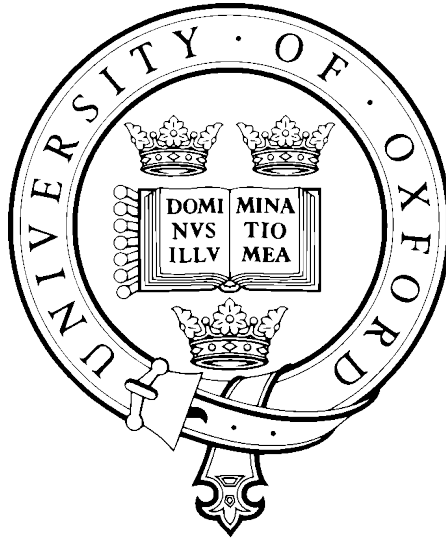


A Bayesian Chromosome Painting Approach to Detect
Signals of Incomplete Positive Selection in Sequence Data:
applications to 1000 Genomes



Christopher Thomas Gamble

Keble College

Department of Statistics

Wellcome Trust Centre for Human Genetics

University of Oxford

Supervised by
Professor Chris Holmes
Dr Chris Spencer

*A Thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy*

Michaelmas Term 2013

Abstract

Methods to detect patterns of variation associated with ongoing positive selection often focus on identifying regions of the genome with extended haplotype homozygosity - indicative of recently shared ancestry. Whilst these have been shown to be powerful they have two major challenges.

First, these methods are constructed to detect variation associated with a classical selective sweep; a single haplotype background gets swept up to a higher than expected frequency given its age. Recently studies have shown that other forms of positive selection, e.g. selection on standing variation, may be more prevalent than previous thought. Under such evolution, a mutation that is already segregating in the population becomes beneficial, possibly as a result of an environmental change. The second challenge with these methods is that they base their inference on non-parametric tests of significance which can result in uncontrolled false positive rates.

We tackle these problems using two approaches. First, by exploiting a widely used model in population genomics we construct a new approach to detect regions where a subset of the chromosomes are much more related than expected genome-wide. Using this metric we show that it is sensitive to both classical selective sweeps, and to soft selective sweeps, e.g. selection on standing variation. Second, building on existing methods, we construct a Bayesian test which bi-partitions chromosomes at every position based on their allelic type and tests for association between chromosomes carrying one allele and significantly reduced time to common ancestor.

Using simulated data we show that this approach results in a powerful, fast, and robust approach to detect signals of positive selection in sequence data. Moreover by comparing our model to existing techniques we show that we have similar power to detect recent classical selective sweeps, and considerably greater power to detect soft selective sweeps.

We apply our method, ABACUS, to three human populations using data from the 1000 Genome Project. Using existing and novel candidates of positive selection, we show that the results between ABACUS and existing methods are comparable in regions of classical selection, and are arguably superior in regions that show evidence for recent selection on standing variation.

Contents

List of Figures	v
-----------------	---

List of Tables	vii
----------------	-----

1 Models in Population Genomics	1
1.1 Population Genetics	3
1.1.1 Kingman’s Coalescent	6
1.1.2 The Coalescent with Recombination	8
1.1.3 The Coalescent with Variable Population Size	11
1.1.4 Coalescent Inference	13
1.2 Natural Selection	14
1.2.1 A Model of Classical Selective Sweeps	18
1.2.2 Effect on genetic variation	20
1.3 Tools for Detecting Classical Sweeps	23
1.3.1 Departures from Neutrality	24
1.3.2 Functional Tests	27
1.3.3 Likelihood-based Approaches	27
1.3.4 Genome Wide Selection Scans	28
1.4 Motivation	32
1.5 Outline	34
2 Classical Approaches to Detect Selective Sweeps	36
2.1 Methods	37
2.1.1 Data Generation	38
2.1.2 Machine Learning Tools	39
2.1.3 Approximate Bayesian Computation	40
2.1.3.1 Choice of Inference Algorithm	42
2.1.3.2 Automatic Choice Dimensionality Reduction	44
2.2 Effect of Selection and Confounding	45
2.2.1 Window v.s. Region	46
2.2.2 Segregating Sites	47
2.2.3 Rare Variants	50
2.3 Predicting σ - Machine Learning	53
2.4 Predicting σ - ABC	56
2.5 Discussion	59

3	A Bayesian Chromosome Painting Approach To Detecting Ge-	62
	nomomic Regions Under Selection	
3.1	Introduction	62
3.2	HMMs in Population Genetics	64
3.2.1	LSHMM	64
3.2.2	HMM Inference	66
3.2.2.1	Forward / Backward Algorithm	67
3.2.2.2	Viterbi Algorithm	69
3.2.2.3	Parameter Inference	69
3.2.3	Examples of HMMs in Population Genetics	70
3.3	Chromosome Painting	71
3.3.1	Local Ancestry	73
3.3.2	Natural Selection	78
3.4	A Bayesian Test of Selection	82
3.4.1	Long Lengths	82
3.4.2	Description of Posterior Distribution	86
3.4.3	A Test of Association	89
3.4.4	Laplace Approximation of Marginal Likelihood	91
3.5	Discussion	92
4	Model Validation and Power to Detect Selective Sweeps	95
4.1	Methods	96
4.1.1	Data Generation	96
4.1.2	Estimating power	98
4.1.3	Calculating metrics in simulated data.	100
4.2	Metric characteristics under neutrality	101
4.2.1	Effect of Demographic Changes	104
4.2.2	Effect of Population Stucture	106
4.2.3	Effect of Recombination Hot Spots	109
4.3	Power to detect ongoing selective sweeps	109
4.3.1	Classical sweeps	113
4.3.2	Soft Sweep - Recurrent Mutations	115
4.3.3	Soft Sweep - Standing Variation	118
4.4	Effect of Sample Size	121
4.5	Discussion	124
5	An Application to 1000 Genome	128
5.1	Introduction	128
5.2	Data	129
5.2.1	Method Implementations	132
5.2.2	Comparing ABACUS and iHS	133
5.3	Existing Candiates for Positive Selection	133
5.3.1	Signals in Six Candiate Regions	136
5.3.1.1	CEU Candidates	137

5.3.1.2	CHB Candidates	145
5.3.1.3	YRI Candidate	149
5.3.2	Summary	150
5.3.3	Overlap with Previous Studies	151
5.4	Novel Candidates for Positive Selection	154
5.5	Discussion	160
6	Conclusion and Future Direction	162
6.1	Conclusions	162
6.1.1	Exact model, approximate inference	162
6.1.2	Approximate model, exact inference	164
6.1.3	Power to Detect Selection	165
6.1.4	New candidates	167
6.2	Future work	169
6.2.1	Alternative inference methods	169
6.2.2	New applications	171
6.2.3	A final word	172
A	Laplace Approximation Equations	173
B	Simulations	176
B.1	Data Generation	176
B.1.1	Classical Sweep	176
B.1.1.1	Null	176
B.1.1.2	Alternative	177
B.1.2	Soft Sweep	177
B.1.2.1	Null	177
B.1.2.2	Alternative	177
B.2	Calculating ABACUS	178
B.3	Calculating uHS	178
B.4	Calculating H_{12}	178
C	ABACUS Implementation	180
D	ABACUS and iHS Results Tables	182
E	Existing Candidate Region Analysis	185
F	Novel Candidate Region Analysis	191
	Bibliography	195

List of Figures

1.1	Example of the Wright Fisher model	4
1.2	Wright Fisher model and Kingman's coalescent	6
1.3	Example of Coalescent and Haplotype Structure	7
1.4	Schematic of Eukaryotic Recombination	9
1.5	Example of Ancestral Recombination Graph	10
1.6	Two models of Demography	12
1.7	Interpolated map of Old World Lactose Persistence frequencies . . .	16
1.8	Correlation Between Malaria Endemics and SCA in Africa	17
1.9	Coop's Model of a Classical Selective Sweep	19
1.10	A Schematic Highlighting Alternative Models of Positive Selection .	21
1.11	Impact of positive selection in simulated variation data	22
1.12	Extended Haplotype Homozygosity in <i>LCT</i>	30
2.1	Effect of Selection and Demography on Watterson's Estimate	48
2.2	Effect of Selection and Demography on Tajima's D	49
2.3	Effect of Selection and Demography on D^*	51
2.4	Effect of Selection and Demography on H	52
2.5	Correlation Matrix of Summary Statistics.	54
2.6	Three Models for Predicting the Strength of Selection.	55
2.7	ABC Density Plots	58
3.1	Hidden Markov Model Schematic	65
3.2	Illustration of chromosome painting	72
3.3	Chromosome Painting and Recent Ancestry	75
3.4	Chromosome Painting and Recombination Schematic	77
3.5	Lengths of Most Like Painting vs Coalescent Times	78
3.6	Schematic Representation of Model	79
3.7	Effect of σ on Lengths of Painted Chunks	80
3.8	Test of Exponential Assumption	84
3.9	Comparison of Posterior Probability and Average Length	88
4.1	Power Schematic	99
4.2	Distribution of Metrics under neutral evolution	102
4.3	Comparison of ABACUS under Different Demographic Scenarios . .	105
4.4	Distribution of Metrics under neutral evolution	106
4.5	Distribution of ABACUS under population structure	108

4.6	Comparison of ABACUS under Recombination Hotspot Scenarios .	110
4.7	Example of iHS on Classical and Soft Sweeps	111
4.8	Power to Detect Hard Sweeps	114
4.9	ROC Curves for Hard Sweeps	115
4.10	Power to Detect Soft Sweeps - Recurrent Mutation	116
4.11	ROC Curves for Soft Sweeps - Recurrent Mutation	117
4.12	Power Curves for Soft Sweeps - Standing Variation	119
4.13	ROC Curves for Soft Sweeps - Standing Variation	120
4.14	Impact of Size on Power	123
5.1	Comparison of Allele Frequency Spectrum for CEU, CHB, and YRI	130
5.2	Genome-wide evidence for positive selection.	135
5.3	Distribution of BF by Derived Frequency	136
5.4	Variation analysis of <i>LCT</i> locus	138
5.5	Diagnostic plot for <i>LCT</i> locus on chromosome 2	140
5.6	Variation analysis of <i>ASPM</i> locus	142
5.7	Variation analysis of <i>TRPV6</i> locus	144
5.8	Variation analysis of <i>ADH1B</i> locus	146
5.9	Variation analysis of <i>EDAR</i> locus	148
5.10	Variation analysis of <i>PSCA</i> locus	149
5.11	Comparison of iHS and ABACUS in CEU	153
5.12	Variation analysis of <i>DPY19L2</i> locus	156
5.13	Variation analysis of <i>FADS</i> locus	157
5.14	Variation analysis of <i>IGH</i> locus	159
C.1	ABACUS Pipeline	181
E.1	Diagnostic plot for <i>ASPM</i> on chromosome 1	186
E.2	Diagnostic plot for <i>TRPV6</i> on chromosome 7	187
E.3	Diagnostic plot for <i>ADH1B</i> on chromosome 4	188
E.4	Diagnostic plot for <i>EDAR</i> on chromosome 2	189
E.5	Diagnostic plot for <i>PSCA</i> on chromosome 8	190
F.1	Diagnostic plot for <i>DPY19L2</i> on chromosome 12	192
F.2	Diagnostic plot for <i>FADS</i> on chromosome 11	193
F.3	Diagnostic plot for <i>IGH</i> on chromosome 14	194

List of Tables

2.1	Simulation Summary	45
2.2	Table of Metrics Used in ML Models	53
5.1	1000 Genome Subpopulations	129
5.2	SNP Characteristics in CEU, YRI, CHB	132
5.3	Processing Time of ABACUS	133
5.4	Candidates from Peter et al. [2012]	137
5.5	Summary of Results in Six Candidates from Peter et al. [2012] . . .	151
5.6	Existing Genome-wide Scans of Selection	152
D.1	Top 10 ABACUS regions in CEU, YRI, and CHB	183
D.2	Top 10 iHS regions in CEU, YRI, and CHB	184

Acknowledgements

This work has been supported by studentship awarded by the Engineering and Physical Sciences Research Council. I am grateful to the University of Oxford, The Wellcome Trust, the Department of Statistics, and Keble College for providing the means for me to complete this work and develop my research skills.

I would like to acknowledge the use of the Oxford Advanced Research Computing (ARC) services at the University of Oxford, and the IRIDIS High Performance Computing Facility at the University of Southampton, in the completion of this work.

My thanks go to all my incredible friends and colleagues from Keble MCR, Keble College Boat Club, and Oxford University Lightweight Rowing Club for all their support and guidance during my time at Oxford. Also thanks to all my friends from Warwick University for their continual reminder of what is important in life.

Particular thanks go to my examiners Dr Simon Myers and Dr Garrett Hellenthal from University College London for their insightful comments and useful advice.

Also, my thanks go to all members of the Donnelly, Holmes, McVean, Myers, and Spencer groups for both their academic and personal support over the past three years. In particular Hilary Martin, Davis McCarthy, Robbie Davies, Oliver Venn, Gavin Band, Si Li Quang, George Busby, Maddy Mitchell, James Watson, Aimee Taylor, Tristan Gray-Davies, and Chris Yau for their support both inside and outside the office.

A special thanks to my supervisors Chris Holmes and Chris Spencer, for their patience, guidance, and support. I could not have asked for two better supervisors.

I would like to thank my family; Tina, Billy, Pat, Malcolm, William, Lily, Steven, Julie, Charlotte, and Rachel, for giving me support and guidance throughout my long years in higher education.

And finally to Laurie. I cannot thank you enough for all that you have done for me. Despite all the challenges that we have faced your support and guidance is what has gotten me to this point.

For mum.

Chapter 1

Models in Population Genomics

“The natural history of these [Galapagos] islands is eminently curious, and well deserves attention.”

Charles Robert Darwin, FRS

In August 1858, Charles Darwin and Alfred Wallace submitted an essay to the Linnean Society of London, [Darwin and Wallace, 1858], describing their joint discovery of the theory of evolution by means of natural selection. One year later Darwin extensively detailed their theory in his book, *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* [Darwin, 1859]. In this book he discusses how beneficial traits, over time, will become more frequent in the population, as individuals with these traits are more likely to live longer and pass the mutations conferring them to their progeny.

After Darwin published *On the Origin of the Species*, Galton [1869] proposed an alternative theory of discontinuous evolution produced by natural selection acting on large discontinuous variation, and for much of the latter half of the 19th Century the two theories of evolution were considered at odds to one another [Provine, 1987]. The development of the field of population genetics lead to the

synthesis of the two competing schools of thought [Provine, 1987]. Early 20th century mathematicians, Wright, Fisher, and Haldane, provided the theoretical foundations of this field between 1918 and 1932, culminating in their description of a process by which generations pass information to the next; the Wright Fisher (WF) model [Wright, 1931].

The latter half of the 20th century saw research efforts in population genetics focus on quantifying aspects of the WF model; [Kimura and Crow, 1964, Kimura, 1969, Jukes and Cantor, 1969, Watterson, 1975]. John Kingman noted, however, that the prospective nature of the WF model can result in an inefficient description of the evolution of a contemporary sample of chromosomes [Kingman, 1982a]. This, in part, motivated him to develop the coalescent model which describes the evolution of a sample of chromosomes retrospectively using the WF model.

Despite extending the coalescent to allow for more realistic evolutionary scenarios [Griffiths, 1981, Hudson, 1983, Griffiths and Marjoram, 1996, Krone and Neuhauser, 1997, Coop and Griffiths, 2004], the intractable nature of the model means that full likelihood inference on large-scale datasets is infeasible. To help alleviate some of these challenges alternative methods were proposed to make scaleable inference in modern genome-wide variation datasets, e.g. Li and Stephens [2003]. This has become increasingly important as the scale of readily obtainable data is now genome-wide [Kandpal et al., 2009].

In recent years, improvements in DNA sequencing technologies has lead to a significant increase in the speed and accuracy of data collection, at the same time reducing the cost. To put this in perspective, a first draft of an assembled map of the whole human genome was completed in 2001 by the International Human Genome Sequencing Consortium [Lander et al., 2001]; the project took 10 years to complete and cost a total of \$3 billion. In contrast, using a contemporary

Illumina machine, it is now possible to sequence five human genomes in approximately one week, for less than \$5,000 per genome (<http://www.illumina.com/>). These improvements have allowed geneticists to use Whole Genome Sequencing (WGS) to understand the role of rare variants and their effect on common diseases [Gonzalez-Angulo et al., 2010, Zhang et al., 2011, Novelli et al., 2010, Cirulli and Goldstein, 2010].

The advent of genomics has given geneticists and biologists the information to begin to answer some of the most pressing questions in their areas. Not only in genetics, but in increasingly more diverse areas, the role of a statistician in recent years has moved towards constructing principled methods that are capable of making accurate, efficient inference that can scale to so called ‘Big Data’ problems. Throughout this thesis we will come across a number of challenges in which we attempt to provide a solution, but as we shall see, there are still many open questions.

We begin this chapter by discussing, in more detail, a number of important population genetics models, some of which we mentioned above. This will lead to discussion of the different types of positive selection and the models used to describe them. We then look at how the problem of inferring signals of positive selection has been tackled in the past. Using this discussion we motivate the rest of this thesis in which we focus on attempting to construct a scalable technique for detecting positive selection in genome wide sequence data.

1.1 Population Genetics

Population genetics is believed by many to have been responsible for the synthesis of Mendelian [Galton, 1869] and Darwinian [Darwin, 1859] evolution [Provine, 1987]. Between 1918 and 1932, Haldane, Wright and Fisher published a series

of papers detailing various models which describe the evolution of organisms. As noted by Provine [1987], their work accounted for the differences between the two schools of thought and culminated in a simple model of populations, describing the genealogical relationship among individuals known as the Wright Fisher (WF) model [Wright, 1931].

