R.M., et al., *Evidence for Variable Selective Pressures at MC1R.* The American Journal of Human Genetics, 2000. **66**(4): p. 1351-1361.

211. Alan R. Rogers, David Iltis, and Stephen Wooding, *Genetic Variation at the MC1R Locus and the Time since Loss of Human Body Hair.* Current Anthropology, 2004. **45**(1): p. 105-108.

212. Lamason, R.L., et al., *SLC24A5, a Putative Cation Exchanger, Affects Pigmentation in Zebrafish and Humans.* Science, 2005. **310**(5755): p. 1782-1786.

213. Soejima, M. and Y. Koda, *Population differences of two coding SNPs in pigmentation-related genes SLC24A5 and SLC45A2.* International Journal of Legal Medicine, 2007. **121**(1): p. 36-39.

214. Soejima, M., et al., *Evidence for Recent Positive Selection at the Human AIM1 Locus in a European Population.* Molecular Biology and Evolution, 2005. **23**(1): p. 179-188.

215. Norton, H.L., et al., *Genetic Evidence for the Convergent Evolution of Light Skin in Europeans and East Asians.* Molecular Biology and Evolution, 2006. **24**(3): p. 710-722.

216. Yang, Z., et al., *A Genetic Mechanism for Convergent Skin Lightening during Recent Human Evolution.* Molecular Biology and Evolution, 2016. **33**(5): p. 1177-1187.

217. Yi, X., et al., *Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude.* Science, 2010. **329**(5987): p. 75-78.

218. Huerta-Sánchez, E., et al., *Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA.* Nature, 2014. **512**(7513): p. 194-197.

219. Wu, D.D., et al., *Convergent genomic signatures of high-altitude adaptation among domestic mammals.* National Science Review, 2020. **7**(6): p. 952-963.

220. McKenna, H.T., A.J. Murray, and D.S. Martin, *Human adaptation to hypoxia in critical illness.* Journal of Applied Physiology, 2020. **129**(4): p. 656-663.

221. Liu, L., et al., *Global, regional, and national causes of child mortality in 2000–13, with projections to inform post-2015 priorities: an updated systematic analysis.* The Lancet, 2015. **385**(9966): p. 430-440.

222. Kwiatkowski, D.P., *How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria.* The American Journal of Human Genetics, 2005. **77**(2): p. 171-192.

223. Miller, L.H., et al., *The Resistance Factor to Plasmodium vivax in Blacks.* New England Journal of Medicine, 1976. **295**(6): p. 302-304.

224. Liu, W., et al., *African origin of the malaria parasite Plasmodium vivax.* Nature Communications, 2014. **5**(1): p. 3346.

225. Hamblin, M.T. and A. Di Rienzo, *Detection of the Signature of Natural Selection in Humans: Evidence from the Duffy Blood Group Locus.* The American Journal of Human Genetics, 2000. **66**(5): p. 1669-1679.

226. Edenberg, H.J. and J.N. McClintick, *Alcohol Dehydrogenases, Aldehyde Dehydrogenases, and Alcohol Use Disorders: A Critical Review.* Alcoholism: Clinical and Experimental Research, 2018. **42**(12): p. 2281-2297.

227. Wang, L.-X., et al., *Molecular adaption of alcohol metabolism to agriculture in East Asia.* Quaternary International, 2016. **426**: p. 187-194.

228. Polimanti, R. and J. Gelernter, *ADH1B: From alcoholism, natural selection, and cancer to the human phenome.* American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 2018. **177**(2): p. 113-125.

229. Li, H., et al., *Diversification of the ADH1B Gene during Expansion of Modern Humans.* Annals of Human Genetics, 2011. **75**(4): p. 497-507.

230. Mizoi, Y., et al., *Alcohol sensitivity related to polymorphism of alcohol-metabolizing enzymes in Japanese.* Pharmacology Biochemistry and Behavior, 1983. **18**: p. 127-133.

231. Higuchi, S., et al., *Aldehyde dehydrogenase genotypes In Japanese alcoholics.* The Lancet, 1994. **343**(8899): p. 741-742.

232. Liu, R., et al., *Homozygous Defect in HIV-1 Coreceptor Accounts for Resistance of Some Multiply-Exposed Individuals to HIV-1 Infection.* Cell, 1996. **86**(3): p. 367-377.

233. Faria, N.R., et al., *The early spread and epidemic ignition of HIV-1 in human populations.* Science, 2014. **346**(6205): p. 56-61.

234. Libert, F., et al., *The Δccr5 Mutation Conferring Protection Against HIV-1 in Caucasian Populations Has a Single and Recent Origin in Northeastern Europe.* Human Molecular Genetics, 1998. **7**(3): p. 399-406.

235. Stephens, J.C., et al., *Dating the Origin of the CCR5-Δ32 AIDS-Resistance Allele by the Coalescence of Haplotypes.* The American Journal of Human Genetics, 1998. **62**(6): p. 1507-1515.

236. Galvani, A.P. and M. Slatkin, *Evaluating plague and smallpox as historical selective pressures for the CCR5-Δ32 HIV-resistance allele.* Proceedings of the National Academy of Sciences, 2003. **100**(25): p. 15276-15279.

237. Hütter, G., et al., *Long-Term Control of HIV by CCR5 Delta32/Delta32 Stem-Cell Transplantation.* New England Journal of Medicine, 2009. **360**(7): p. 692-698.

238. Allers, K., et al., *Evidence for the cure of HIV infection by CCR5Δ32/Δ32 stem cell transplantation.* Blood, 2011. **117**(10): p. 2791-2799.

239. Tebas, P., et al., *Gene Editing of CCR5 in Autologous CD4 T Cells of Persons Infected with HIV.* New England Journal of Medicine, 2014. **370**(10): p. 901-910.

240. Greely, H.T., *CRISPR'd babies: human germline genome editing in the 'He Jiankui affair'*.* Journal of Law and the Biosciences, 2019. **6**(1): p. 111-183.

241. Karlsson, E.K., D.P. Kwiatkowski, and P.C. Sabeti, *Natural selection and infectious disease in human populations.* Nature Reviews Genetics, 2014. **15**(6): p. 379-393.

242. Everitt, A.R., et al., *IFITM3 restricts the morbidity and mortality associated with influenza.* Nature, 2012. **484**(7395): p. 519-523.

243. Albright, F.S., et al., *Evidence for a Heritable Predisposition to Death Due to Influenza.* The Journal of Infectious Diseases, 2008. **197**(1): p. 18-24.

244. Spreeuwenberg, P., M. Kroneman, and J. Paget, *Reassessing the Global Mortality Burden of the 1918 Influenza Pandemic.* American Journal of Epidemiology, 2018. **187**(12): p. 2561-2567.

245. Dong, E., H. Du, and L. Gardner, *An interactive web-based dashboard to track COVID-19 in real time.* The Lancet Infectious Diseases, 2020. **20**(5): p. 533-534.

246. Zeberg, H. and S. Pääbo, *The major genetic risk factor for severe COVID-19 is inherited from Neanderthals.* Nature, 2020. **587**(7835): p. 610-612.

247. Zhou, F., et al., *Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study.* The Lancet, 2020. **395**(10229): p. 1054-1062.

248. Souilmi, Y., et al., *An ancient viral epidemic involving host coronavirus interacting genes more than 20,000 years ago in East Asia.* Current Biology, 2021. **31**(16): p. 3504-3514.e9.

249. Lewontin, R.C. and J. Krakauer, *DISTRIBUTION OF GENE FREQUENCY AS A TEST OF THE THEORY OF THE SELECTIVE NEUTRALITY OF POLYMORPHISMS.* Genetics, 1973. **74**(1): p. 175-195.

250. Jensen, J.D., M. Foll, and L. Bernatchez, *The past, present and future of genomic scans for selection.* Molecular Ecology, 2016. **25**(1): p. 1-4.

251. Fu, W. and J.M. Akey, *Selection and Adaptation in the Human Genome.* Annual Review of Genomics and Human Genetics, 2013. **14**(1): p. 467-489.

252. Sabeti, P.C., et al., *Positive Natural Selection in the Human Lineage.* Science, 2006. **312**(5780): p. 1614-1620.

253. Fagny, M., et al., *Exploring the Occurrence of Classic Selective Sweeps in Humans Using Whole-Genome Sequencing Data Sets.* Molecular Biology and Evolution, 2014. **31**(7): p. 1850-1868.

254. Akbari, A., et al., *Identifying the favored mutation in a positive selective sweep.* Nature Methods, 2018. **15**: p. 279.

255. Schrider, D.R. and A.D. Kern, *Supervised Machine Learning for Population Genetics: A New Paradigm.* Trends in Genetics, 2018. **34**(4): p. 301-312.