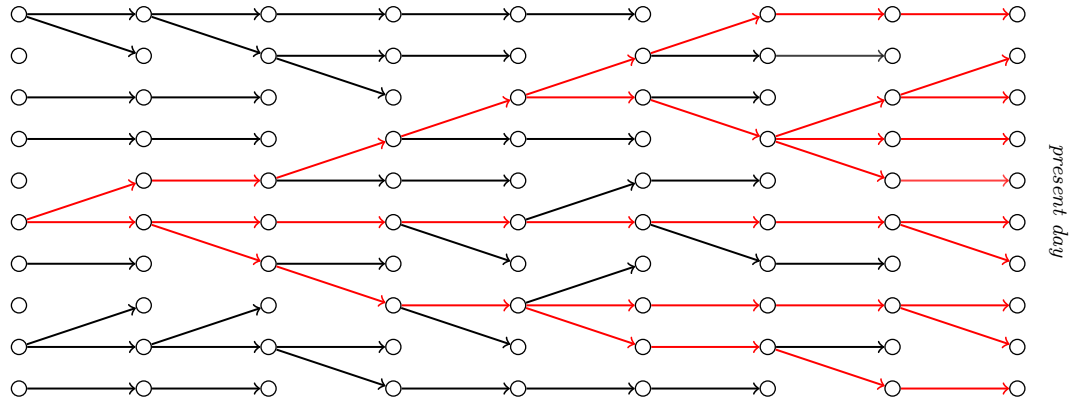


FIGURE 1.1: An example of the Wright Fisher model with nine generations. The arrows indicate the transmission of information from one individual in a particular generation to another individual in the next. The red lines indicated a lineage that has survived to the current population on the right hand side.

The WF model of evolution formalises how neutral traits are passed from one generation to the next, and forms the basis for many models in modern population genetics; see references therein [Hein et al., 2005]. It describes the evolution of an idealised population as a prospective, finite-state Markov chain. Initially, six assumptions were made in order to simplify the model.

First, each state in the Markov chain represents a **discrete generation**, i.e. reproduction and death are simultaneous and occur synchronously across the population. Alternative models have been developed to avoid this assumption, notably the Moran model [Moran, 1958], however both approaches typically result in probabilistically similar genealogies. The second assumption is that each chromosome comes from a **haploid population**. Fortunately that has little quantitative impact and in practice it is straightforward to relax this assumption [Hein et al., 2005]. These assumptions tend not to limit the application of the WF model to

real data, however the other four assumptions do have a significant impact on the distributions of genealogies which can limit its application.

- **Fixed population size.** The probabilities in the model will change when there is a change in the size of the population. In turn this will impact the distribution of allele frequencies.
- **Homogenous fitness.** Increased fitness of an allele increases the probability that that allele gets passed to the next generation. This assumption ensures that the genotype doesn't affect the probability, i.e. each genotype has the same probability of being passed to the next generation.
- **Random mating.** The assumption of uniform random mating does not hold when subpopulation structure is present in the population of chromosomes.
- **No recombination.** Genetic crossovers affect the genealogy of a chromosome, and hence the tree at one position may be different from another position.

A schematic of the WF model can be seen in Figure 1.1. In this example, we observe the evolution of ten chromosomes over nine generations. At each generation or state, members of the homogenous population copy a member of the previous generation with replacement, which can occur imperfectly by introducing mutations drawn at random from a geometric probability distribution.

Using the WF model to perform inference can be inefficient due to its prospective nature [Hein et al., 2005]; it is not efficient to model individuals who do not have direct descendants in the current population. The process by which a genealogy is generated results in many branches not surviving to the present day. As observed in Figure 1.1 it is only the red lineages that survive to the current day, all the black lineages died out at some point in the past.

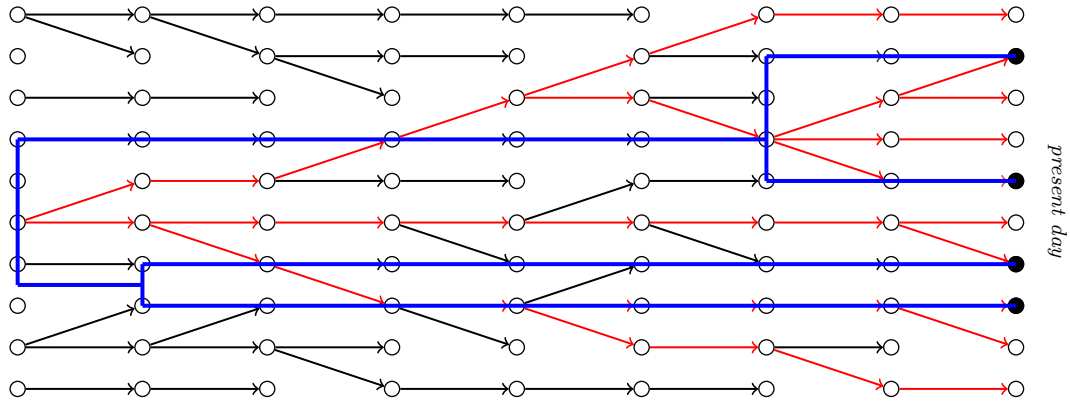


FIGURE 1.2: An Example of the Wright Fisher model with nine generations. Retrospectively the lineages that are relevant to the contemporary population are highlight in red. The black circles indicate when a member of the current population is ‘observed’. The blue tree shows how the lineages of the contemporary sample coalesce back to a common ancestor.

In 1982 the inference approach moved toward the behaviour of the process in reverse time, allowing geneticists to focus on the genealogical tree relating a sample of the current day population. Kingman [Kingman, 1982a,b] showed that in the limit of population size this results in a process that he called the coalescent.

Figure 1.2 shows how Kingman’s coalescent relates to the WF model. Suppose the current day population consists of ten chromosomes, and we take a sample of size four. The lineage relating these chromosomes results in a tree. The point at which the tree joins or coalesces is the point in time at which the two child lineages find a common ancestor, as shown by the blue tree in Figure 1.2.

1.1.1 Kingman’s Coalescent

Kingman’s coalescent [Kingman, 1982a] describes a retrospective, continuous time genealogy of a sample of chromosomes from a large population. Because the foundation of the coalescent is based on the WF model, it originally had many of the same assumptions: no recombination, no selection and no population structure [Kingman, 1982a]. Much like the WF model this made it a somewhat idealised

description of the evolution of a sample of chromosomes. Over time, however, a considerable body of work was developed to provided a theoretical basis for extending the model to more realistic evolutionary scenarios: recombination [Griffiths, 1981, Hudson, 1983], selection [Krone and Neuhauser, 1997, Coop and Griffiths, 2004], and demography [Slatkin and Hudson, 1991, Griffiths and Marjoram, 1996].

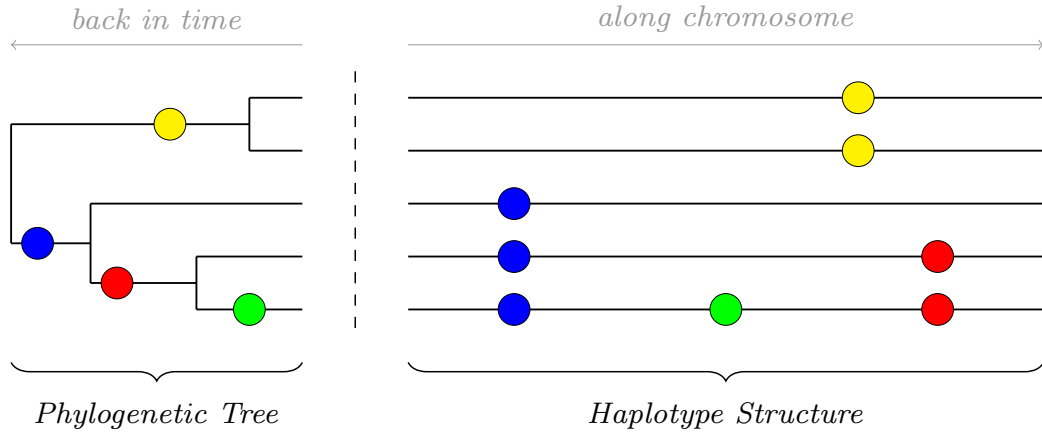


FIGURE 1.3: Example of Coalescent and Haplotype Structure. The left panel shows an example of the a coalescent tree with four mutations. The right panel shows a example of the variation that might be produced by the coalescent on the left side.

Figure 1.3 shows a simple schematic of a coalescent genealogy relating five chromosomes, represented by the five horizontal lines on the right panel. On the left panel the coloured nodes indicate different mutations that have occurred in the past. Their impact on the variation can be seen on the right.

For n individuals, Kingman [1982a] modelled the time to the first coalescence $t^{(1)}$ as having an the exponential distribution with rate $\binom{n}{2}$. This comes from the fact that in WF model the time for lineages to coalesce is modelled as a Geometric distribution, i.e. a waiting time model. As with the WF model, due to the assumption of no population structure, two individuals are chosen at random from the sample, i.e. there are $\binom{n}{2}$ possibilities. These two individuals or lineages are then coalesced into one lineage and the process continues with the sample size reduced by one. The algorithm stops when there is only one lineage left. The time

of this event is known as the time to most recent common ancestor, $T_{MRC A}$ [Hein et al., 2005].

The assumption of no selection allows one to separate the coalescent process from the mutation process, i.e. the allelic state does not affect the probability that an individual or lineage will have progeny [Hartl and Clark, 2007]. The implication of this de-coupling significantly reduces the complexity of the model, allowing one to construct a tree and randomly distribute mutations along it to determine the corresponding chromosomes [Hein et al., 2005].

There are three commonly used mutation models; infinite alleles [Kimura and Crow, 1964], infinite sites [Kimura, 1969], and finite sites [Jukes and Cantor, 1969]. Each is parameterised by the mutation rate, $\theta = 4N_e\mu$; where μ represents the probability from one generation to the next of a change of allele at a single loci, and N_e represents the effective population size from which the samples are drawn.

In an infinite allele mutation model, each mutation creates a new, previously unobserved, allele. Contrast this with the infinite sites model, where we assume that each mutation occurs at a random position in a continuous interval, resulting in no repeat mutations. Finally, in a finite sites model, which is most akin to a observed genetic sequence, a sequence of nucleotides are mutated and transition to a new state via a probability transition model; JC69 [Jukes and Cantor, 1969], K80 [Kimura, 1980], F81 [Felsenstein, 1981].

Many of the methods that we discuss later in this chapter are based on distributional assumptions formulated using the infinite sites model. Since these assumptions tend only to be valid in Kingman's model of the coalescent, the presence of evolutionary processes like demography can alter the allele frequency distribution, thus reducing their accuracy.

1.1.2 The Coalescent with Recombination

Genetic recombination is a fundamental process that occurs in the majority of species [Hartl and Clark, 2007]. As noted by Hein et al. [2005], the process by which genetic information is exchanged differs between eukaryotes, bacteria and viruses, so we will focus in this section on the general case. In eukaryotes, recombination typically occurs during meiosis where parts of the maternal DNA are exchanged with the paternal DNA to construct a new sequence [Hein et al., 2005].

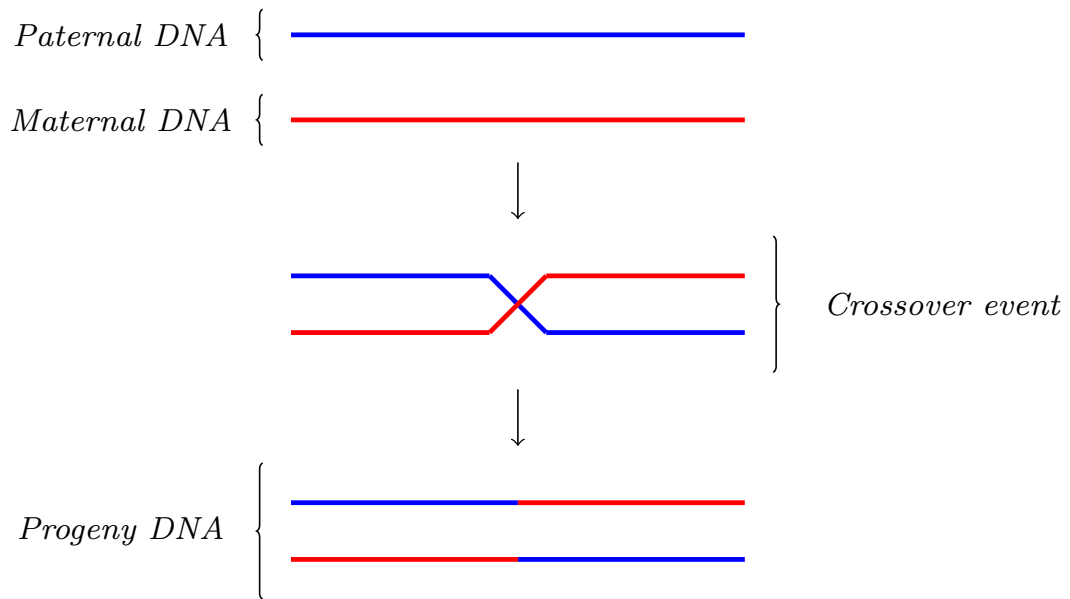


FIGURE 1.4: A schematic showing the effect of a recombination during meiosis on the DNA. The paternal and maternal DNA is crossed to produce a new progenic genome.

Calculating the rate of recombination and understanding how it varies across the human genome is considered to be one of the most important challenges in modern population genomics [Hartl and Clark, 2007], and in recent years many recombination maps have been constructed [Kong et al., 2002, McVean et al., 2004, Myers et al., 2005, Hinch et al., 2011]. One of the main conclusions from these studies is that recombination does not occur uniformly across the genome but tends to be concentrated in specific regions known as recombination hotspots.

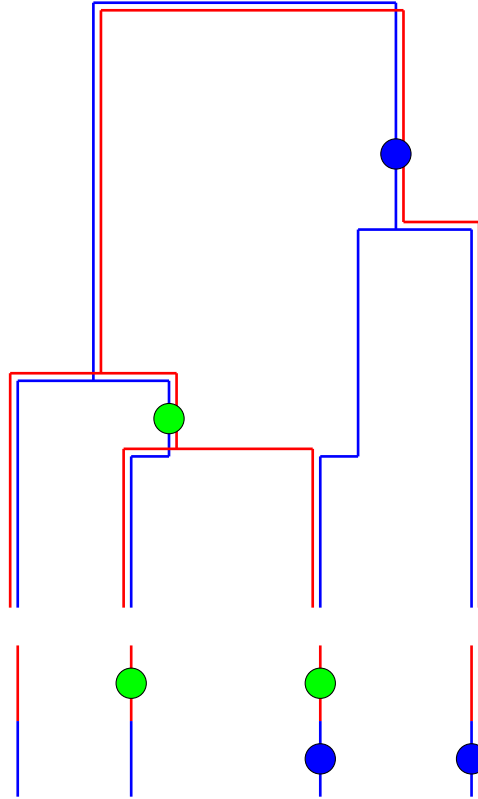


FIGURE 1.5: Simple schematic of Ancestral Recombination Graph showing the ancestral relationships of four sequences with one crossover event. The blue tree represents the ancestry of the four sequences in the blue part of their variation. Likewise for the red tree. The circles indicate mutations and sequences below each tree branch depict the mutations that they carry.

The absence of recombination was one of the original assumptions in Kingman's coalescent and it was one of the first to be relaxed using the model termed the Ancestral Recombination Graph (ARG) [Griffiths, 1981, Hudson, 1983]. Figure 1.4 shows the effect of a single recombination event on the ancestry of diploid progeny. The phylogeny relating the blue part of the genome comes from the paternal phylogeny and the phylogeny relating the red part comes from the maternal phylogeny. As a result there will be two different ancestral relationships for the progeny, one for the red half of the chromosome and another for the blue.

An example of the ARG is shown in Figure 1.5; the red part of the chromosome is related by the red coalescent, likewise for the blue. The construction of the ARG means that these two phylogenies can be related in one model. One of the

advantages of this model, like the coalescent, is that it is a retrospective model that maintains the independence of the coalescent and mutation process [Hein et al., 2005]. The ARG captures the effect of recombination by altering the dependence between nucleotide positions in the chromosome. If the recombination rate is higher, there are more coalescent trees along the genome and thus the dependence between nucleotide positions, known as linkage disequilibrium (LD) will decrease [Hartl and Clark, 2007].

It was shown by Hill and Robertson [1968] for a pair of loci (i, j) , with joint allele frequencies $f_{a,b}$ that the linkage disequilibrium can be measured by the following expression:

$$R^2 = \frac{(f_{1,1} - f_{.,1}f_{1,.})^2}{f_{0,.}f_{1,.}f_{.,0}f_{.,1}}. \quad (1.1)$$

The $f_{.,a}$ notation denotes a marginalisation over the other loci, i.e. the frequency of that allele at that loci. McVean [2002] showed how this quantity, in expectation, can be approximately expressed as a function of the recombination rate, $\rho = 4N_e r$, where r is the probability of a cross-over event between two lineages at a single loci.

The algorithm to construct an ARG is considerably more involved than the neutral coalescent, and its inclusion does not add value to this discussion. Suffice to say that in order to incorporate different coalescent trees in one model the ARG introduces a retrospective branching processes, so rather than n branches coalescing and monotonically falling to a single branch, the ARG can introduce branches to account for the changing ancestry [Hartl and Clark, 2007].

1.1.3 The Coalescent with Variable Population Size

The assumption of constant size in the original WF model can limit the application of it and its descendants to real data. Indeed, as we shall see it can often be difficult to separate patterns of variation associated with varying population size and with positive selection. Figures 1.6(A) & 1.6(B) show two commonly used demographic models in population genetics; the exponential expansion, and the bottleneck.

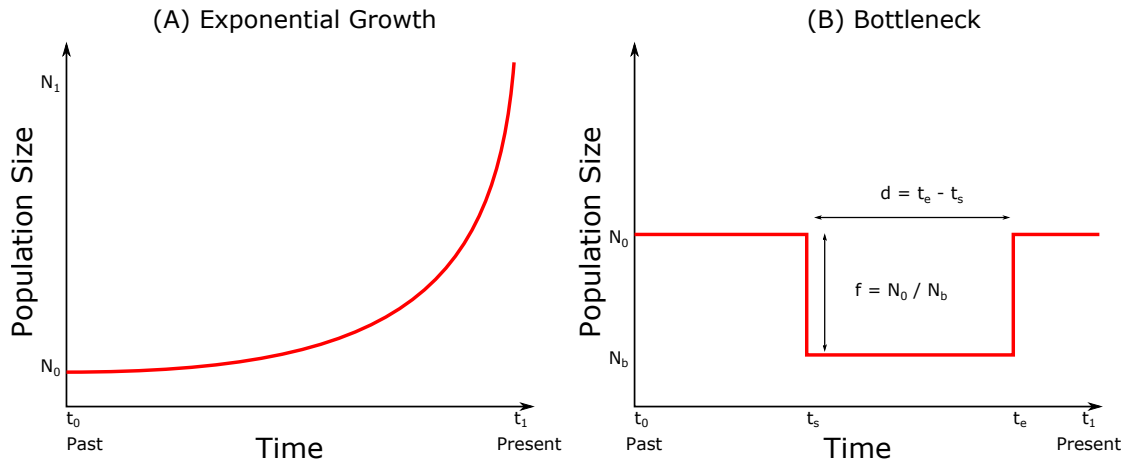


FIGURE 1.6: Schematic representing two models of demography. **(A)** Population experiencing a exponential growth from a size of N_0 at t_0 , to N_1 at t_1 . The growth parameter, α , can be calculated using $N_t = N_0 \exp(\alpha(t - t_0))$. **(B)** Population experiencing a bottleneck between t_s and t_e , were the population size dropped from N_0 to N_b .

If we assume an absence of recombination, the rate of coalescent events in demographic models are influenced by the instantaneous population size relative to the current day population. As a result, in the exponential expansion we see that the rate of coalescent events decrease further back in time resulting in a star-like phylogeny [Hein et al., 2005]. We can see the effects of such expansion in human populations, e.g. European ancestry [Gutenkunst et al., 2009].

The bottleneck model describes the evolution of a population experiencing an abrupt decrease in population size lasting for a given period of time. This could be as a result of an environmental catastrophe, e.g. ice age in populations of

non African ancestry [1000 Genomes Project Consortium et al., 2012]. During the bottleneck the rate of coalescent events will increase due to the reduction in population size. As a result, depending on the age of the bottleneck one might expect to observe a reduction in low frequency mutations relative to a population experiencing a neutral, constant size evolution.

We will see in Chapter 2 how a number of statistics that are used to detect patterns of variation associated with positive selection are also confounded with non constant population sizes.

1.1.4 Coalescent Inference

As noted by Hein et al. [2005], the fact that the mutation and coalescent processes can be separated in Kingman's coalescent significantly simplifies the model. Unfortunately, whilst the model is simple to describe, inference is not as straightforward. One of the major reasons for this intractability is the space of phylogenetic trees. To perform any inference, whether frequentist or Bayesian, our first approach would be to construct a likelihood. In the coalescent, to make inference about the evolutionary variables from a sample of chromosomes requires us to integrate over all possible trees that could have generated this sample. Define H to be the observed patterns of variation, and G to be the unobserved genealogy taking values in the space of topologies, $\mathcal{G}$. Then the likelihood for the mutation rate θ , and recombination rate ρ can be expressed as follows,

$$\mathcal{L}(\rho, \theta | H) = \int_{\mathcal{G}} \mathbb{P}(H | G, \rho, \theta) \mathbb{P}(G | \rho, \theta) dG. \quad (1.2)$$

Suppose there are n chromosomes in the sample. Then at the $k + 1^{th}$ coalescent event there are $\binom{n-k}{2}$ possible coalescences. If we ignore the times of the coalescent events then we calculate that there are $\prod_{i=0}^{n-2} \binom{n-i}{2}$ possible coalescent trees. If, for example, there are just six chromosomes then there are 2,700 possible topologies,

excluding times of coalescent events. In reality, however, the space of trees is infinite as the times in the tree impact the probability of mutation.

A number of analytical results exist for the original coalescent, some of which we shall see in the next section. These results become much more challenging to establish in alternative populations, e.g. recombination, variable population sizes and selection [Hartl and Clark, 2007]. In general, however, inference in the coalescent is intractable for most real datasets [Hein et al., 2005, Hartl and Clark, 2007].

A number of methods have been developed to make inference under the coalescent model. One of the first approaches was a ‘brute-force’ simulation method and resulted in a very successful application to mutation rate inference in *D. melanogaster* [Hudson et al., 1994]. Another approach developed by Tavaré et al. [1997], termed Approximate Bayesian Computation, resulted in another successful application of simulation based methods for inferring parameters in the coalescent.

More recently composite likelihood sequential Monte Carlo methods have been developed to infer specific aspects of the tree, e.g. local recombination rates [Hudson, 2001, McVean et al., 2004].

1.2 Natural Selection

Natural selection has played an important role in the development of all species. The term *natural selection* can be used to describe a number of different types of adaptive evolution, e.g. positive selection, negative selection, and balancing selection.

All forms of natural selection affect the allele frequency trajectory of the beneficial mutation [Hartl and Clark, 2007]. The model of random genetic drift proposed by

Wright [1931] describes how the frequency of a neutral allele varies over time in a Hardy-Weinberg population [Hartl and Clark, 2007]. The way in which selection affects the trajectory as described by the WF model [Wright, 1931] of genetic drift differs according to the type of selection. Furthermore, it has been observed that if the frequency of the allele was known in every generation then it would contain all the information about selection [Coop and Griffiths, 2004].

The difference between models of selection can be seen via a simple bi-allelic model. Suppose at a nucleotide there exists in a population a single allele, A . At some point a mutation occurs and introduces a new allele, a , into the population. In a diploid setting there are now three possible genotypes, homozygous AA or aa , or heterozygous Aa . The relative fitness of each genotype is parameterised by the vector $(1, 1 + s, 1 + hs)$, where s represents the strength of the selected allele, and h represents the heterozygous effect; references herein [Akey, 2009]. As with the mutation rate, s is often re-parameterised by the population scale selection coefficient $\sigma = 2N_e s$.

Positive selection, $(s > 0, h \in (0, 1))$ occurs when an allele rises in frequency faster than one might expect under genetic drift [Smith and Haigh, 1974]. The speed of the rise of this mutation depends on many factors, not least on how beneficial the phenotype encoded by this mutation is. The patterns of variation in the current day population can also be significantly altered depending on whether the mutation occurred before the presence of the selective pressure. Two models have been developed to describe these two scenarios: selection on *de novo* mutation and selection on standing variation, respectively [Barrett and Schluter, 2008].

The classical model of positive selection, i.e. selection on a new or *de novo* mutation describes the ancestry of a sample of chromosomes that have evolved with one mutation experiencing a constant selective pressure. This is often referred to as the hitchhiking model by Smith and Haigh [1974] and we refer to this model as a

‘classical’ sweep. Models which relax the assumption that the beneficial allele derives from a single ancestral background, e.g. selection on standing variation, have been developed to describe less stringent scenarios. In line with previous studies e.g. Messer and Petrov [2013], we refer to these models of selection collectively as ‘soft’ sweeps.

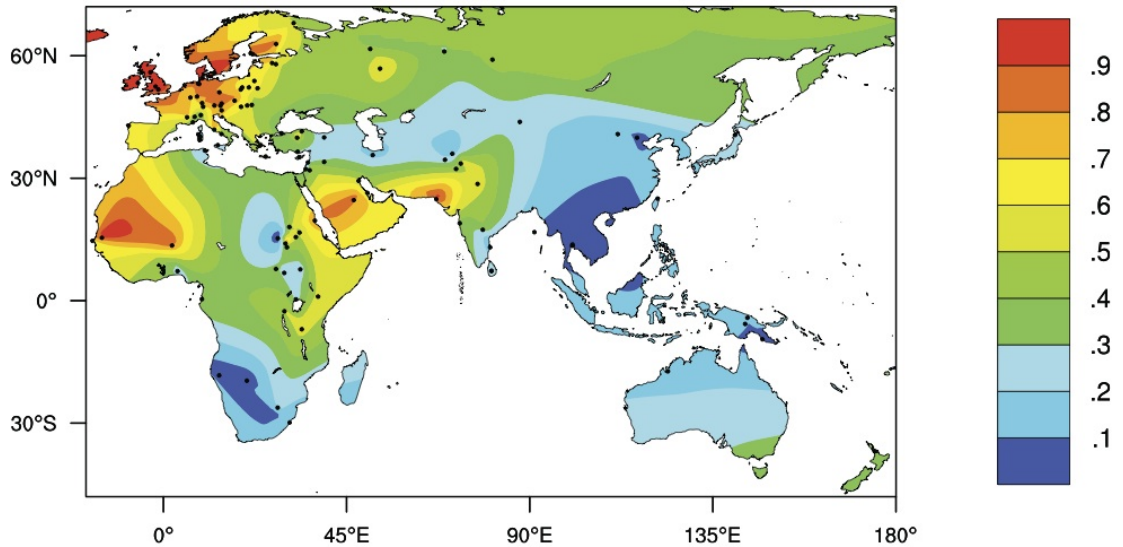


FIGURE 1.7: Interpolated map of Old World Lactose Persistence frequencies; Dots represent collection locations. Colours and colour key show the frequencies of the Lactose Persistence phenotype estimated by surface interpolation in Europe, Africa and Australasia. [Itan et al., 2010]

A well studied region in the human genome that is believed to be under strong positive selection is the Lactase (*LCT*) gene on chromosome 2 [Schlebusch et al., 2013]. Mutations at the *LCT* locus can result in persistence of the ability to digest lactose in to adulthood [Hollox et al., 2001, Enattah et al., 2002]. Figure 1.7 shows an interpolated map of lactose persistence across Asia, Africa, and Europe. Hollox et al. [2001] propose that the high frequency of lactose persistence in European populations was as a result of agricultural developments around the same time that humans were expanding into southern and eastern Europe.

Another form of selection, balancing selection ($s > 0$, $h > 1$), can occur when the heterozygous effect is over dominant, e.g. a beneficial allele is linked to another detrimental allele, i.e. homozygous and heterozygous alleles [Charlesworth, 2006]. In such a setting the beneficial allele is expected to remain polymorphic in the population. An example of a region in the human genome that is thought to be under balancing selection is the haemoglobin beta gene *HBB* on chromosome 11 [Leffler et al., 2013]. It is associated with the disease sickle cell anaemia (SCA) which distorts the structure of red blood cells [Ashley-Koch et al., 2000]. Homozygous carriers of the *HbS* variant of the *HBB* gene suffer SCA, but importantly heterozygous carriers of this variant have a 10-fold reduced risk of severe malaria [Hill et al., 1991, Ackerman et al., 2005]. Figure 1.8 shows the comparison between the frequency of mutation in the *HBB* gene and the extent of malaria in Africa.

In this thesis we focus on the challenges of detecting positive selection, specifically the impact of classical sweeps and soft sweeps that have yet to reach fixation within a population. Ongoing positive selection has a significant role to play in the emergence of drug resistant traits in various pathogens [Sabeti, 2008]. Rapid adaptive evolution to artificial environmental factors, such as public immunisation programs, presents a significant risk to patients where a pathogen is endemic, e.g. influenza in North America and western Europe [Bright et al., 2006], HIV in Africa [Clavel and Hance, 2004], and malaria Africa and south east Asia [Wellems and Plowe, 2001, Miotto et al., 2013]. Indeed there is increasing evidence of the role of positive selection in cancer [Goymer, 2008]. The role of selection in the interaction between pathogens and drugs is not just limited to positive selection; for example, in addition to *HbS* above, there is evidence for balancing selection in *G6PD* and *CD40L* [Sabeti et al., 2002a] which have been associated with malarial resistance. Nevertheless, because the literature on detecting selection is so widespread, we decided to limit the scope of this thesis to those concerned with detecting positive

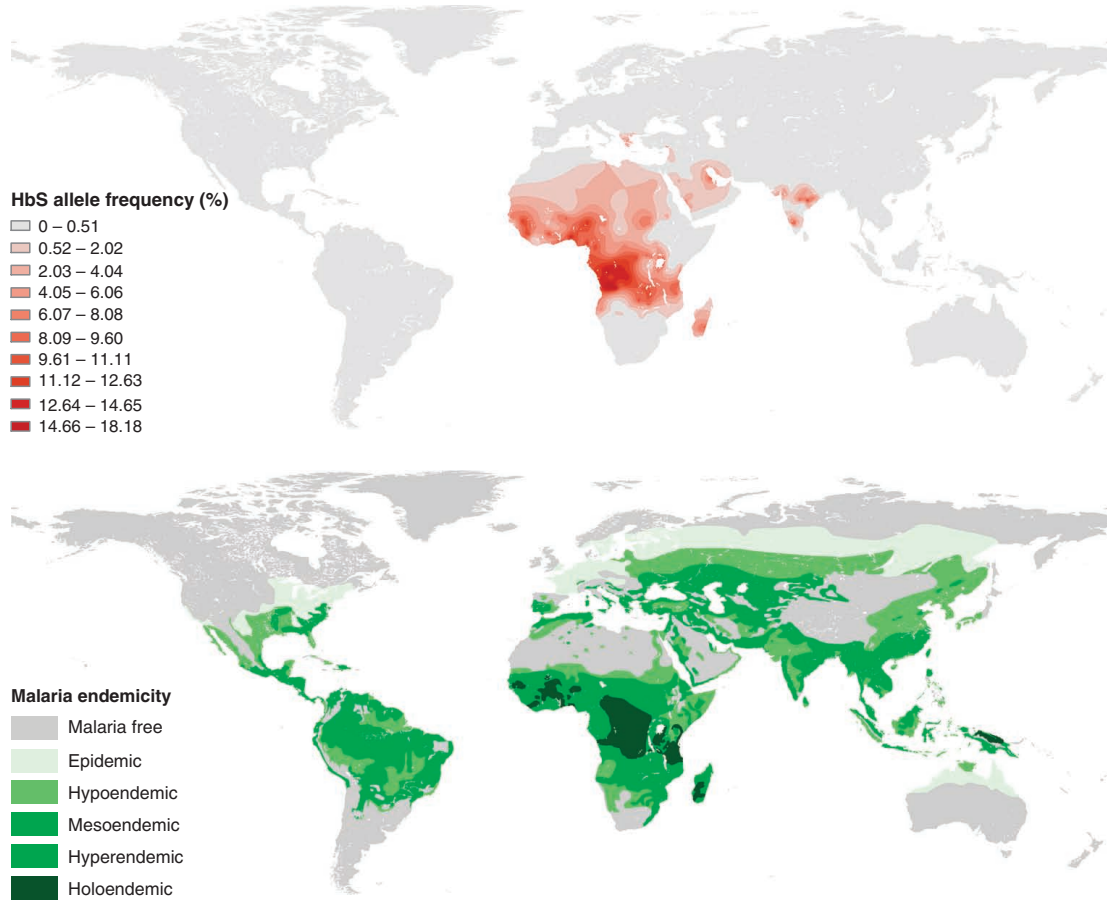


FIGURE 1.8: Correlation Between Malaria Endemics and SCA in Africa. The top panel shows the frequency of a *HbS* variant that confers increased protection against malaria. The bottom panel shows the observed extent of the Malaria endemic in Africa.[Piel et al., 2010].

selection, so that we might explore the area as extensively as possible.

1.2.1 A Model of Classical Selective Sweeps

Selection was one of the latter assumptions to be relaxed in Kingman’s coalescent model. As noted before, the original model, and indeed many of the later extensions, had the property that the distribution governing the mutations was independent of the distribution governing the coalescent times. This significantly reduced the complexity of the model and allowed computationally cheap simulation algorithms to be designed. The relaxation of homogeneous fitness by its

very nature alters the rate of coalescence, since it depends on the allelic states of the whole population with respect to the selected allele. Thus, the distributions governing mutations and coalescent times are no longer independent.

Krone and Neuhauser [1997] and Neuhauser [1999] made one of the early attempts to include selection in the coalescent, publishing a series of papers describing a model called the Ancestral Selection Graph (ASG). Borrowing ideas from the ARG, the ASG uses a branching and coalescent process back in time to produce hypothetical lineages. A mutation model is then applied and the coalescent is produced by following a series of decision rules making it dependent on the allelic states. Unfortunately, as noted by Coop and Griffiths [2004], the model proved to be very complex and lacked the simplicity that made Kingman's original model so popular.

Building on earlier work by Slatkin [2001], Coop and Griffiths [2004] proposed that to build a coalescent model that was dependent on the mutation process, one could construct a coalescent tree conditional on the frequency trajectory of the selected allele. This model is very similar to those used to incorporate variable population sizes, however the difference is that they act across all haplotypes, whereas this model acts only on those that carry the beneficial allele.

Using the WF binomial model [Wright, 1931] of allelic frequencies and its continuous-time approximation as a basis, Coop and Griffiths [2004] constructed a continuous time stochastic partial differential equation that conditions on the allele going to fixation within the population, i.e. a classical sweep. Building on the work of Coop and Griffiths [2004], Prezeworski et al. [2005] developed a similar model to describe the construction of a genealogical tree that experiences selection on standing variation. A schematic of the Coop and Griffiths [2004] model is shown in Figure 1.9.