256. Vitti, J.J., S.R. Grossman, and P.C. Sabeti, *Detecting Natural Selection in Genomic Data*, in *Annual Review of Genetics, Vol 47*, B.L. Bassler, M. Lichten, and G. Schupbach, Editors. 2013, Annual Reviews: Palo Alto. p. 97-120.

257. Jacobs, G.S., T.J. Sluckin, and T. Kivisild, *Refining the Use of Linkage Disequilibrium as a Robust Signature of Selective Sweeps.* Genetics, 2016. **203**(4): p. 1807-25.

258. Kelly, J.K., *A test of neutrality based on interlocus associations.* Genetics, 1997. **146**(3): p. 1197-1206.

259. Alachiotis, N., A. Stamatakis, and P. Pavlidis, *OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets.* Bioinformatics, 2012. **28**(17): p. 2274-2275.

260. Garud, N.R., et al., *Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps.* PLOS Genetics, 2015. **11**(2): p. e1005004.

261. Garud, N.R. and N.A. Rosenberg, *Enhancing the mathematical properties of new haplotype homozygosity statistics for the detection of selective sweeps.* Theoretical Population Biology, 2015. **102**(Supplement C): p. 94-101.

262. Harris, A.M., N.R. Garud, and M. DeGiorgio, *Detection and Classification of Hard and Soft Sweeps from Unphased Genotypes by Multilocus Genotype Identity.* Genetics, 2018. **210**(4): p. 1429-1452.

263. Voight, B.F., et al., *A Map of Recent Positive Selection in the Human Genome.* PLoS Biology, 2006. **4**(3): p. e72.

264. Ferrer-Admetlla, A., et al., *On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure.* Molecular Biology and Evolution, 2014. **31**(5): p. 1275-1291.

265. Sabeti, P.C., et al., *Detecting recent positive selection in the human genome from haplotype structure.* Nature, 2002. **419**(6909): p. 832-837.

266. Tajima, F., *Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.* Genetics, 1989. **123**(3): p. 585-595.

267. Fay, J.C. and C.I. Wu, *Hitchhiking under positive Darwinian selection.* Genetics, 2000. **155**(3): p. 1405-1413.

268. Field, Y., et al., *Detection of human adaptation during the past 2000 years.* Science, 2016. **354**(6313): p. 760-764.

269. Lange, J.D. and J.E. Pool, *A haplotype method detects diverse scenarios of local adaptation from genomic sequence variation.* Molecular Ecology, 2016. **25**(13): p. 3081-3100.

270. Hudson, R.R., M. Slatkin, and W.P. Maddison, *ESTIMATION OF LEVELS OF GENE FLOW FROM DNA-SEQUENCE DATA.* Genetics, 1992. **132**(2): p. 583-589.

271. Sabeti, P.C., et al., *Genome-wide detection and characterization of positive selection in human populations.* Nature, 2007. **449**(7164): p. 913-918.

272. Carvajal-Rodríguez, A., *HacDivSel: Two new methods (haplotype-based and outlier-based) for the detection of divergent selection in pairs of populations.* PLOS ONE, 2017. **12**(4): p. e0175944.

273. Rivas, M.J., S. Dominguez-Garcia, and A. Carvajal-Rodriguez, *Detecting the Genomic Signature of Divergent Selection in Presence of Gene Flow.* Current Genomics, 2015. **16**(3): p. 203-212.

274. Foll, M. and O. Gaggiotti, *A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective.* Genetics, 2008. **180**(2): p. 977-993.

275. Hunter-Zinck, H. and A.G. Clark, *Aberrant Time to Most Recent Common Ancestor as a Signature of Natural Selection.* Molecular Biology and Evolution, 2015. **32**(10): p. 2784-2797.

276. Hudson, R.R., M. Kreitman, and M. Aguade, *A TEST OF NEUTRAL MOLECULAR EVOLUTION BASED ON NUCLEOTIDE DATA.* Genetics, 1987. **116**(1): p. 153-159.

277. DeGiorgio, M., et al., *Sweep Finder 2: increased sensitivity, robustness and flexibility.* Bioinformatics, 2016. **32**(12): p. 1895-1897.

278. Huber, C.D., et al., *Detecting recent selective sweeps while controlling for mutation rate and background selection.* Molecular Ecology, 2016. **25**(1): p. 142-156.

279. McVicker, G., et al., *Widespread Genomic Signatures of Natural Selection in Hominid Evolution.* Plos Genetics, 2009. **5**(5): p. 16.

280. Nielsen, R., et al., *Genomic scans for selective sweeps using SNP data.* Genome Research, 2005. **15**(11): p. 1566-1575.

281. Vy, H.M.T. and Y. Kim, *A Composite-Likelihood Method for Detecting Incomplete Selective Sweep from Population Genomic Data.* Genetics, 2015. **200**(2): p. 633-649.

282. Ma, Y., et al., *Properties of different selection signature statistics and a new strategy for combining them.* Heredity, 2015. **115**(5): p. 426-436.

283. Utsunomiya, Y.T., et al., *Detecting Loci under Recent Positive Selection in Dairy and Beef Cattle by Combining Different Genome-Wide Scan Methods.* Plos One, 2013. **8**(5): p. 11.

284. Randhawa, I.A.S., et al., *Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep.* Bmc Genetics, 2014. **15**: p. 19.

285. Sheehan, S. and Y.S. Song, *Deep Learning for Population Genetic Inference.* PLOS Computational Biology, 2016. **12**(3): p. e1004845.

286. Schrider, D.R. and A.D. Kern, *S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning.* PLOS Genetics, 2016. **12**(3): p. e1005928.

287. Kern, A.D. and D.R. Schrider, *diploS/HIC: An Updated Approach to Classifying Selective Sweeps.* G3 Genes|Genomes|Genetics, 2018. **8**(6): p. 1959-1970.

288. Pybus, M., et al., *Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations.* Bioinformatics, 2015. **31**(24): p. 3946-3952.

289. Lin, K., et al., *Distinguishing Positive Selection From Neutral Evolution: Boosting the Performance of Summary Statistics.* Genetics, 2011. **187**(1): p. 229-244.

290. Grossman, S.R., et al., *Identifying Recent Adaptations in Large-Scale Genomic Data.* Cell, 2013. **152**(4): p. 703-713.

291. Grossman, S.R., et al., *A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection.* Science, 2010. **327**(5967): p. 883-886.

292. Ronen, R., et al., *Learning Natural Selection from the Site Frequency Spectrum.* Genetics, 2013. **195**(1): p. 181-+.

293. Palamara, P.F., et al., *High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability.* Nature Genetics, 2018. **50**(9): p. 1311-1317.

294. Speidel, L., et al., *A method for genome-wide genealogy estimation for thousands of samples.* Nature Genetics, 2019. **51**(9): p. 1321-1329.

295. Edge, M.D. and G. Coop, *Reconstructing the History of Polygenic Scores Using Coalescent Trees.* Genetics, 2018. **211**(1): p. 235-262.

296. Sugden, L.A., et al., *Localization of adaptive variants in human genomes using averaged one-dependence estimation.* Nature Communications, 2018. **9**(1): p. 703.

297. Mughal, M.R. and M. DeGiorgio, *Localizing and Classifying Adaptive Targets with Trend Filtered Regression.* Molecular Biology and Evolution, 2018. **36**(2): p. 252-270.

298. Torada, L., et al., *ImaGene: a convolutional neural network to quantify natural selection from genomic data.* BMC Bioinformatics, 2019. **20**(9): p. 337.

299. Luu, K., E. Bazin, and M.G.B. Blum, *pcadapt: an R package to perform genome scans for selection based on principal component analysis.* Molecular Ecology Resources, 2017. **17**(1): p. 67-77.

300. Privé, F., et al., *Performing Highly Efficient Genome Scans for Local Adaptation with R Package pcadapt Version 4.* Molecular Biology and Evolution, 2020. **37**(7): p. 2153-2154.

301. Bonhomme, M., et al., *Detecting Selection in Population Trees: The Lewontin and Krakauer Test Extended.* Genetics, 2010. **186**(1): p. 241-262.

302. Whitlock, M.C. and K.E. Lotterhos, *Reliable Detection of Loci Responsible for Local Adaptation: Inference of a Null Model through Trimming the Distribution of FST.* The American Naturalist, 2015. **186**(S1): p. S24-S36.

303. Frichot, E., et al., *Fast and Efficient Estimation of Individual Ancestry Coefficients.* Genetics, 2014. **196**(4): p. 973-983.

304. Chen, G.B., et al., *EigenGWAS: finding loci under selection through genome-wide association studies of eigenvectors in structured populations.* Heredity, 2016. **117**(1): p. 51-61.

305. Browning, S.R. and B.L. Browning, *Probabilistic Estimation of Identity by Descent Segment Endpoints and Detection of Recent Selection.* The American Journal of Human Genetics, 2020. **107**(5): p. 895-910.