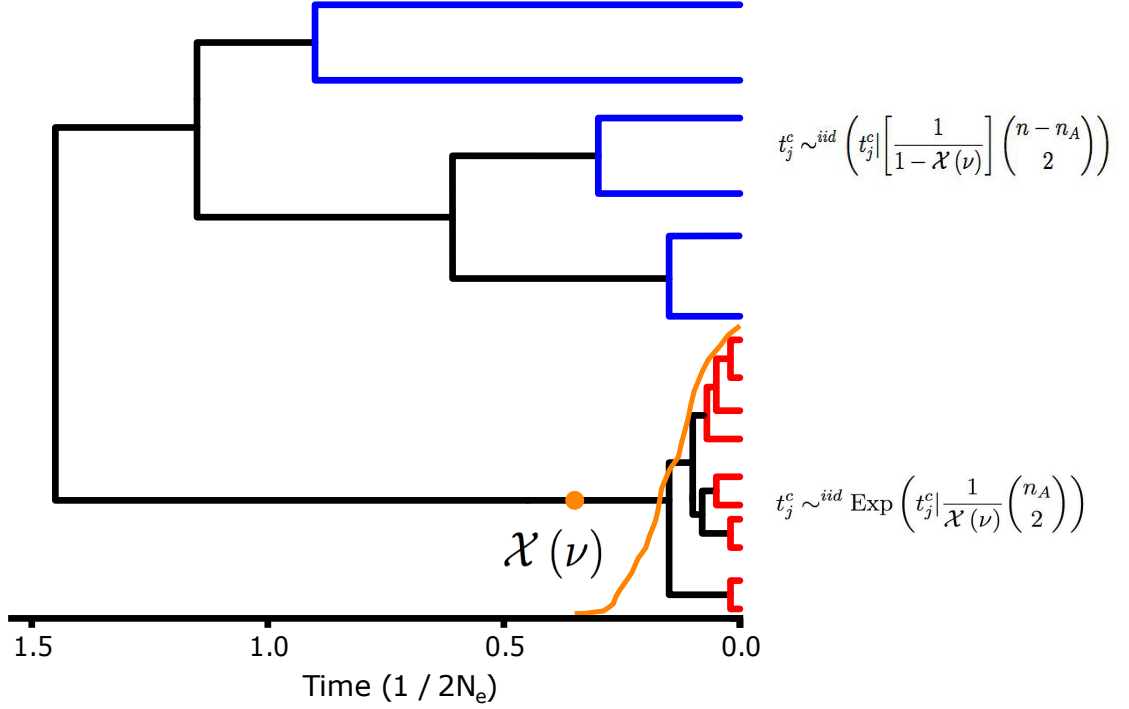


FIGURE 1.9: A schematic showing a retrospective genealogical tree under a classical sweep [Coop and Griffiths, 2004]. The rate of coalescence under the two alleles; ancestral - blue, derived - red, is governed by the distributions detailed on the right hand side. The allele frequency trajectory of the derived mutation is shown by the orange line. Time is measured in units of the effective population size and in reverse, i.e. current day has a value of 0.

Under the neutral coalescent the time to the next coalescent event with n chromosomes is exponentially distributed with mean $\frac{2}{n(n-1)}$ [Hein et al., 2005]. If we assume that at any given point ν in time, n_A individuals carry the selected mutation and the population frequency of the selected allele is $\mathcal{X}(\nu) > 0$ then the model of selection proposed by Coop and Griffiths [2004] describes the time to the next coalescent event for the selected individuals as exponentially distributed with infinitesimal mean $\frac{2\mathcal{X}(\nu)}{n_A(n_A-1)}$.

One effect of reducing the coalescent time for those individuals carrying the selected allele, assuming an infinite sites mutation model, is a reduction in the number of mutations that can accumulate between them, which will necessarily increase the similarities between them. McVean [2007] analytically details the effect of positive selection on patterns of LD around selected sites. He concludes that

while classical sweeps alter the frequency spectrum, the patterns of LD surrounding the beneficial allele look very different from those associated with recombination hotspots.

In their original model, Coop and Griffiths [2004] assume that the individuals are non-recombining, however a number of simulation models have been developed in the last ten years which allowed increasingly flexible descriptions of the evolution of a sample of chromosomes; *SelSim* [Spencer and Coop, 2004], *mb*s [Teshima and Innan, 2009], and *msms* [Ewing and Hermisson, 2010].

1.2.2 Effect on genetic variation

The effect of different modes of positive selection on genetic variation is helpfully described through the coalescent. Figure 1.10 shows schematics of three models of positive selection; classical sweep, soft sweep by recurrent mutation, and soft sweep by standing variation.

Under the classical sweep, the variation in the current day population comes from a single mutation event after the selective pressure. In this setting, as soon as a beneficial mutation arises, the hitchhiking effect [Smith and Haigh, 1974] and the selective pressures will mean that the current day population will likely have a single haplotype on the derived background. Contrast with the soft sweep scenario, where in general, the mutation event happens before the selective pressure. Therefore the fact that the beneficial allele is already segregating in the population will likely mean there will be multiple haplotypes on the derived background.

To see this in practice, we used the coalescent simulator *msms* [Ewing and Hermisson, 2010], to investigate the patterns of genetic variation under a classical sweep, and soft sweep by recurrent mutation.

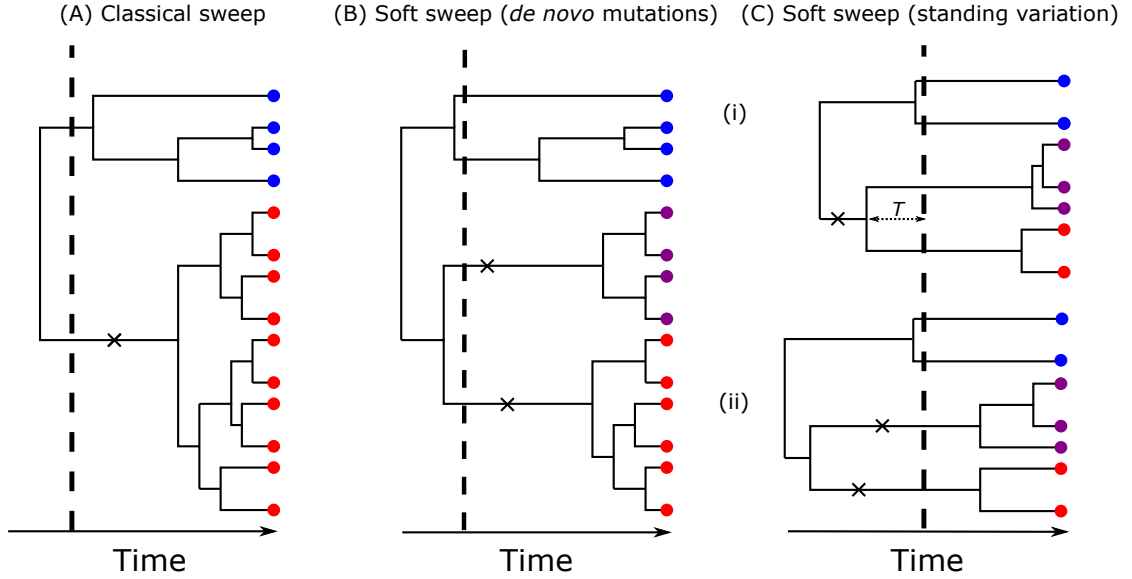


FIGURE 1.10: A schematic modified from Messer and Petrov [2013] showing the difference between models of classical selective sweeps and soft sweeps. Each detail the lineages of those evolving neutrally (coloured blue) and those evolving under a selective pressure. **(A)** As in Coop and Griffiths [2004], the model of a classical sweep describes a beneficial allele that arises from a single mutation (marked by X), after the selective pressure (the dashed line) affects the population. **(B)** Like the classical sweep, a model of a soft sweep by *de novo* mutation describes how the beneficial allele arose from multiple mutation events, after the presence of the selective pressure. **(C)** The model of a soft sweep on standing variation describes how the beneficial allele is already segregating in the population before the onset of the selective pressure. This can occur either from **(i)** a single mutation, similar to **(A)** or **(ii)** multiple mutation events, similar to **(B)**.

We simulated two recombining 2Mb regions from a population of constant size that experience a selective sweep with current day frequency of 50%, and population scaled selection coefficient $\sigma = 2N_e s = 500$. The other parameters were chosen so that they reflected observed characteristics of human genetic variation; mutation rate per generation per base pair, $\mu = 1.2 \times 10^{-8}$ [Sally and Durbin, 2012], recombination rate per generation per base pair $r = 1 \times 10^{-8}$ [Kong et al., 2002], and $N_e = 10,000$ [Kim et al., 2010]. For a discussion about the values of these parameters please refer to the Methods in Chapter 2.

The plots for the haplotype, ordered depending on the allele they carry at the selected site, is shown in Figure 1.11. As expected in the top left panel we see

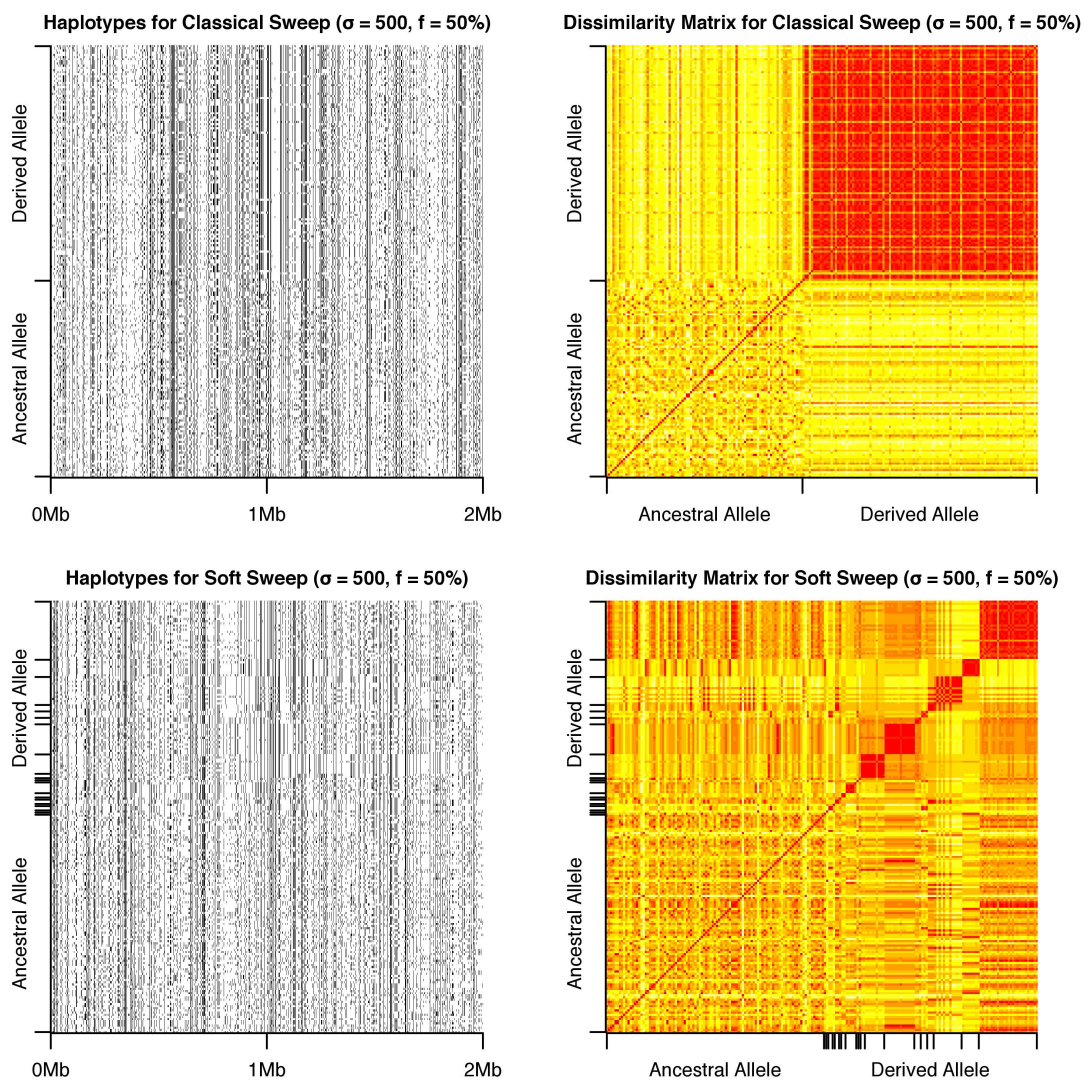


FIGURE 1.11: Example of impact of varying strengths of positive selection in simulated genetic data using *msms* Ewing and Hermisson [2010]. The left panels show the haplotype variation in the region surrounding the selected loci - haplotypes are coded black or white depending on the allele that they carry. The right panels show the heat map representations of the pairwise difference between all chromosomes - the black lines in the axes represent the boundaries of the ancestral backgrounds of the chromosomes.

that under a classical sweep there is a reduction in heterozygosity in a region surrounding the selected locus for those carrying the beneficial allele. This can also be observed by the dissimilarity matrix in the top right panel of Figure 1.11, which shows that those haplotypes carrying the selected allele (i.e. derived allele) have fewer positions that differ between them. Contrast this with a soft sweep in bottom panels where we observe multiple haplotypes on the derived allele background.

Observing patterns of variation in simulated data when the beneficial allele is known, is helpful, however making inference on the strength of the selected coefficient, or the location of the selected allele in real data is not possible. To explore, in detail, some of the important selection detection and inference tools we focus, in the next section, on those that have been shown in the past to be sensitive to classical sweeps. In later chapters we look at different methods that have been developed to detect patterns of variation associated with soft sweeps.

1.3 Tools for Detecting Classical Sweeps

Tools for detecting regions of a genome under positive selection have until recently been limited to candidate gene studies, e.g. *LCT* [Bersaglieri et al., 2004]. In general, however, these types of studies can suffer from two key limitations [Akey, 2009].

First they require a priori knowledge about which genes are likely to be under selection, which requires knowledge of genotype-phenotype relationships. Akey [2009] argue that this is possible in some circumstances, e.g. Mendelian inheritance, when a single base change changes the phenotype; e.g. *LCT* or *HBB*. An alternative approach might be to focus on loci belonging to gene classes that appear to be targets of selection, e.g. immune response [Bamshad and Wooding,

2003], however that can lead to biased sets of loci [Akey, 2009]. Akey [2009] argues that candidate gene studies can be inefficient study designs for identifying selective sweeps in regulatory regions at large distances from the target loci.

Akey [2009] also identifies another challenge with gene candidate studies; as we discussed before there are potential confounding effects of genetic drift that may occur as a result of population demography. Therefore interpreting patterns of variation at loci adds additional complexity to the challenge of detecting selective sweeps.

1.3.1 Departures from Neutrality

Prior to genome wide tests for selection a common approach involved using the distribution of summaries of genetic variation to test for non-neutral evolution. These include, θ_W [Watterson, 1975], D [Tajima, 1989], D^* [Fu and Li, 1993] and H [Fay and Wu, 2000]. Many of these statistics are part of a family of metrics based on the frequency spectrum. For a sample, $\left\{H_j^{(i)}\right\}_{i,j=1}^{n,L} \in \{0,1\}^{n \times L}$, of n chromosomes at L segregating sites the frequency distribution is defined as $\{X_l\}_{l=1}^{L-1}$, where

$$X_l = \left| \left\{ j : \sum_i H_j^{(i)} = l \right\} \right|.$$

Watterson [1975] showed, in an infinite sites model, that the expected number of segregating sites, K , was $a_1(n)\theta$, where θ is the mutation rate, and $a_1(n) = \sum_{i=1}^{n-1} \frac{1}{i}$. Hence, he produced an unbiased estimator $\hat{\theta}_W = \frac{K}{a_1(n)}$, now termed Watterson's estimate. Under a neutrally evolving population, this is an unbiased estimate of the mutation rate, however in the presence of a selective sweep, the

number of segregating sites will tend to be reduced due to the shorter than expected time to most recent common ancestor amongst the chromosomes carrying the selected haplotype.

Tajima [1989] argued that without knowing the value of θ_W in the neutral case, it is difficult to establish whether a region is under selection. Using properties of the distribution of pairwise differences under the neutrally evolving coalescent, Tajima [1989] showed that his statistic (Equation 1.3) is centred around 0 and is sensitive to changes in variation. Define $\hat{k}$ to be the average of the pairwise difference distribution, and K to be the number of segregating sites in the sample. Then D can be expressed as

$$D = \frac{\hat{k} - \frac{K}{a_1(n)}}{\sqrt{e_1(n)K + e_2(n)K(K-1)}}. \quad (1.3)$$

$$a_1(n) = \sum_{i=1}^{n-1} \frac{1}{i}, \quad e_1(n) = \frac{n+1}{3a_1(n)(n-1)} - \frac{1}{a_1^2(n)}$$

$$a_2(n) = \sum_{i=1}^{n-1} \frac{1}{i^2}, \quad e_2(n) = \frac{1}{a_1^2(n) + a_2(n)} \left(\frac{2(n^2 + n + 3)}{9n(n-1)} - \frac{n+2}{na_1(n)} + \frac{a_2(n)}{a_1^2(n)} \right).$$

Using a combination of simulation studies and numerical examples from *D. melanogaster*, Tajima [1989] showed that the D statistic was sensitive to conserved regions of DNA which can be associated with positive selection. Since D has an intractable distribution the critical values of D are computed by assuming it is distributed as a scaled beta distribution [Tajima, 1989]. As noted by Simonsen et al. [1995] these values tend to be somewhat conservative. Therefore it may be necessary to determine the critical values of D [Tajima, 1989] using simulations. This becomes even more important when the organism undergoes recombination [Nielsen, 2001].