306. Setter, D., et al., *VolcanoFinder: Genomic scans for adaptive introgression.* PLOS Genetics, 2020. **16**(6): p. e1008867.

307. DeGiorgio, M., K.E. Lohmueller, and R. Nielsen, *A Model-Based Approach for Identifying Signatures of Ancient Balancing Selection in Genetic Data.* PLOS Genetics, 2014. **10**(8): p. e1004561.

308. Racimo, F., *Testing for Ancient Selection Using Cross-population Allele Frequency Differentiation.* Genetics, 2015. **202**(2): p. 733-750.

309. Chen, H., N. Patterson, and D. Reich, *Population differentiation as a test for selective sweeps.* Genome Research, 2010. **20**(3): p. 393-402.

310. Librado, P. and L. Orlando, *Detecting Signatures of Positive Selection along Defined Branches of a Population Tree Using LSD.* Molecular Biology and Evolution, 2018. **35**(6): p. 1520-1535.

311. Cheng, X., C. Xu, and M. DeGiorgio, *Fast and robust detection of ancestral selective sweeps.* Molecular Ecology, 2017. **26**(24): p. 6871-6891.

312. Alachiotis, N. and P. Pavlidis, *RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors.* Communications Biology, 2018. **1**(1): p. 79.

313. Pavlidis, P., et al., *SweeD: Likelihood-Based Detection of Selective Sweeps in Thousands of Genomes.* Molecular Biology and Evolution, 2013. **30**(9): p. 2224-2234.

314. Fujito, N.T., et al., *A new inference method for detecting an ongoing selective sweep.* Genes & Genetic Systems, 2018. **93**(4): p. 149-161.

315. Yang, Z., et al., *Detecting Recent Positive Selection with a Single Locus Test Bipartitioning the Coalescent Tree.* Genetics, 2018. **208**(2): p. 791-805.

316. Fu, Y.X. and W.H. Li, *Statistical tests of neutrality of mutations.* Genetics, 1993. **133**(3): p. 693-709.

317. Tournebize, R., et al., *McSwan: A joint site frequency spectrum method to detect and date selective sweeps across multiple population genomes.* Molecular Ecology Resources, 2019. **19**(1): p. 283-295.

318. Yang, Z.H. and J.P. Bielawski, *Statistical methods for detecting molecular adaptation.* Trends in Ecology & Evolution, 2000. **15**(12): p. 496-503.

319. McDonald, J.H. and M. Kreitman, *Adaptive protein evolution at the Adh locus in Drosophila.* Nature, 1991. **351**(6328): p. 652-654.

320. Parsch, J., J.F. Baines, and Z. Zhang, *The Influence of Demography and Weak Selection on the McDonald–Kreitman Test: An Empirical Study in Drosophila.* Molecular Biology and Evolution, 2009. **26**(3): p. 691-698.

321. Fay, J.C., *Weighing the evidence for adaptation at the molecular level.* Trends in Genetics, 2011. **27**(9): p. 343-349.

322. Nei, M. and W.H. Li, *Mathematical model for studying genetic variation in terms of restriction endonucleases.* Proceedings of the National Academy of Sciences, 1979. **76**(10): p. 5269-5273.

323. Corbett-Detig, R.B., D.L. Hartl, and T.B. Sackton, *Natural Selection Constrains Neutral Diversity across A Wide Range of Species.* PLOS Biology, 2015. **13**(4): p. e1002112.

324. Holsinger, K.E. and B.S. Weir, *FUNDAMENTAL CONCEPTS IN GENETICS Genetics in geographically structured populations: defining, estimating and interpreting F-ST.* Nature Reviews Genetics, 2009. **10**(9): p. 639-650.

325. Weir, B.S. and C.C. Cockerham, *Estimating F-Statistics for the Analysis of Population Structure.* Evolution, 1984. **38**(6): p. 1358-1370.

326. Pavlidis, P. and N. Alachiotis, *A survey of methods and tools to detect recent and strong positive selection.* Journal of Biological Research-Thessaloniki, 2017. **24**: p. 17.

327. Kim, Y. and R. Nielsen, *Linkage Disequilibrium as a Signature of Selective Sweeps.* Genetics, 2004. **167**(3): p. 1513-1524.

328. O'Reilly, P.F., E. Birney, and D.J. Balding, *Confounding between recombination and selection, and the Ped/Pop method for detecting selection.* Genome Research, 2008. **18**(8): p. 1304-1313.

329. Comeron, J.M., R. Ratnappan, and S. Bailin, *The Many Landscapes of Recombination in Drosophila melanogaster.* Plos Genetics, 2012. **8**(10): p. 21.

330. Ferretti, L., S.E. Ramos-Onsins, and M. Pérez-Enciso, *Population genomics from pool sequencing.* Molecular Ecology, 2013. **22**(22): p. 5561-5576.

331. Li, H. and R. Durbin, *Inference of human population history from individual whole-genome sequences.* Nature, 2011. **475**(7357): p. 493-U84.

332. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nature Genetics, 2006. **38**(8): p. 904-909.

333. Harpak, A., A. Bhaskar, and J.K. Pritchard, *Mutation Rate Variation is a Primary Determinant of the Distribution of Allele Frequencies in Humans.* PLOS Genetics, 2016. **12**(12): p. e1006489.

334. Pfeifer, B., et al., *Genome scans for selection and introgression based on k-nearest neighbour techniques.* Molecular Ecology Resources, 2020. **20**(6): p. 1597-1609.

335. Flagel, L., Y. Brandvain, and D.R. Schrider, *The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference.* Molecular Biology and Evolution, 2018. **36**(2): p. 220-238.

336. Webb, G.I., J.R. Boughton, and Z. Wang, *Not so naive Bayes: aggregating one-dependence estimators.* Machine learning, 2005. **58**(1): p. 5-24.

337. Service, S., et al., *Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies.* Nature Genetics, 2006. **38**(5): p. 556-560.

338. International Human Genome Sequencing, C., *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-945.

339. Kwong, J.C., et al., *Whole genome sequencing in clinical and public health microbiology.* Pathology, 2015. **47**(3): p. 199-210.

340. Yates, A.D., et al., *Ensembl 2020.* Nucleic Acids Research, 2019. **48**(D1): p. D682-D688.

341. Lewin, H.A., et al., *Earth BioGenome Project: Sequencing life for the future of life.* Proceedings of the National Academy of Sciences, 2018. **115**(17): p. 4325-4333.

342. Paabo, S., et al., *Genetic analyses from ancient DNA.* Annual Review of Genetics, 2004. **38**: p. 645-679.

343. Reich, D., et al., *Genetic history of an archaic hominin group from Denisova Cave in Siberia.* Nature, 2010. **468**(7327): p. 1053-1060.

344. Sankararaman, S., et al., *The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans.* Current Biology, 2016. **26**(9): p. 1241-1247.

345. Hill, W.G., *ESTIMATION OF LINKAGE DISEQUILIBRIUM IN RANDOMLY MATING POPULATIONS.* Heredity, 1974. **33**(OCT): p. 229-239.

346. Stephens, M., N.J. Smith, and P. Donnelly, *A new statistical method for haplotype reconstruction from population data.* American journal of human genetics, 2001. **68**(4): p. 978-989.

347. Browning, S.R. and B.L. Browning, *Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.* American journal of human genetics, 2007. **81**(5): p. 1084-1097.

348. McVean, G.A.T., et al., *The Fine-Scale Structure of Recombination Rate Variation in the Human Genome.* Science, 2004. **304**(5670): p. 581-584.

349. Kemper, K.E., et al., *Selection for complex traits leaves little or no classic signatures of selection.* Bmc Genomics, 2014. **15**: p. 14.

350. Thornton, K.R. and J.D. Jensen, *Controlling the false-positive rate in multilocus genome scans for selection.* Genetics, 2007. **175**(2): p. 737-750.

351. Kelley, J.L., et al., *Genomic signatures of positive selection in humans and the limits of outlier approaches.* Genome Research, 2006. **16**(8): p. 980-989.

352. Comeron, J.M., *Background Selection as Baseline for Nucleotide Variation across the Drosophila Genome.* Plos Genetics, 2014. **10**(6): p. 18.

353. Schrider, D.R., *Background Selection Does Not Mimic the Patterns of Genetic Diversity Produced by Selective Sweeps.* Genetics, 2020. **216**(2): p. 499-519.

354. Mariño-Ramírez, L., et al., *Transposable elements donate lineage-specific regulatory sequences to host genomes.* Cytogenetic and Genome Research, 2005. **110**(1-4): p. 333-341.

355. Villanueva-Cañas, J.L., et al., *Beyond SNPs: how to detect selection on transposable element insertions.* Methods in Ecology and Evolution, 2017. **8**(6): p. 728-737.

356. Merenciano, M., et al., *Multiple Independent Retroelement Insertions in the Promoter of a Stress Response Gene Have Variable Molecular and Functional Effects in Drosophila.* PLOS Genetics, 2016. **12**(8): p. e1006249.

357. Schlamp, F., et al., *Evaluating the performance of selection scans to detect selective sweeps in domestic dogs.* Molecular ecology, 2016. **25**(1): p. 342-356.

358. Randhawa, I.A.S., et al., *Composite Selection Signals for Complex Traits Exemplified Through Bovine Stature Using Multibreed Cohorts of European and African Bos taurus.* G3: Genes|Genomes|Genetics, 2015. **5**(7): p. 1391-1401.

359. Jain, K. and W. Stephan, *Modes of Rapid Polygenic Adaptation.* Molecular Biology and Evolution, 2017. **34**(12): p. 3169-3175.

360. Guo, J., J. Yang, and P.M. Visscher, *Leveraging GWAS for complex traits to detect signatures of natural selection in humans.* Current Opinion in Genetics & Development, 2018. **53**: p. 9-14.

361. Berg, J.J., et al., *Reduced signal for polygenic adaptation of height in UK Biobank.* eLife, 2019. **8**: p. e39725.

362. Boyrie, L., et al., *A linkage disequilibrium-based statistical test for Genome-Wide Epistatic Selection Scans in structured populations.* Heredity, 2021. **126**(1): p. 77-91.

363. Jones, A.G., S.J. Arnold, and R. Bürger, *The Effects of Epistasis and Pleiotropy on Genome-Wide Scans for Adaptive Outlier Loci.* Journal of Heredity, 2019. **110**(4): p. 494-513.

364. Csilléry, K., et al., *Detecting the genomic signal of polygenic adaptation and the role of epistasis in evolution.* Molecular Ecology, 2018. **27**(3): p. 606-612.

365. Bitarello, B.D., et al., *Signatures of Long-Term Balancing Selection in Human Genomes.* Genome Biology and Evolution, 2018. **10**(3): p. 939-955.

366. Siewert, K.M. and B.F. Voight, *Detecting Long-Term Balancing Selection Using Allele Frequency Correlation.* Molecular Biology and Evolution, 2017. **34**(11): p. 2996-3005.

367. Poh, Y.-P., et al., *On the prospect of identifying adaptive loci in recently bottlenecked populations.* PloS one, 2014. **9**(11): p. e110579-e110579.

368. Provine, W.B., *Ernst Mayr.* Genetics and Speciation, 2004. **167**(3): p. 1041-1046.

369. Jensen, J.D., et al., *Distinguishing Between Selective Sweeps and Demography Using DNA Polymorphism Data.* Genetics, 2005. **170**(3): p. 1401-1410.

370. Schrider, D.R. and A.D. Kern, *Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome.* Molecular Biology and Evolution, 2017. **34**(8): p. 1863-1877.

371. Harris, R.B., A. Sackman, and J.D. Jensen, *On the unfounded enthusiasm for soft selective sweeps II: Examining recent evidence from humans, flies, and viruses.* PLOS Genetics, 2018. **14**(12): p. e1007859.

372. Vy, H.M.T., Y.-J. Won, and Y. Kim, *Multiple Modes of Positive Selection Shaping the Patterns of Incomplete Selective Sweeps over African Populations of Drosophila melanogaster.* Molecular Biology and Evolution, 2017. **34**(11): p. 2792-2807.

373. Garud, N.R., P.W. Messer, and D.A. Petrov, *Detection of hard and soft selective sweeps from Drosophila melanogaster population genomic data.* PLOS Genetics, 2021. **17**(2): p. e1009373.

374. Schrider, D. and A. Kern, *On the well-founded enthusiasm for soft sweeps in humans: a reply to Harris, Sackman, and Jensen*. 2018, Zenodo.

375. McCoy, R.C. and J.M. Akey, *Selection plays the hand it was dealt: evidence that human adaptation commonly targets standing genetic variation.* Genome Biology, 2017. **18**(1): p. 139.

376. Sugden, L.A., et al., *Assessing the validity and reproducibility of genome-scale predictions.* Bioinformatics, 2013. **29**(22): p. 2844-2851.

377. Pavlidis, P., et al., *A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans.* Mol Biol Evol, 2012. **29**(10): p. 3237-48.

378. Stamatakis, A. and F. Izquierdo-Carrasco, *Result verification, code verification and computation of support values in phylogenetics.* Briefings in Bioinformatics, 2011. **12**(3): p. 270-279.

379. Barrett, R.D.H. and H.E. Hoekstra, *Molecular spandrels: tests of adaptation at the genetic level.* Nature Reviews Genetics, 2011. **12**(11): p. 767-780.

380. Vatsiou, A.I., E. Bazin, and O.E. Gaggiotti, *Detection of selective sweeps in structured populations: a comparison of recent methods.* Molecular Ecology, 2016. **25**(1): p. 89-103.

381. Fumagalli, M. and M. Sironi, *Human genome variability, natural selection and infectious diseases.* Current Opinion in Immunology, 2014. **30**(Supplement C): p. 9-16.

382. Cadzow, M., et al., *A bioinformatics workflow for detecting signatures of selection in genomic data.* Frontiers in Genetics, 2014. **5**: p. 293.

383. Lotterhos, K.E., et al., *Composite measures of selection can improve the signal-to-noise ratio in genome scans.* Methods in Ecology and Evolution, 2017. **8**(6): p. 717-727.

384. Leonardi, M., et al., *The evolution of lactase persistence in Europe. A synthesis of archaeological and genetic evidence.* International Dairy Journal, 2012. **22**(2): p. 88-97.

385. Hudson, R.R., *Generating samples under a Wright–Fisher neutral model of genetic variation.* Bioinformatics, 2002. **18**(2): p. 337-338.

386. Teshima, K.M. and H. Innan, *mbs: modifying Hudson's ms software to generate samples of DNA sequences with a biallelic site under selection.* BMC Bioinformatics, 2009. **10**: p. 166-166.

387. Shlyakhter, I., P.C. Sabeti, and S.F. Schaffner, *Cosi2: an efficient simulator of exact and approximate coalescent with selection.* Bioinformatics, 2014. **30**(23): p. 3427-3429.

388. Ewing, G. and J. Hermisson, *MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus.* Bioinformatics, 2010. **26**(16): p. 2064-2065.

389. R Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, 3.4.2, https://www.R-project.org/, 2017.

390. Hunter-Zinck, H., *tsel: Time to most recent common ancestor selection inference (TSel).* R package version 0.5, 2014.

391. Paradis, E., J. Claude, and K. Strimmer, *APE: Analyses of Phylogenetics and Evolution in R language.* Bioinformatics, 2004. **20**(2): p. 289-290.

392. Erikson, Galina A., et al., *Whole-Genome Sequencing of a Healthy Aging Cohort.* Cell, 2016. **165**(4): p. 1002-1011.

393. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-2158.

394. Hollander, M., D.A. Wolfe, and E. Chicken, *Nonparametric Statistical Methods*. 2013, Somerset, UNITED STATES: John Wiley & Sons, Incorporated.

395. Robin, X., et al., *pROC: an open-source package for R and S+ to analyze and compare ROC curves.* BMC Bioinformatics, 2011. **12**(1): p. 77.

396. Dunn, O.J., *Multiple Comparisons among Means.* Journal of the American Statistical Association, 1961. **56**(293): p. 52-64.

397. Metz, C.E., *BASIC PRINCIPLES OF ROC ANALYSIS.* Seminars in Nuclear Medicine, 1978. **8**(4): p. 283-298.

398. McClish, D.K., *ANALYZING A PORTION OF THE ROC CURVE.* Medical Decision Making, 1989. **9**(3): p. 190-195.

399. Jabalameli, M.R., et al., *Gene-dense autosomal chromosomes show evidence for increased selection.* Heredity, 2019. **123**(6): p. 774-783.

400. Hunter, N., *Meiotic Recombination: The Essence of Heredity.* Cold Spring Harbor Perspectives in Biology, 2015. **7**(12).

401. Broman, K.W., et al., *Comprehensive Human Genetic Maps: Individual and Sex-Specific Variation in Recombination.* The American Journal of Human Genetics, 1998. **63**(3): p. 861-869.

402. Coop, G., et al., *High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns Among Humans.* Science, 2008. **319**(5868): p. 1395-1398.

403. Hussin, J., et al., *Age-Dependent Recombination Rates in Human Pedigrees.* PLOS Genetics, 2011. **7**(9): p. e1002251.

404. Otto, S.P. and T. Lenormand, *Resolving the paradox of sex and recombination.* Nature Reviews Genetics, 2002. **3**(4): p. 252-261.