Fu and Li [1993] published two alternative statistics, D^* , and F^* , to test for non-neutral evolution. Following convention we use the $*$ notation to distinguish the methods from previous statistics. Define η to be the number of mutations on the

tree then, in a neutrally evolving population, Fu and Li [1993] showed the expected number of singletons, η_s , is

$$\mathbb{E}(\eta_s) = \frac{n\theta}{n-1}.$$

Under the infinite sites mutation model η is exactly the number of segregating sites. D^* , and F^* were constructed to assess the changes in η_s by assessing the number of mutations in the internal (older), and external (younger) branches. Fu and Li [1993] propose that selection will effect these branches in different ways; ‘in purifying or negative selection there will tend to be increase mutations in the external branches because deleterious alleles are present at low frequencies’. F^* , unlike D^* , compares the distribution of mutations between these branches by looking across related species, e.g. out-groups. When it is not possible to establish out-groups Fu and Li [1993] propose using the D^* statistic

$$D^* = \frac{\left(\frac{n}{n-1}\right) \eta - a_1(n) \eta_s}{u(n) \eta + v(n) \eta^2}. \quad (1.4)$$

Due to the complexity and number of dependencies that are involved in defining $u(n)$, and $v(n)$ we do not detail them here but refer to the appendices of Fu and Li [1993].

Later Fay and Wu [2000] proposed the H statistic which is designed to adjust for the demographic effects by combining multiple pieces of information from different loci. The H statistic gives weight to high frequency derived alleles that can result from a selective sweep.

$$H = \sum_{i=1}^{n-1} \frac{2X_i i (n-i)}{n(n-1)} - \sum_{i=1}^{n-1} \frac{2X_i i^2}{n(n-1)}. \quad (1.5)$$

Simonsen et al. [1995] performed a series of power simulations and concluded that D [Tajima, 1989] is, among the tests in this class, most powerful against the

alternative hypothesis of selective sweeps, and demography. More recently Achaz [2009] showed that these statistics were in fact specific instances of a general model, and that this framework can be used to devise new more powerful tests that better detect deviations from the neutral hypothesis. He illustrates the utility of the model by showing how it supports a selection hypothesis in the human *LCT* gene.

As noted by Nielsen [2001], whilst these tests have been successful in many applications, the interpretation of when results are significant isn't always clear. As we mentioned before, it is widely acknowledged that the neutral hypothesis is a composite hypothesis, i.e. significant deviations from the null does not provide evidence against selective neutrality [Nielsen, 2001]. Therefore great care must be taken when using these neutrality departure test for the presence of a mutation undergoing positive selection.

1.3.2 Functional Tests

An alternative approach, that is more robust to alternative demographic models, uses the difference in functional substitutions in protein coding genes to detect evidence of positive selection. In particular the McDonald and Kreitman [1991] test counts the number of non-synonymous and synonymous mutations within and between species. McDonald and Kreitman [1991] argue that if the observed substitutions are neutral, the ratio of non synonymous to synonymous polymorphism between species should be the same within and between species. McDonald and Kreitman [1991] show evidence of more non-synonymous mutations between species at the *Adh* gene in samples from three species of *Drosophila*, indicating evidence to support the hypothesis of positive selection.

Nielsen [2001] notes that despite observing that the McDonald and Kreitman [1991] test can provide unambiguous evidence for selection, it is not always clear

which type of selection is acting on the gene. Nevertheless, he argues that these types of tests are robust to changes in population sizes, because they are constructed by comparing different types of variability within the same locus, or genomic region.

1.3.3 Likelihood-based Approaches

The model developed by Coop and Griffiths [2004] of selective sweeps was shown by the authors to be a full likelihood solution to the challenge of estimating the selection coefficient, σ , in a specific scenario. Due to the complexity of the likelihood, the authors tackle the problem using an importance sampling technique [Gelman and Meng, 1998]. Coop and Griffiths [2004] demonstrate the utility of their importance sampling method on small simulated and real datasets.

The real dataset in question was from Verrelli et al. [2002] and focused a region surrounding the *G6PD* gene. Among others, Sabeti et al. [2002a] observed that this gene was associated with malarial resistance and could be under balancing selection. The region spanned 5.2Kb and involved 216 subjects. Coop and Griffiths [2004] show that their method works well on simulated data sets but observed that the maximum likelihood estimate of σ in the *G6PD* study was effectively infinite. Coop and Griffiths [2004] argue that the merits of their method lie primarily in the development of a full likelihood model of selection, not in the importance sampling inference model.

Zhu and Bustamante [2005] proposed an alternative model based on a composite likelihood ratio test that attempts to establish whether a gene is under directional selection. Zhu and Bustamante [2005] conclude, however, that their method has good power when the selection is weakly negative, but has only moderate power to detect positive selection. They argue that their method is robust to biased

recombination rate estimates, but not to population growth or bottlenecks, and they conclude that their estimates of the selection coefficient can be conservative.

1.3.4 Genome Wide Selection Scans

After the publication of the human genome there was paradigm shift towards using large samples of genomes to make inference in population genetics. The phrase, ‘population genomics’, was coined by Black et al. [2001] to describe the approach of using genome-wide samples to infer loci-specific population genetic distributions.

Akey [2009] argues that this approach offers solutions to the two challenges associated with gene candidate studies. First, the genome can be surveyed without prior knowledge of which genes are under selection. Second, population genomics offers, at least in principle, the ability to distinguish between the effect demography and positive selection. Demographic pressures are equally likely to affect the whole genome whereas selection acts locally and only on those haplotypes carrying a beneficial allele.

The shift towards population genomics was led in part by the huge gains in technology allowing geneticists to sequence organisms more accurately, and at a significantly reduced price [Mardis, 2011]. Akey [2009] describes the development of genome-wide scans of selection and constructs a meta-map of recent signals of positive selection in the human genome. He argues, in general, that population genomics models are based on identifying outlier loci from empirical distributions of metrics that are sensitive to patterns of selection. Highlighting the similarities of hypothesis testing in population genomics with that of rejecting the neutral hypothesis, Akey [2009] argues that an observation in the tails of the empirical distribution doesn’t prove that that loci has been influenced by positive selection. Instead, he proposes that all one can conclude is that this loci exhibits patterns

of genetic variation that are unusual in some respect relative to the rest of the genome.

One of the first metrics developed to leverage a signal left by positive selection in genomic data is Extended Haplotype Homozygosity (EHH) [Sabeti et al., 2002b]. EHH is defined as the probability that two randomly chosen haplotypes carrying the target allele are identical by descent (IBD) for the entire interval from that locus to the point x . Two haplotypes are IBD if they are replicas of a haplotype that is present in a previous generation [Hartl and Clark, 2007]. Using this definition Sabeti et al. [2002b] argue that EHH is designed to leverage the observation of elevated regions of reduced LD for haplotypes carrying the selected allele, i.e. the ‘hitchhiking’ effect [Smith and Haigh, 1974].

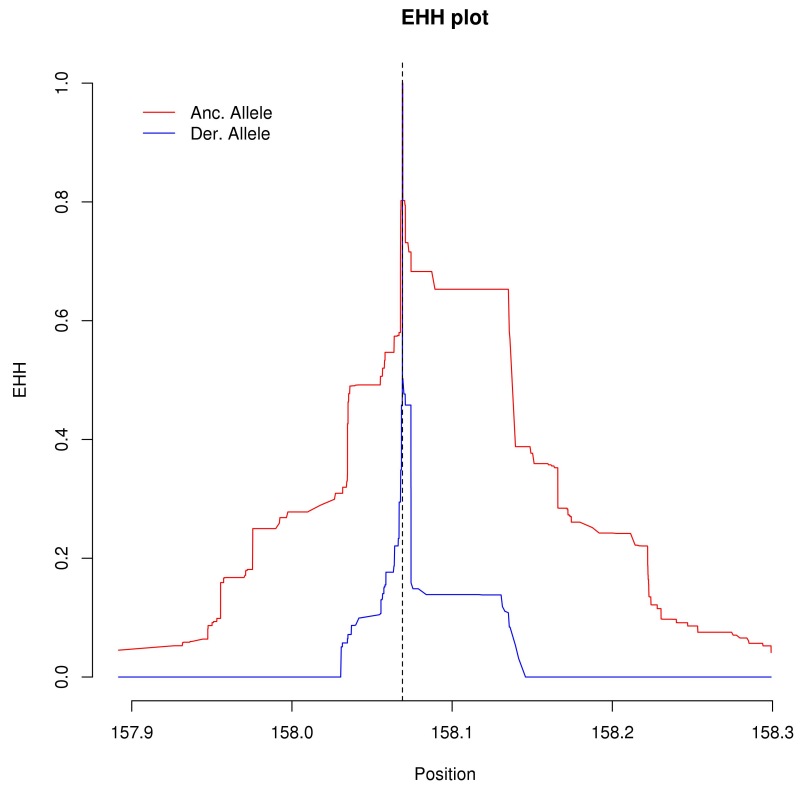


FIGURE 1.12: Example of Extended Haplotype Homozygosity decay at the *LCT* gene in a sample from the CEU population using data from 1000 Genomes Project Consortium et al. [2012] and *R* package *rehh*. Position measures in units of genetic distance (cM).

Sabeti et al. [2002b] note that the signal of haplotype homozygosity may also be observed in regions of high LD as a result of low recombination rates. Therefore to construct their test for positive selection, referred to as the Long Range Haplotype (LRH) test, Sabeti et al. [2002b] calculate EHH on the ancestral allele background and on the derived, and establish the relative difference. Sabeti et al. [2002b] argue that LRH will be biased towards the derived alleles as they are younger therefore more similar and more likely to have reduced LD. Using empirical distributions, as noted in Akey [2009], Sabeti et al. [2002b] establish the cut off point for significant scores.

The authors demonstrate the utility of this method by looking at genes that have been implicated in resistance to severe malaria (*Plasmodium falciparum*) in humans: *G6PD* [Ruwende and Hill, 1998] and *CD40* [Sabeti et al., 2002a]. In their investigation of alternative methods, Sabeti et al. [2002b] looked at many of the metrics that we discussed above on their sample of *G6PD* material. They show that none of the statistics show evidence of a non-neutral evolution, suggesting that they are not sensitive to balancing selection, the majority view on the evolution of *G6PD* [Verrelli et al., 2002].

Rather than looking at the relative value of EHH, Voight et al. [2006] propose looking at its decay to improve the resolution of the genome wide maps of positive selection in humans. Voight et al. [2006] propose an approach that numerically integrates under the EHH curves for both alleles and determines whether there is a significant difference. Define uIHS as the unstandardised iHS score and iHH_A as the integrated EHH metric under the ancestral allele A , equivalently iHH_D for the derived allele. Then

$$\text{uIHS} = \log_e \left(\frac{\text{iHH}_A}{\text{iHH}_D} \right). \quad (1.6)$$

Like Sabeti et al. [2002b], Voight et al. [2006] observed that the unstandardised

iHS score is correlated with derived allele frequency therefore they propose standardising the metric within frequency bins to produce the final iHS metric. In noting that the standardised iHS score appears to have Gaussian qualities, Voight et al. [2006] propose using the ubiquitous threshold of two standard deviations away from the mean to establish how unusual the haplotypes around a given SNP are. The quality of this approximation depends on a number of factors, e.g. number of observations within frequencies, and whilst it may, in practice, produce results, it is not the most powerful approach as there is no parametric form for the likelihood. Akey [2009] argues that not placing stringent assumptions on the distribution of metrics calculated using variation data can be an advantage, however not having a parametric form for the likelihood means that it is difficult to assess the false positive rate as it necessarily depends on the fit of the observations to the distribution.

Methods based on EHH tend to only target one specific aspect of the signal left by adaptive evolution: extended haplotype homozygosity. There are other non-parametric methods that attempt to exploit other observations, e.g. XP-EHH Sabeti et al. [2007] which leverages the extended haplotype signal across populations to improve the resolution of the selection maps and estimate of the nature and age of the selection.

A significant impact of the approach is that EHH will tend to produce false negatives in the presence of soft sweeps. One reason for this is that soft sweeps tend to produce multiple haplotypes on the derived background, as the beneficial allele is segregating in the population before it became advantageous. As a result EHH, will decay much quicker because of the increased variation present in the derived background due to the multiple haplotypes.

Alternative methods that more directly estimate the level of IBD sharing across the genome have recently been developed. In Albrechtsen et al. [2010] and Han

and Abney [2013] the authors use a combination of theory and simulations to show how IBD changes in the presence of ongoing positive selection. The first paper, Albrechtsen et al. [2010], develops the framework removing loci in low LD, and shows the utility of the method on the *MHC* region of the human genome. The followup paper Han and Abney [2013] incorporates the LD information and builds on the model by Albrechtsen et al. [2010] to scan samples of Kenyan (MKK) chromosomes from the HapMap Project [International HapMap Consortium, 2005]. The results from the MKK study show patterns that agree with previous studies, however there are significantly fewer signals of recent positive selection using this method when compared to alternative methods [Voight et al., 2006] on similar populations.

We will be returning to many of the techniques that we have mentioned in this section throughout this thesis.

1.4 Motivation

Turning our attention away from a general discussion, we note that the literature on detecting signals of adaptive evolution appears to have a divide. That divide appears to be between models that scale but are non-parametric in that their results don't have a direct evolutionary interpretation e.g. iHS [Voight et al., 2006] and models that are very parametric but do not scale e.g. Coop and Griffiths [2004].

These population genomic methods, e.g. EHH [Sabeti et al., 2002b], and iHS [Voight et al., 2006], appear to be able to detect signals of classical positive selection, however it is not clear whether they would be robust to other forms of positive selection. Moreover, the non-parametric nature of their inference means

that comparisons are difficult due to the fact that the false positive rates are difficult to control. Indeed from a Bayesian point of view they are limited in the sense that they are not easily able to combine information from different sources e.g. previous studies.

Whilst this provides us with a strong statistical motivation, we are also motivated by the biological consequences of positive selection. Darwinian evolution tells us all organisms are a product of adaptive evolution. The ability of an organism to adapt to its environment and pass those traits on to its progeny with high fidelity, is a well understood process, however as we have seen detecting these traits using genetic variation has its challenges. As we alluded to earlier in the chapter, and as noted in more general terms by Darwin, a organism's ability to adapt to changing environments can give it a distinct advantage. This is perhaps especially true when we consider the problem of protecting humans against pathogens.

Take Malaria as an example. The World Health Organisation (WHO) estimated that, in 2010, malaria was responsible for the deaths of 660,000 people, the majority of which were young children in Africa. Over the last five decades, a number of drugs have been developed to help tackle the spread of the disease in humans. Chloroquine was used extensively in the global effort to eradicate malaria in 1950s and 1960s, however in the late 1950s chloroquine-resistant strains of *P. falciparum* were discovered on the Thailand-Cambodian border, and later around the world [Payne, 1987]. *P. falciparum*'s resistance to pyrimethamine [Roper et al., 2004] and sulfadoxine [Mita et al., 2011] is also now well established in global populations. Recently strains of the parasite, resistant to the new anti-malarial drug artemisinin, were discovered in western Cambodia [Noedl et al., 2008]. As a result, there is now an urgent need to discover genetic markers associated with artemisinin resistance in *P. falciparum* [Miotto et al., 2013].

It is for these reasons that we are motivated to construct a fully Bayesian model

that use well established population genomic models of evolution and ancestry to detect signals of positive selection in genomic data. Moreover, we want to construct the model in such a way that it is robust to demographic history and population structure, but at the same time is capable of detecting both classical and soft selective sweeps. Finally, we require that the method scales to contemporary genomic datasets, and is capable of running on modern computing infrastructure, i.e. multicore compute nodes and distributed systems.

1.5 Outline

Since direct inference using the coalescent likelihood is challenging, we explore, in Chapter 2, methods that have been developed to make inference in intractable likelihood settings. We begin by looking at the four frequency spectrum statistics from this chapter and assess the degree to which each statistic is affected by strength of positive selection using simulated data. We also illustrate the confounding effect of variable population sizes on these statistics. Then using a variety of vanilla machine learning and simulation techniques, e.g. Approximate Bayesian Computation, we establish the feasibility of using these methods to make inference on positive selection. We show this type of approximate inference to detect patterns of variation associated with positive selection is very challenging and will not scale well to genome wide variation data.

In Chapter 3 we investigate more approximate models that are often used genomics to perform inference; specifically Hidden Markov Models. We show how one such method, chromosome painting [Lawson et al., 2012] can be used to infer changes in the ancestry of a sample of chromosomes. We show through a series of simulations and theoretical arguments that it is possible to identify patterns of variation consistent with positive selection by detecting regions of the genome where a subset of the individuals have reduced time to common ancestor. Using

parallels with iHS [Voight et al., 2006], we construct a simple Bayesian model that can probabilistically identify regions of a genome that exhibit patterns of variation associated with positive selection.

In Chapter 4 we perform a series of power simulations and compare the results to existing models that have been used to detect positive selection, e.g. iHS and H_{12} [Garud et al., 2013]. We show that our method has comparable power to iHS and H_{12} to detect strong classical sweeps, and increased power to detect soft sweeps. We also discuss the effect of demography and population size of the distribution of the results from our method, and we explore the effect of changing the sample size.

In Chapter 5 we demonstrate the utility of our model by applying it to genome wide sequence variation data from 1000 Genomes Project Consortium et al. [2012]. We show that our method and iHS identify many of the previously identified signals of positive selection, e.g. *LCT* [Hollox et al., 2001] and we show that there is a high degree of overlap between these results and the integrated map of positive selection constructed by Akey [2009]. In addition we highlight three novel candidates for positive selection in the human genome that have some of the largest values under our method and are considerably lower down the ranks under iHS. We show that two of these candidates show patterns of variation consistent with a soft selective sweep.