405. Tapper, W., *Linkage disequilibrium maps and location databases.* Methods Mol Biol, 2007. **376**: p. 23-45.

406. Bherer, C., C.L. Campbell, and A. Auton, *Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales.* Nature Communications, 2017. **8**: p. 9.

407. Tapper, W., et al., *A map of the human genome in linkage disequilibrium units.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(33): p. 11835-11839.

408. Jeffreys, A.J., L. Kauppi, and R. Neumann, *Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex.* Nature Genetics, 2001. **29**: p. 217.

409. Myers, S., et al., *A common sequence motif associated with recombination hot spots and genome instability in humans.* Nature Genetics, 2008. **40**(9): p. 1124-1129.

410. Borde, V., et al., *Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites.* The EMBO Journal, 2009. **28**(2): p. 99-111.

411. Parvanov, E.D., P.M. Petkov, and K. Paigen, *Prdm9 Controls Activation of Mammalian Recombination Hotspots.* Science, 2010. **327**(5967): p. 835-835.

412. Powers, N.R., et al., *The Meiotic Recombination Activator PRDM9 Trimethylates Both H3K36 and H3K4 at Recombination Hotspots In Vivo.* PLOS Genetics, 2016. **12**(6): p. e1006146.

413. Paigen, K. and P.M. Petkov, *PRDM9 and Its Role in Genetic Recombination.* Trends in genetics : TIG, 2018. **34**(4): p. 291-300.

414. Baker, C.L., et al., *PRDM9 binding organizes hotspot nucleosomes and limits Holliday junction migration.* Genome Research, 2014. **24**(5): p. 724-732.

415. Kohl, K.P. and J. Sekelsky, *Meiotic and Mitotic Recombination in Meiosis.* Genetics, 2013. **194**(2): p. 327-334.

416. Baudat, F., et al., *PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice.* Science, 2010. **327**(5967): p. 836-840.

417. Berg, I.L., et al., *Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(30): p. 12378-12383.

418. Hinch, A.G., et al., *The landscape of recombination in African Americans.* Nature, 2011. **476**: p. 170.

419. Oliver, P.L., et al., *Accelerated Evolution of the Prdm9 Speciation Gene across Diverse Metazoan Taxa.* Plos Genetics, 2009. **5**(12): p. 14.

420. Ponting, C.P., *What are the genomic drivers of the rapid evolution of PRDM9?* Trends in Genetics, 2011. **27**(5): p. 165-171.

421. Hochwagen, A. and Gabriel A.B. Marais, *Meiosis: A PRDM9 Guide to the Hotspots of Recombination.* Current Biology, 2010. **20**(6): p. R271-R274.

422. Hayashi, K., K. Yoshida, and Y. Matsui, *A histone H3 methyltransferase controls epigenetic events required for meiotic prophase.* Nature, 2005. **438**: p. 374.

423. Muñoz-Fuentes, V., A. Di Rienzo, and C. Vilà, *Prdm9, a major determinant of meiotic recombination hotspots, is not functional in dogs and their wild relatives, wolves and coyotes.* PloS one, 2011. **6**(11): p. e25498-e25498.

424. Brick, K., et al., *Genetic recombination is directed away from functional genomic elements in mice.* Nature, 2012. **485**(7400): p. 642-645.

425. Narasimhan, V.M., et al., *Health and population effects of rare gene knockouts in adult humans with related parents.* Science, 2016. **352**(6284): p. 474-477.

426. Boulton, A., R.S. Myers, and R.J. Redfield, *The hotspot conversion paradox and the evolution of meiotic recombination.* Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(15): p. 8058-8063.

427. Myers, S., et al., *Drive Against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination.* Science, 2010. **327**(5967): p. 876-879.

428. Jeffreys, A.J., et al., *Human recombination hot spots hidden in regions of strong marker association.* Nature Genetics, 2005. **37**(6): p. 601-606.

429. Lesecque, Y., et al., *The Red Queen Model of Recombination Hotspots Evolution in the Light of Archaic and Modern Human Genomes.* PLOS Genetics, 2014. **10**(11): p. e1004790.

430. Ubeda, F., T.W. Russell, and V.A.A. Jansen, *PRDM9 and the evolution of recombination hotspots.* Theoretical Population Biology, 2019. **126**: p. 19-32.

431. Wigginton, J.E., D.J. Cutler, and G.R. Abecasis, *A note on exact tests of Hardy-Weinberg equilibrium.* American Journal of Human Genetics, 2005. **76**(5): p. 887-893.

432. Purcell, S., et al., *PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.* American Journal of Human Genetics, 2007. **81**(3): p. 559-575.

433. Yang, J., et al., *GCTA: A Tool for Genome-wide Complex Trait Analysis.* American Journal of Human Genetics, 2011. **88**(1): p. 76-82.

434. Hartigan, J.A. and M.A. Wong, *Algorithm AS 136: A K-Means Clustering Algorithm.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 1979. **28**(1): p. 100-108.

435. Fearnhead, P. and P. Donnelly, *Estimating recombination rates from population genetic data.* Genetics, 2001. **159**(3): p. 1299-1318.

436. McVean, G., P. Awadalla, and P. Fearnhead, *A coalescent-based method for detecting and estimating recombination from gene sequences.* Genetics, 2002. **160**(3): p. 1231-1241.

437. Green, P.J., *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination.* Biometrika, 1995. **82**(4): p. 711-732.

438. Booker, T.R., R.W. Ness, and P.D. Keightley, *The Recombination Landscape in Wild House Mice Inferred Using Population Genomic Data.* Genetics, 2017. **207**(1): p. 297-309.

439. Kuo, T.-Y., W. Lau, and A.R. Collins, *LDMAP*, in *Linkage Disequilibrium and Association Mapping: Analysis and Applications*, A.R. Collins, Editor. 2007, Humana Press: Totowa, NJ. p. 47-57.

440. Maniatis, N., et al., *The first linkage disequilibrium (LD) maps: Delineation of hot and cold blocks by diplotype analysis.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(4): p. 2228-2233.

441. Malécot, G., *Les mathématiques de l'hérédité*. 1948, Paris: Masson.

442. Zhang, W., et al., *Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(52): p. 18075-18080.

443. Chan, A.H., P.A. Jenkins, and Y.S. Song, *Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster.* PLOS Genetics, 2012. **8**(12): p. e1003090.

444. Dapper, A.L. and B.A. Payseur, *Effects of Demographic History on the Detection of Recombination Hotspots from Linkage Disequilibrium.* Molecular Biology and Evolution, 2018. **35**(2): p. 335-353.

445. Tapper, W., et al., *A comparison of methods to detect recombination hotspots.* Human Heredity, 2008. **66**(3): p. 157-169.

446. Stumpf, M.P.H. and G.A.T. McVean, *Estimating recombination rates from population-genetic data.* Nature Reviews Genetics, 2003. **4**: p. 959.

447. Reed, F.A. and S.A. Tishkoff, *Positive selection can create false hotspots of recombination.* Genetics, 2006. **172**(3): p. 2011-2014.

448. McVean, G., *The Structure of Linkage Disequilibrium Around a Selective Sweep.* Genetics, 2007. **175**(3): p. 1395-1406.

449. Gao, F., et al., *New Software for the Fast Estimation of Population Recombination Rates (FastEPRR) in the Genomic Era.* G3-Genes Genomes Genetics, 2016. **6**(6): p. 1563-1571.

450. Auton, A. and G. McVean, *Recombination rate estimation in the presence of hotspots.* Genome Research, 2007. **17**(8): p. 1219-1227.

451. Fearnhead, P., *SequenceLDhot: detecting recombination hotspots.* Bioinformatics, 2006. **22**(24): p. 3061-3066.

452. Barrett, J.C., et al., *Haploview: analysis and visualization of LD and haplotype maps.* Bioinformatics, 2005. **21**(2): p. 263-265.

453. Yang, T., H.W. Deng, and T.H. Niu, *Critical assessment of coalescent simulators in modeling recombination hotspots in genomic sequences.* Bmc Bioinformatics, 2014. **15**: p. 14.

454. Spence, J.P. and Y.S. Song, *Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations.* Science Advances, 2019. **5**(10): p. eaaw9206.

455. Zhou, Y., B.L. Browning, and S.R. Browning, *Population-Specific Recombination Maps from Segments of Identity by Descent.* The American Journal of Human Genetics, 2020. **107**(1): p. 137-148.

456. V. Barroso, G., N. Puzović, and J.Y. Dutheil, *Inference of recombination maps from a single pair of genomes and its application to ancient samples.* PLOS Genetics, 2019. **15**(11): p. e1008449.

457. Percival, D.B. and A.T. Walden, *Wavelet methods for time series analysis*. Cambridge series in statistical and probabilistic mathematics. 2000.

458. He, L., Y. Wang, and Z. Xiang, *Support driven wavelet frame-based image deblurring.* Information Sciences, 2019. **479**: p. 250-269.

459. Truchetet, F. and O. Laligant, *Review of industrial applications of wavelet and multiresolution-based signal and image processing.* Journal of Electronic Imaging, 2008. **17**(3): p. 11.

460. Bruckstein, A.M., D.L. Donoho, and M. Elad, *From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images.* Siam Review, 2009. **51**(1): p. 34-81.

461. Wall, J.D. and L.S. Stevison, *Detecting Recombination Hotspots from Patterns of Linkage Disequilibrium.* G3-Genes Genomes Genetics, 2016. **6**(8): p. 2265-2271.

462. Liò, P., *Wavelets in bioinformatics and computational biology: state of art and perspectives.* Bioinformatics, 2003. **19**(1): p. 2-9.

463. Cai, H.M., et al., *WaveDec: A Wavelet Approach to Identify Both Shared and Individual Patterns of Copy-Number Variations.* Ieee Transactions on Biomedical Engineering, 2018. **65**(2): p. 353-364.

464. Jang, H., Y. Hur, and H. Lee, *Identification of cancer-driver genes in focal genomic alterations from whole genome sequencing data.* Scientific Reports, 2016. **6**: p. 15.

465. Wiedenhoeft, J., E. Brugel, and A. Schliep, *Fast Bayesian Inference of Copy Number Variants using Hidden Markov Models with Wavelet Compression.* Plos Computational Biology, 2016. **12**(5): p. 28.

466. Harvey, B.S. and S.Y. Ji, *Cloud-Scale Genomic Signals Processing for Robust Large-Scale Cancer Genomic Microarray Data Analysis.* Ieee Journal of Biomedical and Health Informatics, 2017. **21**(1): p. 238-245.

467. Sedlar, K., et al., *Set of rules for genomic signal downsampling.* Computers in Biology and Medicine, 2016. **69**: p. 308-314.

468. Kvikstad, E.M., F. Chiaromonte, and K.D. Makova, *Ride the wavelet: A multiscale analysis of genomic contexts flanking small insertions and deletions.* Genome research, 2009. **19**(7): p. 1153-1164.

469. Sanderson, J., et al., *Reconstructing Past Admixture Processes from Local Genomic Ancestry Using Wavelet Transformation.* Genetics, 2015. **200**(2): p. 470-+.

470. Sansonnet, L., *Wavelet Thresholding Estimation in a Poissonian Interactions Model with Application to Genomic Data.* Scandinavian Journal of Statistics, 2014. **41**(1): p. 200-226.

471. Paape, T., et al., *Fine-scale population recombination rates, hotspots, and correlates of recombination in the Medicago truncatula genome.* Genome biology and evolution, 2012. **4**(5): p. 726-737.

472. Spencer, C.C.A., et al., *The Influence of Recombination on Human Genetic Diversity.* PLOS Genetics, 2006. **2**(9): p. e148.

473. Nason, G.P., *Wavelet Methods in Statistics with R*. Use R. 2008, New York: Springer.

474. Haar, A., *Zur Theorie der orthogonalen Funktionensysteme.* Mathematische Annalen, 1910. **69**(3): p. 331-371.

475. Constantine, W.P., Donald *wmtsa: Wavelet Methods for Time Series Analysis.* R package version 2.0-3, https://CRAN.R-project.org/package=wmtsa, 2017.

476. Daubechies, I., *Orthonormal bases of compactly supported wavelets.* Communications on Pure and Applied Mathematics, 1988. **41**(7): p. 909-996.

477. Daubechies, I., *Orthonormal Bases of Compactly Supported Wavelets II. Variations on a Theme.* SIAM Journal on Mathematical Analysis, 1993. **24**(2): p. 499-519.

478. Morlet, J., et al., *Wave propagation and sampling theory—Part I: Complex signal and scattering in multilayered media.* GEOPHYSICS, 1982. **47**(2): p. 203-221.

479. Morlet, J., et al., *Wave propagation and sampling theory—Part II: Sampling theory and complex waves.* GEOPHYSICS, 1982. **47**(2): p. 222-236.

480. R Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, 3.6.0, https://www.R-project.org/, 2019.

481. Gouhier, T.C.G., Aslak; Simko, Viliam *R package biwavelet: Conduct Univariate and Bivariate Wavelet Analyses.* Version 0.20.17, https://github.com/tgouhier/biwavelet, 2018.

482. Walker, J.S., *A Primer on Wavelets and Their Scientific Applications, Second Edition*. 2008: Taylor & Francis.

483. Grinsted, A., J.C. Moore, and S. Jevrejeva, *Application of the cross wavelet transform and wavelet coherence to geophysical time series.* Nonlinear Processes in Geophysics, 2004. **11**(5-6): p. 561-566.

484. Foufoula-Georgiou, E. and P. Kumar, *Wavelet Analysis in Geophysics: An Introduction*, in *Wavelet Analysis and Its Applications*, E. Foufoula-Georgiou and P. Kumar, Editors. 1994, Academic Press. p. 1-43.

485. Liu, Y., X. San Liang, and R.H. Weisberg, *Rectification of the Bias in the Wavelet Power Spectrum.* Journal of Atmospheric and Oceanic Technology, 2007. **24**(12): p. 2093-2102.

486. Meyers, S.D., B.G. Kelly, and J.J. O'Brien, *An Introduction to Wavelet Analysis in Oceanography and Meteorology: With Application to the Dispersion of Yanai Waves.* Monthly Weather Review, 1993. **121**(10): p. 2858-2866.

487. Torrence, C. and G.P. Compo, *A Practical Guide to Wavelet Analysis.* Bulletin of the American Meteorological Society, 1998. **79**(1): p. 61-78.

488. Gilman, D.L., F.J. Fuglister, and J.M. Mitchell Jr, *On the power spectrum of "red noise".* Journal of the Atmospheric Sciences, 1963. **20**(2): p. 182-184.

489. Torrence, C. and P.J. Webster, *Interdecadal Changes in the ENSO–Monsoon System.* Journal of Climate, 1999. **12**(8): p. 2679-2690.

490. Dong, X.J., et al., *Wavelets for agriculture and biology: A tutorial with applications and outlook.* Bioscience, 2008. **58**(5): p. 445-453.

491. Cazelles, B., et al., *Wavelet analysis of ecological time series.* Oecologia, 2008. **156**(2): p. 287-304.

492. R Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, 3.3.0, https://www.R-project.org/, 2016.

493. Bourgon, R., *intervals: Tools for Working with Points and Intervals.* R package version 0.15.1, https://CRAN.R-project.org/package=intervals, 2015.

494. Auton, A., *The estimation of recombination rates from population genetic data*. 2007, Oxford University, UK.

495. Sudmant, P.H., et al., *An integrated map of structural variation in 2,504 human genomes.* Nature, 2015. **526**(7571): p. 75-81.

496. Ensembl, *Structural variant: esv3647597*. 2021. [cited 23/05/2021] Available from: http://grch37.ensembl.org/Homo_sapiens/StructuralVariation/Evidence?db=core;r=22:32275157-32638056;sv=esv3647597;svf=50451828;vdb=variation.

497. Jabalameli, M., *Next-Generation Sequencing Analyses in Human Diseases and Population Genomics*, in *Faculty of Medicine*. 2018, University of Southampton, UK.

498. Vergara-Lope, A., et al., *Linkage disequilibrium maps for European and African populations constructed from whole genome sequence data.* Scientific Data, 2019. **6**(1): p. 208.

499. Li, S., C. Schlebusch, and M. Jakobsson, *Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples.* Proceedings of the Royal Society B-Biological Sciences, 2014. **281**(1793): p. 9.

500. Gurdasani, D., et al., *Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa.* Cell, 2019. **179**(4): p. 984-1002.e36.

501. Otto, S.P. and N.H. Barton, *The Evolution of Recombination: Removing the Limits to Natural Selection.* Genetics, 1997. **147**(2): p. 879-906.

502. Rogers, A.R., R.J. Bohlender, and C.D. Huff, *Early history of Neanderthals and Denisovans.* Proceedings of the National Academy of Sciences, 2017. **114**(37): p. 9859-9863.

503. Sirugo, G., S.M. Williams, and S.A. Tishkoff, *The Missing Diversity in Human Genetic Studies.* Cell, 2019. **177**(1): p. 26-31.

504. Payseur, B.A. and M.W. Nachman, *Gene Density and Human Nucleotide Polymorphism.* Molecular Biology and Evolution, 2002. **19**(3): p. 336-340.

505. Cutter, A.D. and B.A. Payseur, *Genomic signatures of selection at linked sites: unifying the disparity among species.* Nature Reviews Genetics, 2013. **14**(4): p. 262-274.