Finally, in Chapter 6 we draw conclusions based on the results presented in this thesis and discuss potential extensions of the model that we have looked at during the development of the method. Finally we provide a short discussion about future directions and application of the method to other species.

Chapter 2

Classical Approaches to Detect Selective Sweeps

“If we knew what it was we were doing, it would not be called research, would it?”

Albert Einstein

The goal for this thesis is to design and apply new techniques to detect regions in genome wide datasets that appear to have the characteristics of an ongoing selective sweep. We highlighted in the previous chapter, that one of the major challenges in detecting selection in real data is that full likelihood inference under the coalescent is intractable for many applications. A number of simple summary statistics have been constructed to detect departures from neutral evolution, e.g. D [Tajima, 1989]. Whilst these have been shown to be sensitive to patterns of variation associated with selection they are also affected by demographic history. For example population expansion or bottleneck.

Recently there has been major advances in both the development of new simulation based inference techniques and computational hardware. A particular method, known as Approximate Bayesian Computation (ABC) [Tavaré et al., 1997], is

designed specifically to make inference in settings where full likelihood inference is intractable. ABC relies on simulating large amounts of data from a model, and making inference on a set of parameters based on proximity measures between the observed and the simulated data. Since ABC has been shown to be an effective inference tool in different population genetic settings, e.g. Tavaré et al. [1997] we decided that it was important to determine the feasibility of using these approaches to make inference on the strength of ongoing positive selection.

We begin by looking at the effects of selection and alternative models of demography on the frequency spectrum statistics from the previous chapter, e.g. $\hat{\theta}_W$ [Watterson, 1975], D [Tajima, 1989], D^* [Fu and Li, 1993] and H [Fay and Wu, 2000]. In line with previous studies [Simonsen et al., 1995, Fay and Wu, 2000], we show that these methods are sensitive to selection but also to changes in demography. We then look at using these metrics in a machine learning (ML) regression framework to predict the true value of the selection coefficient σ , however we conclude that estimates tend to be somewhat unreliable. Then using an ABC approach we establish whether we can improve the quality of the inference. We conclude by showing that whilst the inference quality does improve, the estimates are still highly variable, even in the simplest problems. More generally we show that, even with the hardware available to us, the computational burden associated with these methods means that applying them to real data, at scale, would be very difficult.

2.1 Methods

In this section we detail the methods that we use in this chapter. We begin by detailing the coalescent simulator and the parameters used to generate the data. We then describe a number of machine learning tools that we use later in the

chapter to make inference. Finally, we introduce the Approximate Bayesian Computation paradigm and a number of challenges associated with making inference in this framework.

2.1.1 Data Generation

To generate the genetic variation data in this thesis we had a choice of a number of coalescent simulators. For example *mb*s [Teshima and Innan, 2009] and *msms* [Ewing and Hermisson, 2010]. To maintain continuity throughout this thesis we decided to use the forward in time selection simulator *msms* as it is a fast and efficient coalescent simulator with inbuilt parallelisation, allowing us to take advantage of modern computing infrastructure; multiple compute core, multi-node supercomputers [Ewing and Hermisson, 2010]. *msms* uses a combination of deterministic and stochastic method to create the selected allele frequency trajectory and has a standard input/output (IO) format, [Hudson, 2002], that is backward compatible and allows construction of simple simulation pipelines to perform inference.

Each dataset is simulated for one hundred recombining chromosomes on a 200Kb region. The effective population size N_e for each dataset was fixed at 10,000, which is often cited as an estimate of the effective population size of modern humans [Kim et al., 2010]. We chose $\mu = 1.2 \times 10^{-8}$ per generation per base pair to be the value of the mutation rate ($\theta = 96$) [Sally and Durbin, 2012], and the probability of recombination was set at, $r = 1 \times 10^{-8}$ per generation per base pair [Kong et al., 2002] ($\rho = 80$). These parameters we chosen to reflect the average variation observed in humans, not any specific subpopulation. Note that we do not model recombination heterogeneity due to its increased computational complexity. This will likely reduce the applicability of this method to real data.

We explore two simple demographic scenarios; a population experiencing a bottleneck lasting 800 generations, with instantaneous recovery, and a population experiencing exponential expansion beginning 1000 generations ago. For the bottleneck we looked at the effect of reducing the population size during the bottleneck to 50% of the current day effective population size. For the exponential expansion we chose a growth rate $\alpha = 92.1$, which means that the size of the population at t_0 in Figure 1.6(A) is 10% of the current day population. As before, these parameters were not chosen to match inferred demographic histories of human populations, as undertaking an exhaustive exploration of all potential scenarios is a major undertaking and beyond the scope of this thesis. Note, however, that these parameters are similar to those described in previous studies, e.g. [Gutenkunst et al., 2009].

2.1.2 Machine Learning Tools

Machine learning (ML) is the body of research related to automated large-scale data analysis, e.g. Barber [2012]. To relate a set of observations, to a set of predictors there exist two general model classes; regression and classification. Regression models are primarily used in conjunction with continuous observations for the purposes of prediction, whereas classification models tend to be used in conjunction with discrete observations for the purposes of classification [Barber, 2012].

In this chapter we work with three vanilla ML regression models; generalised linear model, general additive model, and neural network. All of these models are constructed to relate a continuous observation to a set of predictors, however they differ in the way that they perform the inference.

Generalised linear models (GLM) describe the relationship between a response variable, Y , and a vector of explanatory variables, X , through a link function, g :

$g(Y) = \eta = X\beta$. For example in logistic regression $y = f(x) = \text{logit}^{-1}(x)$ [Casella and Berger, 1990].

Like GLMs, general additive models (GAM) describe the relationship between Y and X through a link function, f , however rather than using the raw values X , different parametric, or non-parametric functions can be applied to each variable in X to improve the fit [Venables et al., 1994].

Finally, neural networks (NN) are part of a much larger class of latent variable models, i.e. Y is related to X through a series of latent variables Z [Hastie et al., 2009]. NN have proved to be extremely powerful tools in ML and have been used in many different fields, e.g. [Vanhoucke et al., 2011].

2.1.3 Approximate Bayesian Computation

A goal of Bayesian inference is to describe the posterior probability distribution, $\mathbb{P}(\gamma|D)$, of a vector of parameters, γ , after observing the data, D [Barber, 2012]. In the purest sense, this requires knowing the likelihood, $\mathcal{L}$ and a prior probability distribution, p on γ .

$$\mathbb{P}(\gamma|D) \propto \mathcal{L}(\gamma|D)p(\gamma) \quad (2.1)$$

If $\mathcal{L}$ and p are conjugate, i.e. the prior and posterior have the same algebraic form, then this inference can often be done analytically; e.g. exponential family of random variables [Diaconis and Ylvisaker, 1979]. If not, then a Monte Carlo (MC) approach can often be used to approximate the inference, however this requires that the likelihood, $\mathcal{L}$, is known up to proportional constants and can be point wise evaluated [Chib and Greenberg, 1995]. If, however, this is not the case, i.e. an intractable likelihood, then other inference methods must be considered.

Approximate Bayesian Computation (ABC) [Tavaré et al., 1997], is a MC approach to make inference in intractable likelihood settings by simulating data from a model using parameters drawn from a prior distribution, and discarding parameters that result in observations that are not equal to the observed.

Tavaré et al. [1997] was the first application of ABC to population genetics. In that paper the author developed a method to infer coalescent times from DNA sequence data by taking advantage of the ease with which samples could be generated in the coalescent. The success of this paper lead in part to the development and formalisation of the ABC inference framework [Lopes and Beaumont, 2010].

In practice, however, the probability that two datasets are exactly equal is close to zero, therefore we would have to wait a very long time before one sample was accepted. In order to soften this assumption the ABC methodology allows for some tolerance level and in the case of high dimensional data a distance metric, $\mathcal{D}$, that accepts parameters that are sufficiently close, i.e. $< \epsilon$. The original ABC algorithm is detailed in Algorithm 1. [Lopes and Beaumont, 2010]

Require: N the number of samples required; γ , the parameter vector with prior $\pi(\cdot)$; $f_{\mathcal{M}}(\cdot|\gamma)$ the model; observe Y in data space $\mathcal{X}$; $S : \mathcal{X} \mapsto \mathbb{R}^d$ summary statistic and $\mathcal{D} : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ a distance metric. Finally $\epsilon \in \mathbb{R}$ chosen tolerance level

```

1:  $k = 0$ 
2: while  $k \leq N$  do
3:    $\gamma^* \sim \pi(\cdot)$ 
4:    $X^* \sim f_{\mathcal{M}}(\cdot|\gamma^*)$ 
5:   if  $\tilde{\mathcal{D}}(S(X^*), S(Y)) \leq \epsilon$  then
6:     store  $\gamma^*$ ; set  $k = k + 1$ 
7:   end if
8: end while
9: return stored  $\gamma^*$ s

```

Algorithm 1: General Approximate Bayesian Computation

It is useful to establish at this stage that samples from an ABC algorithm are not samples from the true posterior. They are, in fact, samples from a distribution

that has been referred to as the ABC posterior; e.g. Fearnhead and Prangle [2010].

$$\mathbb{P}(\gamma | \mathcal{D}(\cdot, Y) < \epsilon). \quad (2.2)$$

This distribution is related to the true posterior, through the choice of distance metric $\mathcal{D}$ and tolerance ϵ . Assume that $\mathcal{D}$ has the same properties of a true distance metric; symmetric, non-negative, sub additive and identity of indiscernibles. This implies that the distance metric is equal to zero if and only if the two arguments are identical, i.e.

$$\mathcal{D}(X, Y) = 0 \Leftrightarrow X = Y. \quad (2.3)$$

It can be shown that in the limit as $\epsilon \rightarrow 0$ the ABC posterior tends in distribution to the posterior. In other words, the difference in the cumulative density functions of the ABC posterior and the true posterior tends to 0 as $\epsilon \rightarrow 0$.

$$\mathbb{P}(\gamma | \mathcal{D}(\cdot, Y) < \epsilon) \rightarrow_D \mathbb{P}(\gamma | Y) \text{ as } \epsilon \rightarrow 0. \quad (2.4)$$

2.1.3.1 Choice of Inference Algorithm

All ABC algorithms that we detail here follow the same format as the original ABC approach, detailed in Algorithm 1, however the way in which they choose parameters to generate samples change.

As noted by Csilléry et al. [2010] an advantage of the original ABC approach is its simplicity. The construction of Algorithm 1 enables one to construct trivial parallel implementations that allow the framework to be used in very large problems. Unfortunately this simplicity has a disadvantage, as noted by Marjoram et al. [2003]; it can be inefficient to iteratively sample from the prior distribution. If

the prior is necessarily diffuse to allow one to search the state space, the posterior may have a mass in a small proportion of the space thus reducing the acceptance probability, and increasing the time one has to run the algorithm to obtain the same number of samples [Marjoram et al., 2003]. Despite this the simple rejection algorithm can be used in pilot studies to inform larger, more complex studies [Csilléry et al., 2010].

To solve the problem of inefficient sampling from the prior, Marjoram et al. [2003] introduced the MCMC-ABC method that seeks to reduce the effect of low acceptance probabilities by sampling from a Markov kernel. It is identical, in its implementation, to the Metropolis Hastings [Chib and Greenberg, 1995] algorithm with the added ABC proximity constraint on the kernel [Lopes and Beaumont, 2010]. As noted by Csilléry et al. [2010], MCMC-ABC like full likelihood MCMC analysis, can suffer from mixing problems. A chain is said to have not mixed well when multiple independent chains of the same MCMC appear to result in different posterior distributions [Chib and Greenberg, 1995]. Moreover the MCMC-ABC approach can still perform poorly if the tolerance level ϵ is small, or if the proposal density isn't flexible enough to explore the parameter space [Sisson et al., 2007].

Sisson et al. [2007] and Robert et al. [2008] proposed the alternative SMC-ABC method. The principle behind SMC is to decompose the difficulty of sampling from

$$\mathbb{P}(\gamma | \mathcal{D}(S(X), S(Y)) \leq \epsilon), \quad (2.5)$$

into a series of simpler subproblems. A general SMC algorithm begins at time 0 sampling from $\mathbb{P}(\gamma | \mathcal{D}(S(X), S(Y)) \leq \epsilon_0)$, with ϵ_0 large, then simulating from an increasingly difficult sequence of target distributions $\mathbb{P}(\gamma | \mathcal{D}(S(X), S(Y)) \leq \epsilon_n)$, such that $\epsilon_n < \epsilon_{n-1}$, at subsequent time steps $n = 1, \dots, T$ [Doucet et al., 2001].

The original ABC-SMC algorithm [Sisson et al., 2007] was shown by Robert et al. [2008] to be biased, who introduced an alternative SMC-ABC method known as Population Monte Carlo (PMC).

Despite the developments in SMC a major challenge with this approach is the choice of tolerances $\epsilon_n \forall n \in \{1, \dots, T\}$. Del Moral et al. [2011] argue that if the tolerance levels decrease too slowly the algorithm will be too computationally intensive, whereas if they decrease too quickly the resulting samples will not be a good approximation to the samples from the underlying distribution. Del Moral et al. [2011] proposed the use of an adaptive learning method that can choose an appropriate tolerance level decay, using feed back loops.

2.1.3.2 Automatic Choice Dimensionality Reduction

Aside from the challenge of choosing an inference algorithm another challenge associated with ABC inference is the choice of distance measure $\mathcal{D}$ [Lopes and Beaumont, 2010]. In practice, the notion of distance $\mathcal{D}(\cdot, \cdot)$ is actually split across two methods; a dimensionality reduction function $S(\cdot)$ and a distance metric $\tilde{\mathcal{D}}(\cdot, \cdot)$. The quality of the ABC inference relies heavily on determining whether a simulated dataset is close to an observed, and as noted by Lopes and Beaumont [2010] summaries of high dimension data, $S(\cdot)$, can especially impact the quality of the results.

An original approach for choosing S relied on choosing statistics that the practitioner believed to be most informative of γ [Joyce and Marjoram, 2008]. Since the development of the original ABC algorithm more rigorous techniques to select S have been designed; e.g. using information theory to determine the marginal effect of adding additional summaries of the data [Joyce and Marjoram, 2008]. This approach was further improved by the introduction of Approximate Sufficiency [Nunes and Balding, 2010] which allowed one to find a subset of statistics that was

most informative of γ . Unfortunately both of these methods are not exchangeable, i.e. the order in which statistics are added to the set changes the results [Joyce and Marjoram, 2008, Nunes and Balding, 2010].

Fearnhead and Prangle [2010] proposed an alternative approach that in practice removed the need to chose a subset of summary statistics. Using a theoretical argument the authors show that MC error increases with the dimensionality, d , of the summary statistic function S . Therefore to control the MC error, the authors propose using a model based approach to predict the true value of γ from the summary statistics S , i.e. the posterior mean of this model is a natural estimate of γ

$$\mathbb{E}(\gamma|Y) \approx \hat{\mathbb{E}}(\gamma|S(Y)). \quad (2.6)$$

As a result they propose setting the proximity metric to be

$$\mathcal{D}(S(X^*), S(Y)) = \mathcal{D}(\hat{\mathbb{E}}(\gamma|S(X^*)), \hat{\mathbb{E}}(\gamma|S(Y))) \quad (2.7)$$

where $\hat{\mathbb{E}}(\gamma|S(X))$ is the point estimate of γ from a trained model, learning the parameters γ from $S(X) : \dim(S(.)) > p$. Fearnhead and Prangle [2010] show that they are able to fit a model using many of the metrics that are informative of the parameter of interest without adversely affecting the MC error.

2.2 Effect of Selection and Confounding

In this section we investigate the effect of selection on the distribution of metrics that capture aspects of the frequency distribution. We focus on the family of frequency spectrum statistics, because they summarise particular features of the underlying coalescent tree. We decided not to look at the family of metrics based

on functional changes, however their utility is widely acknowledged and discussed in a number of papers; Yang and Nielsen [2000], Nielsen [2001], McDonald and Kreitman [1991].

Following the discussion in Methods we simulated synthetic variation data under varying strengths of selection, allele frequencies, and demography, as shown in Table 2.1.

Model ($\sigma = 2N_es$)	Fixation Frequency	Demography
Neutral	Drift	Constant
$\sigma = 50$	{25%, 50%, 75%, 100%}	Constant
$\sigma = 100$	{25%, 50%, 75%, 100%}	Constant
$\sigma = 200$	{25%, 50%, 75%, 100%}	Constant
$\sigma = 400$	{25%, 50%, 75%, 100%}	Constant
Exponential	Drift	Exponential
Bottleneck	Drift	Bottleneck

TABLE 2.1: Table detailing the simulation scenarios, the strength of selection, the fixation frequencies in the selected scenarios, and models of demography. When fixation frequency is ‘Drift’ we do not condition on the current day frequency.

For each of the seven scenarios detailed in Table 2.1 and within each frequency bin, i.e. {25%, 50%, 75%, 100%} we simulated 5,000 datasets. The selection scenarios were simulated under the assumption that the selection strength was homogenous throughout the allele frequency trajectory [Coop and Griffiths, 2004], i.e. selection on a *de novo* mutation. For this chapter we decided not to simulate selection on standing variation so that we could focus our initial attention on one type of selection.

In this section we focus specifically on $\hat{\theta}_W$ [Watterson, 1975], D [Tajima, 1989], D^* [Fu and Li, 1993] and H [Fay and Wu, 2000]. Recall that these statistics have been shown to be sensitive to patterns of variation associated with a complete positive selective sweep, however our goal here is to establish whether they are sensitive to ongoing, or incomplete positive selection. Since only a proportion of the chromosomes carry the selected haplotype the ability of these measures to

detect selection will likely be reduced. As these statistics were introduced in the previous chapter we focus on the technical aspects of the signal that each statistic is trying to leverage, and how selection changes the distributions.