506. Veeramah, K.R., et al., *Evidence for Increased Levels of Positive and Negative Selection on the X Chromosome versus Autosomes in Humans.* Molecular Biology and Evolution, 2014. **31**(9): p. 2267-2282.

507. Spataro, N., et al., *Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology.* Human molecular genetics, 2017. **26**(3): p. 489-500.

508. Haller, B.C. and P.W. Messer, *SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model.* Molecular Biology and Evolution, 2019. **36**(3): p. 632-637.

509. Collins, A. and N.E. Morton, *Mapping a disease locus by allelic association.* Proceedings of the National Academy of Sciences, 1998. **95**(4): p. 1741-1745.

510. Horscroft, C., et al., *zalpha: an R package for the identification of regions of the genome under selection.* The Journal of Open Source Software, 2020. **5**(56).

511. Kaplan, N.L., R.R. Hudson, and C.H. Langley, *The "hitchhiking effect" revisited.* Genetics, 1989. **123**(4): p. 887-899.

512. Delignette-Muller, M.L. and C. Dutang, *fitdistrplus: An R Package for Fitting Distributions.* Journal of Statistical Software, 2015. **64**(4): p. 34.

513. Smithson, M. and J. Verkuilen, *A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables.* Psychological Methods, 2006. **11**(1): p. 54-71.

514. Wickham, H., et al., *roxygen2: In-Line Documentation for R.* R package version 7.1.1, https://CRAN.R-project.org/package=roxygen2, 2020.

515. Wickham, H., J. Hester, and W. Chang, *devtools: Tools to Make Developing R Packages Easier.* R package version 2.3.0, https://CRAN.R-project.org/package=devtools, 2020.

516. Wickham, H., *testthat: Get Started with Testing.* The R Journal, 2011. **3**: p. 5--10.

517. Travis CI, 2020. [cited 08/11/2020] Available from: https://travis-ci.com/.

518. Monniaux, D., *The pitfalls of verifying floating-point computations.* ACM Trans. Program. Lang. Syst., 2008. **30**(3): p. Article 12.

519. Pfeifer, B., et al., *PopGenome: an efficient Swiss army knife for population genomic analyses in R.* Mol Biol Evol, 2014. **31**(7): p. 1929-36.

520. Rozas, J., et al., *DNA Variation at the rp49 Gene Region of Drosophila simulans: Evolutionary Inferences From an Unusual Haplotype Structure.* Genetics, 2001. **158**(3): p. 1147-1155.

521. Gautier, M., A. Klassmann, and R. Vitalis, *rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure.* Molecular Ecology Resources, 2017. **17**(1): p. 78-90.

522. DNA my dog, 2020. [cited 22/09/2020] Available from: https://dnamydog.com/.

523. Embark Veterinary, 2020. [cited 22/09/2020] Available from: https://embarkvet.com/.

524. Orivet, 2020. [cited 22/09/2020] Available from: https://www.orivet.com/.

525. Wisdom Panel, 2020. [cited 22/09/2020] Available from: https://www.wisdompanel.com.

526. Hytönen, M.K. and H. Lohi, *Canine models of human rare disorders.* Rare diseases (Austin, Tex.), 2016. **4**(1): p. e1241362-e1241362.

527. Lindblad-Toh, K., et al., *Genome sequence, comparative analysis and haplotype structure of the domestic dog.* Nature, 2005. **438**(7069): p. 803-819.

528. Vilà, C., et al., *Multiple and Ancient Origins of the Domestic Dog.* Science, 1997. **276**(5319): p. 1687-1689.

529. Wang, G.-D., et al., *Out of southern East Asia: the natural history of domestic dogs across the world.* Cell Research, 2016. **26**(1): p. 21-33.

530. Yang, Q., et al., *Genetic Diversity and Signatures of Selection in 15 Chinese Indigenous Dog Breeds Revealed by Genome-Wide SNPs.* Frontiers in Genetics, 2019. **10**(1174).

531. Parker, H.G., et al., *Genetic Structure of the Purebred Domestic Dog.* Science, 2004. **304**(5674): p. 1160-1164.

532. Honeycutt, R.L., *Unraveling the mysteries of dog evolution.* BMC Biology, 2010. **8**(1): p. 20.

533. Druzhkova, A.S., et al., *Ancient DNA Analysis Affirms the Canid from Altai as a Primitive Dog.* PLOS ONE, 2013. **8**(3): p. e57754.

534. Perri, A., *A wolf in dog's clothing: Initial dog domestication and Pleistocene wolf variation.* Journal of Archaeological Science, 2016. **68**: p. 1-4.

535. Pollinger, J.P., et al., *Selective sweep mapping of genes with large phenotypic effects.* Genome Research, 2005. **15**(12): p. 1809-1819.

536. Marsden, C.D., et al., *Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs.* Proceedings of the National Academy of Sciences, 2016. **113**(1): p. 152-157.

537. Perri, A., et al., *NEW EVIDENCE OF THE EARLIEST DOMESTIC DOGS IN THE AMERICAS.* American Antiquity, 2019. **84**(1): p. 68-87.

538. Leonard, J.A., et al., *Ancient DNA Evidence for Old World Origin of New World Dogs.* Science, 2002. **298**(5598): p. 1613-1616.

539. Morey, D.F. and K. Aaris-Sorensen, *Paleoeskimo dogs of the Eastern Arctic.* Arctic, 2002. **55**(1): p. 44-56.

540. Brown, S.K., et al., *Using multiple markers to elucidate the ancient, historical and modern relationships among North American Arctic dog breeds.* Heredity, 2015. **115**(6): p. 488-495.

541. Castroviejo-Fisher, S., et al., *Vanishing native American dog lineages.* BMC Evolutionary Biology, 2011. **11**(1): p. 73.

542. Parker, H.G., et al., *Genomic Analyses Reveal the Influence of Geographic Origin, Migration, and Hybridization on Modern Dog Breed Development.* Cell Reports, 2017. **19**(4): p. 697-708.

543. Huson, H.J., et al., *A genetic dissection of breed composition and performance enhancement in the Alaskan sled dog.* BMC Genetics, 2010. **11**(1): p. 71.

544. Ní Leathlobhair, M., et al., *The evolutionary history of dogs in the Americas.* Science, 2018. **361**(6397): p. 81-85.

545. Bergström, A., et al., *Origins and genetic legacy of prehistoric dogs.* Science, 2020. **370**(6516): p. 557-564.

546. Anderson, T.M., et al., *Molecular and Evolutionary History of Melanism in North American Gray Wolves.* Science, 2009. **323**(5919): p. 1339-1343.

547. Galov, A., et al., *First evidence of hybridization between golden jackal (Canis aureus) and domestic dog (Canis familiaris) as revealed by genetic markers.* Royal Society Open Science, 2015. **2**(12): p. 150450.

548. Axelsson, E., et al., *Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome.* Genome Research, 2012. **22**(1): p. 51-63.

549. Eizirik, E., et al., *Pattern and timing of diversification of the mammalian order Carnivora inferred from multiple nuclear gene sequences.* Molecular Phylogenetics and Evolution, 2010. **56**(1): p. 49-63.

550. Auton, A., et al., *Genetic recombination is targeted towards gene promoter regions in dogs.* PLoS genetics, 2013. **9**(12): p. e1003984-e1003984.

551. Pengelly, R.J., et al., *Commercial chicken breeds exhibit highly divergent patterns of linkage disequilibrium.* Heredity, 2016. **117**(5): p. 375-382.

552. Auton, A., et al., *A Fine-Scale Chimpanzee Genetic Map from Population Sequencing.* Science, 2012. **336**(6078): p. 193-198.

553. Muñoz-Fuentes, V., et al., *Strong Artificial Selection in Domestic Mammals Did Not Result in an Increased Recombination Rate.* Molecular Biology and Evolution, 2014. **32**(2): p. 510-523.

554. Ross-Ibarra, J., *The Evolution of Recombination under Domestication: A Test of Two Hypotheses.* The American Naturalist, 2004. **163**(1): p. 105-112.

555. Lenormand, T. and S.P. Otto, *The Evolution of Recombination in a Heterogeneous Environment.* Genetics, 2000. **156**(1): p. 423-438.

556. Breen, M., *Canine cytogenetics--from band to basepair.* Cytogenetic and genome research, 2008. **120**(1-2): p. 50-60.

557. Liehr, T., *Chapter 2 - CG-CNVs: What Is the Norm?*, in *Benign & Pathological Chromosomal Imbalances*, T. Liehr, Editor. 2014, Academic Press: Oxford. p. 13-24.

558. Campbell, C.L., et al., *A Pedigree-Based Map of Recombination in the Domestic Dog Genome.* G3: Genes|Genomes|Genetics, 2016. **6**(11): p. 3517-3524.