2.2.1 Window v.s. Region

One of the challenges, as mentioned in the introduction, with trying to detect signals of selection in recombining sequences is that the signal left by a selective sweep is observed only in a region surrounding the selected mutation; as observed in Figure 1.11. This presents a challenge when using univariate summaries of the genetic variation to detect selection. It is possible to calculate the statistic over the entire region, however, due to presence of recombination the signal will likely be reduced.

Another approach might be to calculate the statistic in sliding windows and use empirical distributions to establish when a region is ‘unusual’ like in many population genomics models, e.g. Voight et al. [2006]. Whilst this approach has been shown by Carlson et al. [2005] to be quite effective at establishing whether regions in a recombining genome exhibit patterns of variation associated with positive selection there is a question mark over the correct size of window. If the window is too large then the signal of selection will be reduced due to the presence of haplotypes that have recombined away from the selected haplotype. If the window is too small then the false positive rate will likely increase due to the reduction in variation that is likely to be observed in shorter sequences. Moreover the size of a selected haplotype, is affected by both the age of the mutation and the local recombination rates [Hartl and Clark, 2007], therefore a constant window size across the whole genome will likely limit the ability of the method to being able to detect other types of selection.

To avoid these challenges we base our analysis, in this chapter, on a small region (200Kb) of recombining sequences. This allowed us to focus on the task of understanding the effect of selection on summaries of the coalescent, and establishing whether it was possible to make inference.

2.2.2 Segregating Sites

$\hat{\theta}_W$ [Watterson, 1975] was one of the first metrics developed to estimate the mutation rate in a neutrally evolving population. Watterson [1975] showed that in an infinite sites model the expected number of segregating sites, K , was $\theta a_1(n)$, where θ is the mutation rate, and $a_1(n) = \sum_{i=1}^{n-1} \frac{1}{i}$.

Under a neutrally evolving population this is an unbiased estimate of the mutation rate, however as discussed in Chapter 1, in the presence of a selective sweep the number of segregating sites will tend to be reduced due to the shorter than expected time to most recent common ancestor among the chromosomes carrying the selected allele.

Figure 2.1 shows the distribution of $\hat{\theta}_W$ across the nineteen scenarios. Simulations show there is a skew towards smaller values of $\hat{\theta}_W$, as a result of the fall in the number of segregating sites. This skew appears to be most acutely observed when the selected mutation has completed its sweep, i.e. the right most panel in Figure 2.1. Observe that this skew towards lower values is also present in the exponentially expanding demographic model, but not in the bottleneck model. One explanation for this is that, due to the expanding population size, the coalescent time among the sample falls thus reducing the expected number of mutations that can accrue [Hartl and Clark, 2007]. As noted by Luikart et al. [1998] bottlenecks tend to shift the allele frequency away from low frequency alleles towards more intermediate frequencies. As observed in Figure 2.1, if there is an upward modal shift in the

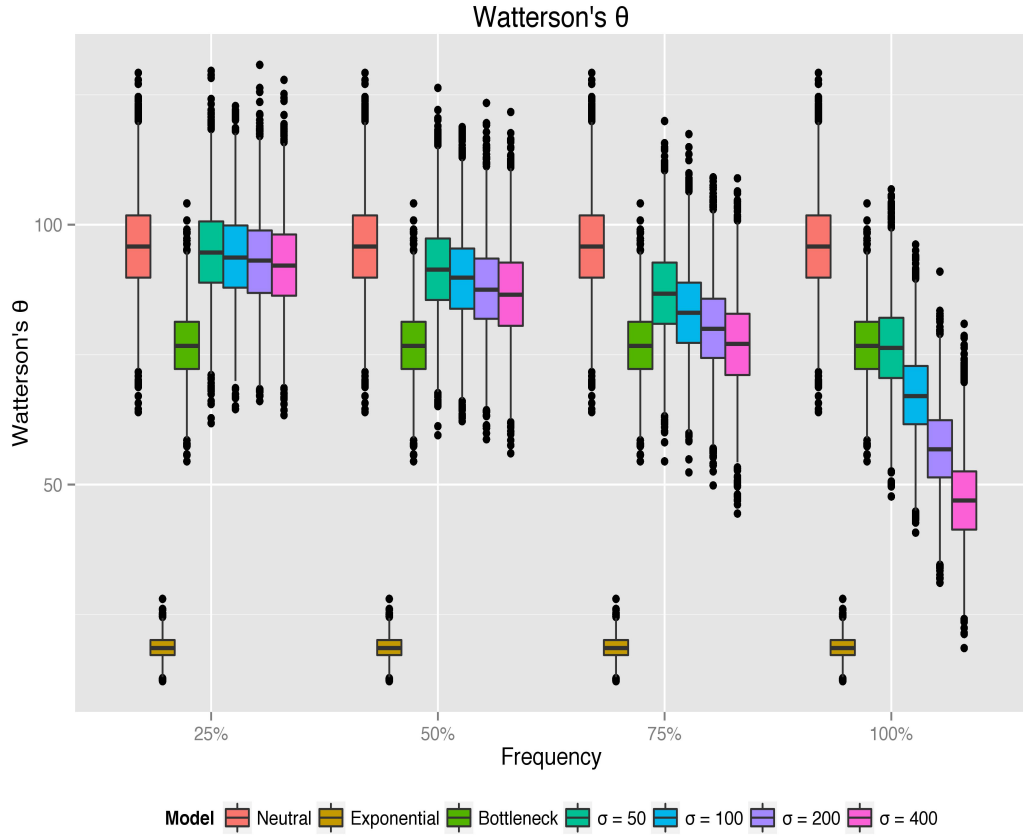


FIGURE 2.1: Effect of Selection and Demography on Watterson's Estimate. The different models of evolution are identified by the different colors, and the effect of different population frequencies of the selected allele is observed by the multiple panels. Note that the distributions for the neutrally evolving populations, and the distributions for the population experiencing the two demographic scenarios are repeated for each frequency bin.

allele frequency distribution then one might expect there to be relatively fewer rare variants.

As noted by Tajima [1989], without knowing the value of $\hat{\theta}_W$ in the neutral case it is quite difficult to establish whether a region is under selection. Using theoretical properties of the distribution of differences between two randomly chosen pairs of chromosomes, K , Tajima [1989] constructed the D statistic. Using the fact that $\mathbb{E}(\hat{K}) = \theta$, Tajima showed that in a neutral model we would expect that $\mathbb{E}(D) = 0$, because $\mathbb{E}\left(\frac{K}{a_1(n)}\right) = \theta$. In Figure 2.2 we show the effect of selection and demography D .

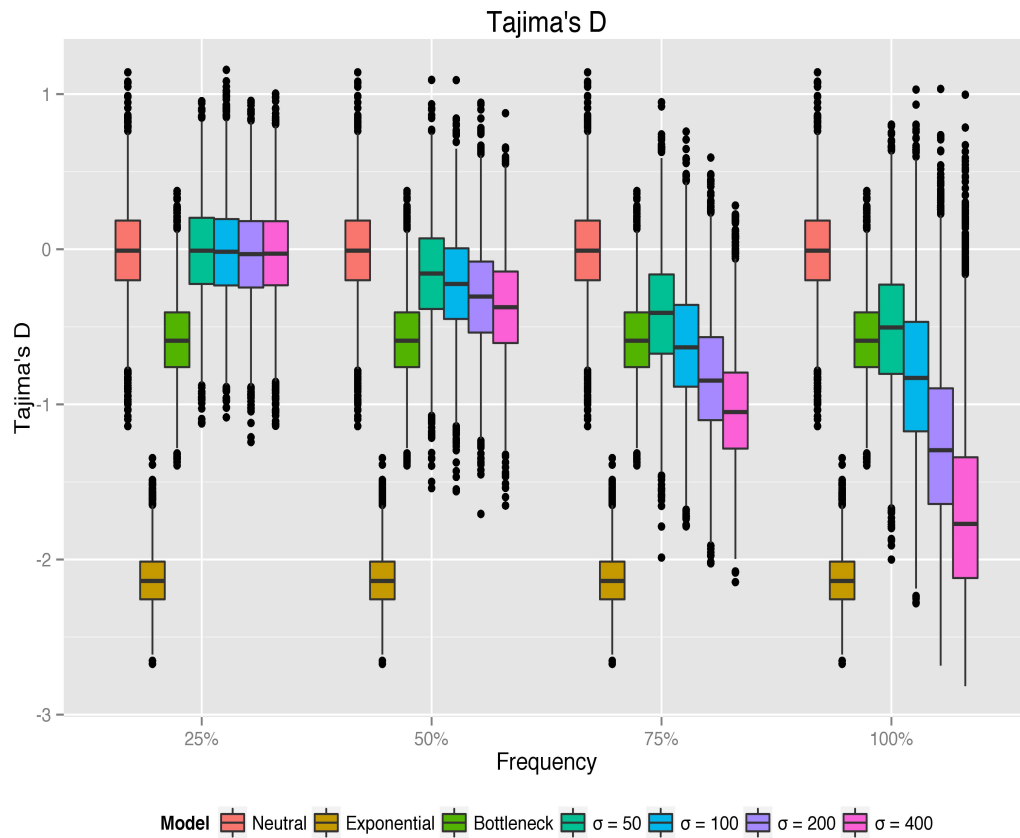


FIGURE 2.2: Effect of Selection and Demography on Tajima's D . The different models of evolution are identified by the different colors, and the effect of different population frequencies of the selected allele is observed by the multiple panels. Note that the distributions for the neutrally evolving populations, and the distributions for the population experiencing the two demographic scenarios are repeated for each frequency bin.

We discussed in Chapter 1 the effect of selection on the pairwise coalescent times as modelled by Coop and Griffiths [2004]. We observed that the time to most recent common ancestor among the selected chromosomes was depressed as a result of the trajectory of the selected allele. One of the effects of this, as noted earlier, is that the number of mutations that can accrue between the chromosomes drops in expectation. This in turn has the effect that the number of differences between the selected chromosomes is lower than we would expect, reducing D .

We observe in Figure 2.2 that there is a considerable skew towards lower values. One explanation as to why we observe a less pronounced skew under partial sweeps

is likely due to the increased heterogeneity as a result of the two groups of individuals; those carrying the derived (beneficial) allele and those carrying the ancestral. The pattern that we observe for demographic scenarios in D is quite similar to $\hat{\theta}_W$ and indeed the explanation for this observation is still valid for D .

2.2.3 Rare Variants

D^* [Fu and Li, 1993] and H [Fay and Wu, 2000] are the final two frequency spectrum metrics that we discussed in Chapter 1. Like $\hat{\theta}_W$ and D these metrics have been shown to be sensitive to the presence of selection, however, as discussed, they try to capture different aspects of the frequency spectrum. We begin by showing the effect of different evolutionary scenarios on the distribution of D^* ; Figure 2.3.

When compared with $\hat{\theta}_W$ and D we observe, in Figure 2.3, a less pronounced skew in the selection scenarios. It would appear from the plot that isn't much evidence to support the argument that D^* is sensitive to partial sweeps. We know, e.g. Coop and Griffiths [2004], that the time to a coalescent event is reduced among the chromosomes carrying the selected haplotype. As a result the number of rare variants among the sample of chromosomes will increase in expectation, because the probability that a variant will occur ancestral to chromosomes carrying the beneficial allele is reduced [Hein et al., 2005]. Increasing the number rare variants has the effect of skewing D^* towards negative values, which we observed in Figure 2.3. Note, because exponential expansion has a similar effect on all the haplotypes we see a large negative skew in those populations experiencing that demography.

It is argued by Fay and Wu [2000] that their H statistic is designed to adjust for the demographic effects by combining multiple pieces of information from different loci. Nielsen [2001] showed that the metric adds weight to new derived alleles that have spread through the population and are now at high frequency in the population.

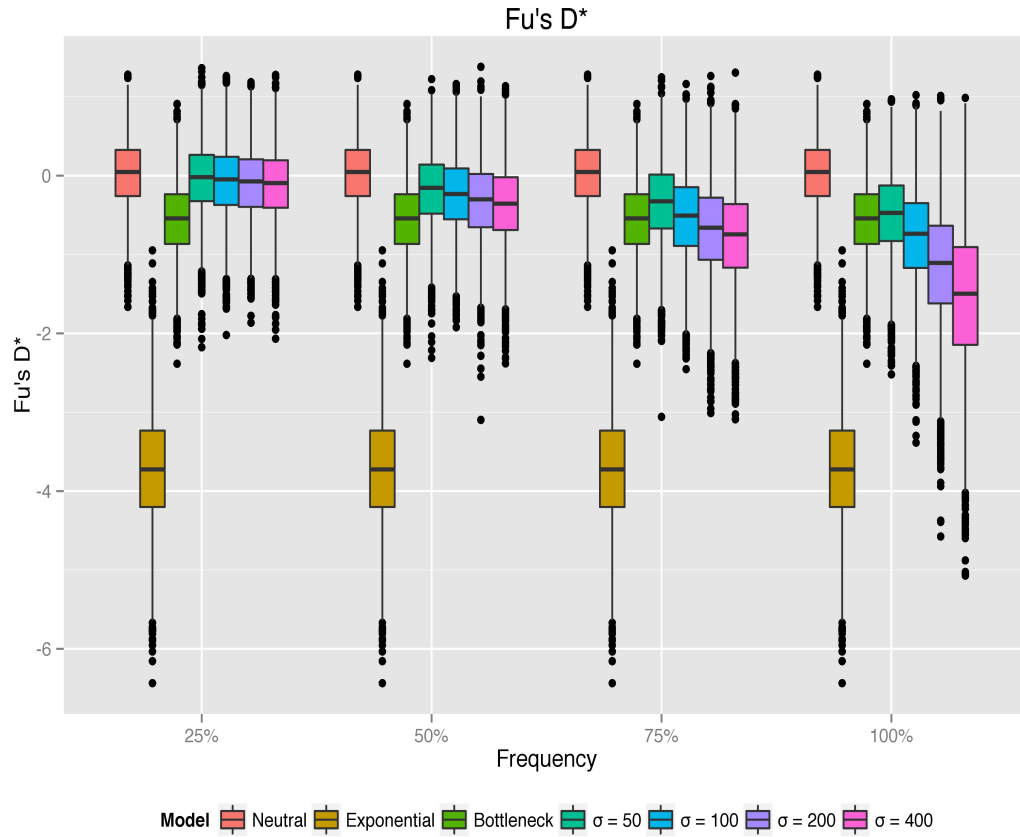


FIGURE 2.3: Effect of Selection and Demography on D^* [Fu and Li, 1993]. The different models of evolution are identified by the different colours, and the effect of different population frequencies of the selected allele is observed by the multiple panels. Note that the distributions for the neutrally evolving populations, and the distributions for the population experiencing the two demographic scenarios are repeated for each frequency bin.

Therefore the effect of selection on H , like many of the statistics that we discussed above, is likely most pronounced when a selective sweep has completed its sweep and has become fixed or nearly fixed in the population.

Our first observation from Figure 2.4 is that the demographic scenarios appear to have very similar distributions to the neutrally evolving populations. Moreover we observe that the skew towards negative values is evident in selection scenarios where the allele has reached intermediate frequencies in the population ($> 50\%$). At least in this small example, it would appear that using H [Fay and Wu, 2000] to infer selection would be a more powerful, and robust method compared to the

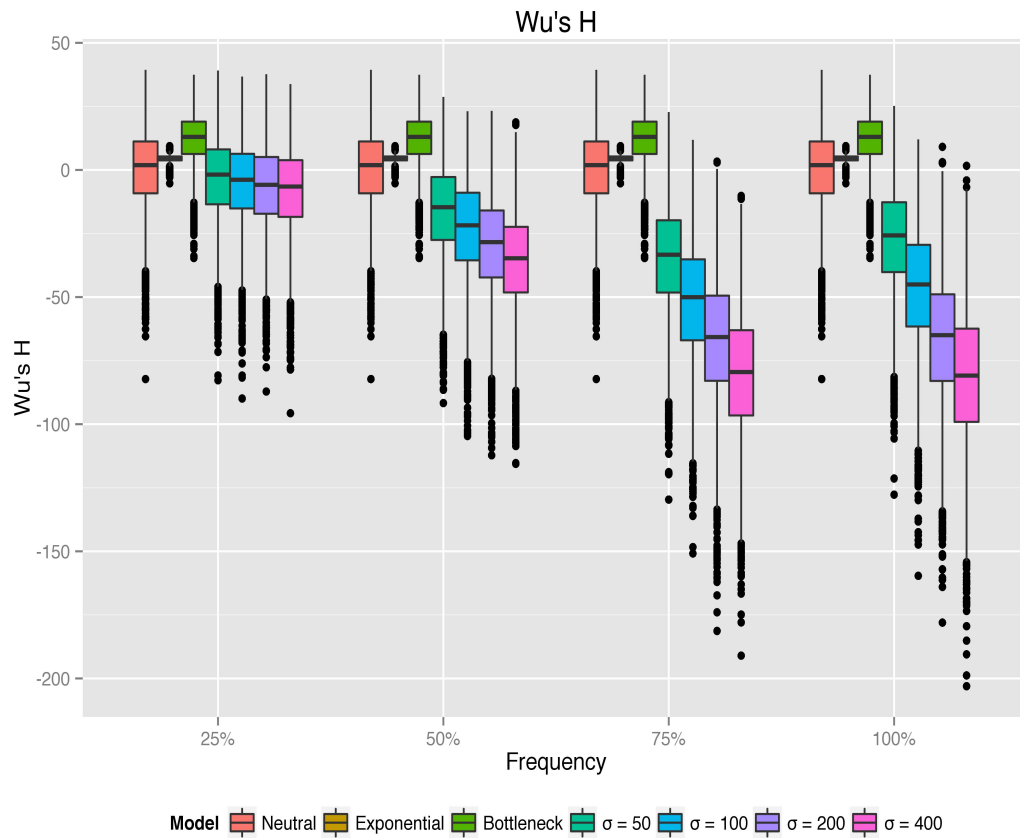


FIGURE 2.4: Effect of Selection and Demography on H [Fay and Wu, 2000]. The different models of evolution are identified by the different colors, and the effect of different population frequencies of the selected allele is observed by the multiple panels. Note that the distributions for the neutrally evolving populations, and the distributions for the population experiencing the two demographic scenarios are repeated for each frequency bin.

other three approaches.