559. Wang, G.-D., et al., *The genomics of selection in dogs and the parallel evolution between dogs and humans.* Nature Communications, 2013. **4**(1): p. 1860.

560. Miao, B., Z. Wang, and Y. Li, *Genomic Analysis Reveals Hypoxia Adaptation in the Tibetan Mastiff by Introgression of the Gray Wolf from the Tibetan Plateau.* Molecular Biology and Evolution, 2016. **34**(3): p. 734-743.

561. vonHoldt, B., et al., *EPAS1 variants in high altitude Tibetan wolves were selectively introgressed into highland dogs.* PeerJ, 2017. **5**: p. e3522.

562. Axelsson, E., et al., *The genomic signature of dog domestication reveals adaptation to a starch-rich diet.* Nature, 2013. **495**(7441): p. 360-364.

563. Perry, G.H., et al., *Diet and the evolution of human amylase gene copy number variation.* Nature Genetics, 2007. **39**(10): p. 1256-1260.

564. Cagan, A. and T. Blass, *Identification of genomic variants putatively targeted by selection during dog domestication.* BMC Evolutionary Biology, 2016. **16**(1): p. 10.

565. Freedman, A.H., et al., *Demographically-Based Evaluation of Genomic Regions under Selection in Domestic Dogs.* PLOS Genetics, 2016. **12**(3): p. e1005851.

566. vonHoldt, B.M., et al., *Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication.* Nature, 2010. **464**(7290): p. 898-902.

567. Akey, J.M., et al., *Tracking footprints of artificial selection in the dog genome.* Proceedings of the National Academy of Sciences, 2010. **107**(3): p. 1160-1165.

568. Vaysse, A., et al., *Identification of Genomic Regions Associated with Phenotypic Variation between Dog Breeds using Selection Mapping.* PLOS Genetics, 2011. **7**(10): p. e1002316.

569. Zeyland, J., et al., *Complete mitochondrial genome of wild aurochs (Bos primigenius) reconstructed from ancient DNA.* Polish Journal of Veterinary Sciences, 2013. **16**(2): p. 265-273.

570. Deane-Coe, P.E., et al., *Direct-to-consumer DNA testing of 6,000 dogs reveals 98.6-kb duplication associated with blue eyes and heterochromia in Siberian Huskies.* PLOS Genetics, 2018. **14**(10): p. e1007648.

571. Sams, A.J. and A.R. Boyko, *Fine-Scale Resolution of Runs of Homozygosity Reveal Patterns of Inbreeding and Substantial Overlap with Recessive Disease Genotypes in Domestic Dogs.* G3: Genes|Genomes|Genetics, 2019. **9**(1): p. 117-123.

572. Short, A.D., et al., *Hardy–Weinberg Expectations in Canine Breeds: Implications for Genetic Studies.* Journal of Heredity, 2007. **98**(5): p. 445-451.

573. Marees, A., et al., *A tutorial on conducting genome-wide association studies: Quality control and statistical analysis.* International Journal of Methods in Psychiatric Research, 2018. **27**: p. e1608.

574. Schrauwen, I., et al., *Identification of novel genetic risk loci in Maltese dogs with necrotizing meningoencephalitis and evidence of a shared genetic risk across toy dog breeds.* PloS one, 2014. **9**(11): p. e112755-e112755.

575. Epskamp, S., et al., *qgraph: Network Visualizations of Relationships in Psychometric Data.* Journal of Statistical Software, 2012. **48**(4): p. 18.

576. Akey, J.M., et al., *Interrogating a High-Density SNP Map for Signatures of Natural Selection.* Genome Research, 2002. **12**(12): p. 1805-1814.

577. Ostrander, E.A. and L. Kruglyak, *Unleashing the Canine Genome.* Genome Research, 2000. **10**(9): p. 1271-1274.

578. Freedman, A.H., et al., *Genome Sequencing Highlights the Dynamic Early History of Dogs.* PLOS Genetics, 2014. **10**(1): p. e1004016.

579. Turner, S.D., *qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots.* bioRxiv, 2014: p. 005165.

580. Chen, H., *VennDiagram: Generate High-Resolution Venn and Euler Plots.* R package version 1.6.20, https://CRAN.R-project.org/package=VennDiagram, 2018.

581. Kuhn, R.M., D. Haussler, and W.J. Kent, *The UCSC genome browser and associated tools.* Briefings in bioinformatics, 2013. **14**(2): p. 144-161.

582. Boyko, A.R., et al., *A Simple Genetic Architecture Underlies Morphological Variation in Dogs.* PLOS Biology, 2010. **8**(8): p. e1000451.

583. McLaren, W., et al., *The Ensembl Variant Effect Predictor.* Genome Biology, 2016. **17**(1): p. 122.

584. Zhu, L.J., et al., *ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data.* BMC Bioinformatics, 2010. **11**(1): p. 237.

585. Durinck, S., et al., *Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.* Nat Protoc, 2009. **4**(8): p. 1184-91.

586. Hahne, F. and R. Ivanek, *Visualizing Genomic Data Using Gviz and Bioconductor*, in *Statistical Genomics: Methods and Protocols*, E. Mathé and S. Davis, Editors. 2016, Springer New York: New York, NY. p. 335-351.

587. Morgan, M. and L. Shepherd, *AnnotationHub: Client to access AnnotationHub resources.* R package version 2.20.1, https://bioconductor.org/packages/release/bioc/html/AnnotationHub.html, 2020.

588. Rainer, J., L. Gatto, and C.X. Weichenberger, *ensembldb: an R package to create and use Ensembl-based annotation resources.* Bioinformatics, 2019. **35**(17): p. 3151-3153.

589. Wong, A.K., et al., *A Comprehensive Linkage Map of the Dog Genome.* Genetics, 2010. **184**(2): p. 595-605.

590. Kieleczawa, J., *Fundamentals of sequencing of difficult templates--an overview.* Journal of biomolecular techniques : JBT, 2006. **17**(3): p. 207-217.

591. Doucet, A.J., et al., *U6 snRNA Pseudogenes: Markers of Retrotransposition Dynamics in Mammals.* Molecular Biology and Evolution, 2015. **32**(7): p. 1815-1832.

592. Brow, D.A. and C. Guthrie, *Spliceosomal RNA U6 is remarkably conserved from yeast to mammals.* Nature, 1988. **334**(6179): p. 213-218.

593. Leroy, G., et al., *Methods to estimate effective population size using pedigree data: Examples in dog, sheep, cattle and horse.* Genetics Selection Evolution, 2013. **45**(1): p. 1.

594. Calboli, F.C.F., et al., *Population Structure and Inbreeding From Pedigree Analysis of Purebred Dogs.* Genetics, 2008. **179**(1): p. 593-601.

595. Rimbault, M. and E.A. Ostrander, *So many doggone traits: mapping genetics of multiple phenotypes in the domestic dog.* Human molecular genetics, 2012. **21**(R1): p. R52-R57.

596. Plassais, J., et al., *Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology.* Nature Communications, 2019. **10**(1): p. 1489.

597. Sutter, N.B., et al., *A Single IGF1 Allele Is a Major Determinant of Small Size in Dogs.* Science, 2007. **316**(5821): p. 112-115.

598. Cadieu, E., et al., *Coat Variation in the Domestic Dog Is Governed by Variants in Three Genes.* Science, 2009. **326**(5949): p. 150-153.

599. Nichols, B.L., et al., *The maltase-glucoamylase gene: Common ancestry to sucrase-isomaltase with complementary starch digestion activities.* Proceedings of the National Academy of Sciences, 2003. **100**(3): p. 1432-1437.

600. Lee, M.-H., et al., *Identification of a gene, ABCG5, important in the regulation of dietary cholesterol absorption.* Nature Genetics, 2001. **27**(1): p. 79-83.

601. Tang, K., K.R. Thornton, and M. Stoneking, *A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome.* PLOS Biology, 2007. **5**(7): p. e171.

602. Rao, Y., et al., *The influence of recombination on SNP diversity in chickens.* Hereditas, 2011. **148**(2): p. 63-69.

603. Nachman, M.W., *Single nucleotide polymorphisms and recombination rate in humans.* Trends in Genetics, 2001. **17**(9): p. 481-485.

604. Varela, M.A. and W. Amos, *Heterogeneous distribution of SNPs in the human genome: Microsatellites as predictors of nucleotide diversity and divergence.* Genomics, 2010. **95**(3): p. 151-159.

605. Charlesworth, B., *Measures of divergence between populations and the effect of forces that reduce variability.* Molecular Biology and Evolution, 1998. **15**(5): p. 538-543.

606. Cruickshank, T.E. and M.W. Hahn, *Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow.* Molecular Ecology, 2014. **23**(13): p. 3133-3157.

607. Neff, M.W. and J. Rine, *A Fetching Model Organism.* Cell, 2006. **124**(2): p. 229-231.