In general, however, using these univariate metrics in real data can be challenging and result in unsuitable conclusions. Our major concern with them is that, whilst the metrics are simple to calculate, their parametric distributions will exist, at best, in simplified settings. This has implications for their general applicability to real data. As, without explicit accounting for false positive rate, it is difficult to establish a level of significance without exhaustive simulations. As a result, and as noted by Akey [2009], if we observe an ‘unusual’ value we cannot conclude that that region has evolved under selective pressure.

2.3 Predicting σ - Machine Learning

The metrics that we discussed above tackle the problem of inferring the presence of selection by summarising different aspects of the allele frequency spectrum. It would appear that inferring the strength of selection from any one of these metrics would be challenging, so an interesting question to ask is whether we might be able to combine these metrics into an inference framework and train a model to estimate the strength of selection.

Summary Statistics	
Measure	Description
$\hat{\theta}_W$ [Watterson, 1975]	Across Entire Region
D [Tajima, 1983]	Across Entire Region
Pairwise Difference K	Median $K_{(\frac{1}{2})}$
	Lower Quartile $K_{(\frac{1}{4})}$
	Upper Quartile $K_{(\frac{3}{4})}$
D^* [Fu and Li, 1993]	Across Entire Region
H [Fay and Wu, 2000]	Across Entire Region

TABLE 2.2: A list of metrics that we used in the ML techniques detailed in Methods.

So that we could establish whether this was possible we limited our investigations to a sample of recombining chromosomes that come from a population of constant size. To make inference we appealed to the machine learning (ML) tools that we detailed in Methods.

We began by simulating 1×10^6 genetic variation datasets; each region was constructed using the same parameters as the datasets in the previous section, however this time we draw the selection rate, σ , from a uniform distribution on the interval $[0, 400]$. Moreover we fixed the allele frequency at 75% so as to consider the problem of inferring the strength of selection in ongoing positive selection. Each region was mapped to a vector of summary statistics which we have detailed in

Table 2.2. Since we cannot include the whole pairwise difference distribution in the model we decided to summarise it using its quartiles.

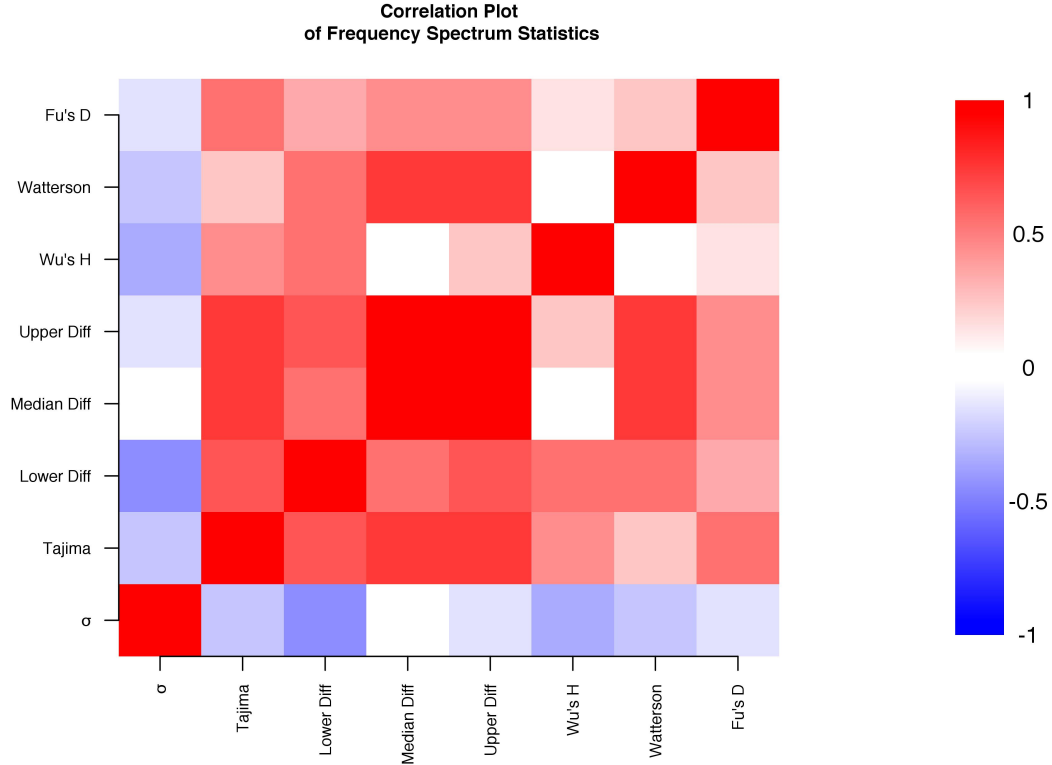


FIGURE 2.5: Estimate of the correlation matrix between the summary statistics in Table 2.2 and the strength of selection $\sigma = 2N_e s$. Calculated using 1×10^6 genomic regions with σ random sampled from Uniform[0, 400].

Using these statistics and the data we calculated an estimate of the correlation matrix between the metrics and the strength of selection; Figure 2.5. Due to the construction of D [Tajima, 1983] we observe strong correlation structure between summaries of the pair difference distribution. The observation that the correlation, outside of $\left(D, K_{(\frac{1}{2})}, K_{(\frac{1}{4})}, K_{(\frac{3}{4})}\right)$, is lower supports our earlier suggestion that these metrics are using different features of the allele frequency distribution to make inference.

Using the simulated σ and the corresponding values of the summary statistics we explored a number of different linear parametric, and non-parametric models. As detailed in Methods we chose three different ML regression models and used the

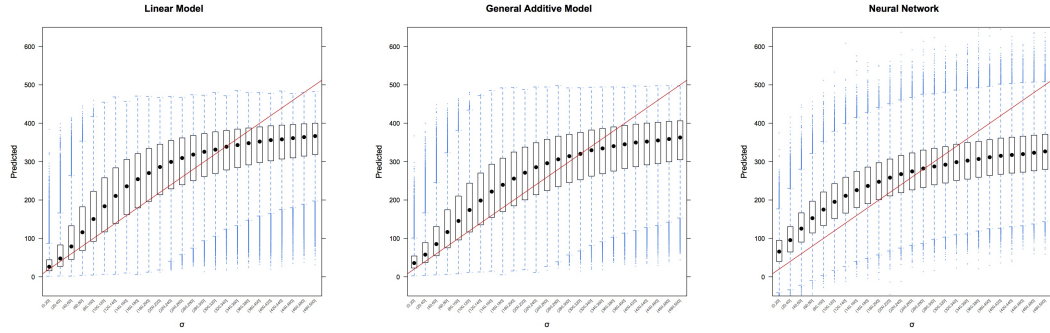


FIGURE 2.6: Each plot shows, for a given model, how the distribution of fitted values change relative to changes in the truth. The left panel shows the results for a General Linear Model (GLM), the center panel shows a General Additive Model (GAM), and right panel left shows a Neural Network (NN). For a discussion of these models refer to Methods. The ends of the boxplots are calculated using lower 2.5% and upper 97.5% quantiles in the distribution.

simulated dataset of 1×10^6 observations to train them. Figure 2.6 shows the observed changes in the relationship between the fitted value of the models, $\hat{\sigma}$, as the true σ is varied.

We observe evidence of decreasing effectiveness for each model to distinguish stronger strengths of selection from intermediate strengths. It may be because the metrics, that the models are based on, are unable to distinguish between the variation of $\sigma = 200$ and $\sigma = 400$, however it could also be an effect of size of the region that we are simulating. Since we have chosen to simulate a 200Kb sequence it is possible that the selected haplotypes under very strong selection will extend beyond the region. This will reduce our ability to detect the marginal differences between scenarios at the upper end of the range of σ .

It would appear that none of the models produce a satisfactory relationship between the fitted and the observed, however of the three models in Figure 2.6 it would appear that the neural network produces the more accurate estimates. As evidenced by the reduced heteroscedasticity.

Despite this, we observe that there is a high degree of variation in the results above. This, together with the observation that the prediction distributions have

a large amount of overlap, suggests that it will be challenging to use either of these methods, together with these summary statistics, to obtain reliable estimates of σ . As a result, we decided to turn our attention towards the Approximate Bayesian Computation (ABC) paradigm and determine whether this method results in any improvement.

2.4 Predicting σ - ABC

ABC has been used very successfully as an inference framework in a number of areas of population genetics [Tavaré et al., 1997, Beaumont et al., 2002, Lopes and Beaumont, 2010]. Since we introduced the approach above, we focus here on trying to answer the question; can ABC be used to detect selection in genetic variation data? Note that we limit our attention to the original ABC rejection method [Tavaré et al., 1997]. We have explored the use of more advanced inference algorithms, however because this is, in essence, a feasibility study we took the view that it was more important to demonstrate the method in the simplest setting.

As with the results from the Machine Learning section from above, we decided to simplify the inference task by reducing the problem to simply inferring the strength of selection, σ , by treating everything else about the evolution of a sample of chromosomes as known, e.g. location and population frequency of selected allele, recombination rate, and mutation rate, etc.

The rejection algorithm was simple to implement and combine with the simulation software *msms* [Ewing and Hermisson, 2010]. The whole pipeline was implemented using the Linux operating system, and the ABC algorithm was constructed using a simple concurrent *Java* application.

One of the most important steps in the design of an ABC experiment is to choose the method that will determine the proximity between a simulated and the true

dataset. In Methods we discussed different approaches that have been developed over the past decade that aim to improve the quality of the ABC inference. In our investigations we explored a number of different options, e.g. automated choice [Fearnhead and Prangle, 2010], however we decided for this thesis to limit our attention to the classical model of ABC [Tavaré et al., 1997]. In this model we choose a set of summary statistics that are informative of the parameter of interest and use a Euclidean distance measure to determine proximity between datasets.

Using the software, *msms* [Ewing and Hermisson, 2010], we simulated four datasets, where the selection rate is known; $\sigma = \{0, 50, 100, 200\}$ and derived allele frequency fixed at 75%. Using our ABC pipeline we calculated the summary statistics, detailed in Table 2.2, on a large number (1×10^6) synthetic regions, sampling σ from Uniform[0, 400]. We used a standardised Euclidean metric to compare the distance between the estimates of the summary statistics under the truth relative to the simulated. As discussed in Methods we accept simulated parameters if the distance between the simulated and observed datasets fall below some tolerance ϵ . Figure 2.7 shows the effect of changing ϵ on the distribution of the accepted values of σ ;

Generally speaking the distributions of the accepted parameter values tend to have modes near to the true values of the parameters, however as with the ML methods the variance of the distributions is considerable. Indeed we see that in the majority of cases the ABC posterior credible interval covers over 50% of the parameter space. One explanation for this is that we only have 100 observations at the lowest tolerance level which will necessarily inflate the variance. Finally, in the lower panel of Figure 2.7 we observe that at the lowest tolerance level there appears to be a reduction in variance of the accepted parameters. Moreover we observe that for strong selection $\sigma = 200$ the method tends to favour accepting parameters that occur in the upper end of the space. Even through the variance is

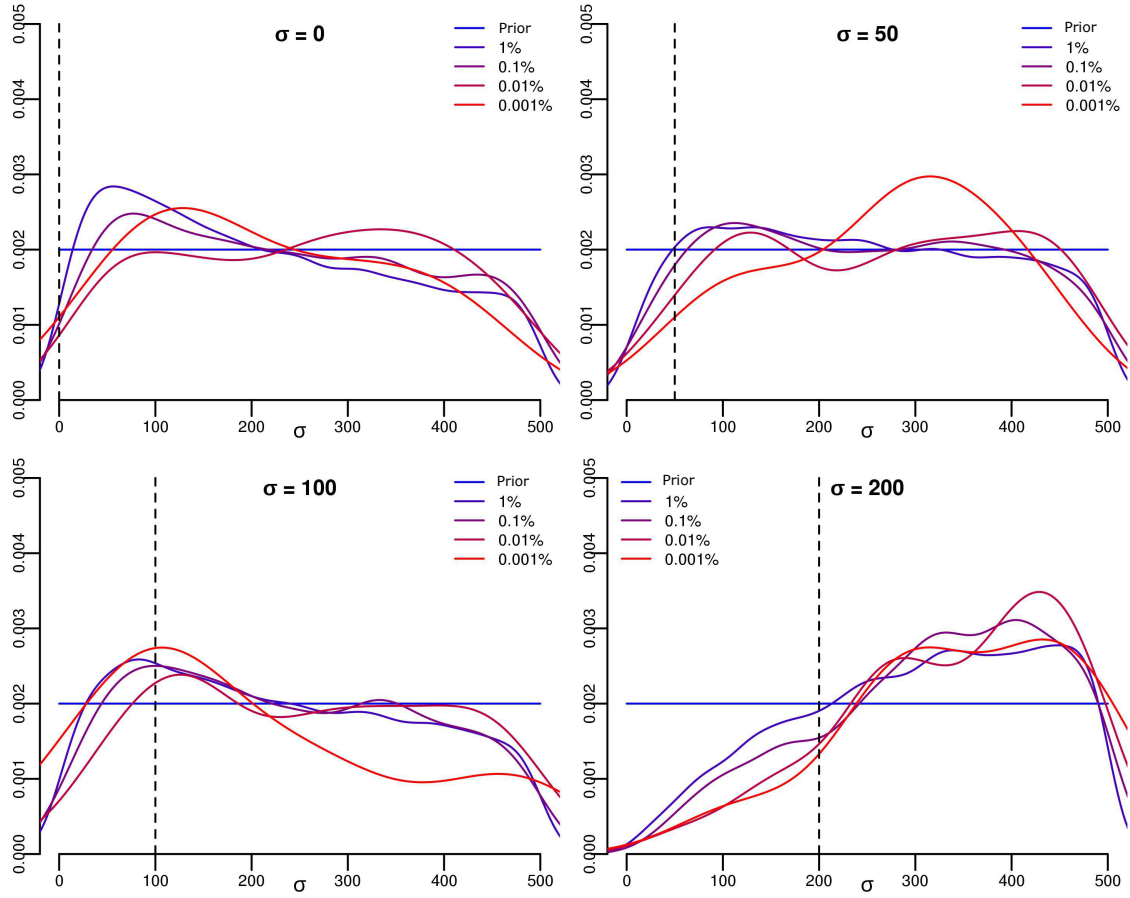


FIGURE 2.7: Effect of changing the value of ϵ on the distribution of the accepted parameter values. The densities for σ were constructed by establishing the distribution of the parameter values if we accepted the top 10%, 1%, 0.1%, 0.01%. The values for each ϵ are detailed in the plots.

still high this observation may suggest that as the strength of selection increases the ability of the method to detect selection improves.

2.5 Discussion

We began this chapter with an aim of better understanding the effect of ongoing selective sweeps on the distribution on four frequency spectrum metrics. We discovered that all of the statistics were sensitive to selection when the allele was fixed in the population, however as the population frequency of the selected allele dropped so did the metrics ability to detect departures from neutrality. As we

discussed above, one explanation for this is that, all other things being equal, only those haplotypes who carry the selected allele will exhibit patterns of variation associated with positive selection. As such the effectiveness of the metrics will be reduced as the signal of selection across all chromosomes will be reduced.

Moreover we noted that, whilst these metrics are sensitive to patterns of variation associated with selection, with the exception of H [Fay and Wu, 2000], they are also sensitive to demographic changes; e.g. exponential expansion and bottlenecks. We explained that, because these demographic scenarios often have a similar impact on the variation as selection it is difficult to disentangle the two features using these summaries of the coalescent.

We then decided to look at whether we could combine these metrics together to make inference on the strength of selection in the simplest setting when the only unknown parameter was σ . We showed that, whilst it was difficult to estimate the exact value of σ because of the large variation in the predictions, the results suggest that these methods are able to detect whether there is strong or weak selection. We then looked at using large scale simulations to estimate σ using the ABC inference framework. As with the ML methods we concluded that whilst the method seems to be capable of detecting the presence of strong selection ($\sigma = 200$) estimating the true value would present a significant challenge.

If we look holistically at the results in this chapter we might conclude that detecting partial sweeps of selection using these summaries of the coalescent will not be an effective, or robust approach to detecting signals of selection in real data. For both the ML methods and ABC we simplified the inference problem by assuming that the only parameter that was not known was σ . This implicitly assumes that we know the effective population size, and how the population evolved through time. Clearly, these values will not be known exactly, and whilst we can use different methods to estimate these characteristics an incorrect assumption could bias

the results, e.g. Figures 2.1 - 2.3.

These results were compiled two years ago, and in that time a number of papers have been published that try to make inference on selection using an ABC approach, e.g. Peter et al. [2012]. In this paper the authors propose an approach that uses ABC to both estimate the selection parameter and infer whether the selected mutation came from a new mutation or from standing variation. Using a rejection based ABC sampling method [Tavaré et al., 1997], in conjunction with a set of summary statistics the authors present their results from both simulated data and real data. In addition to the statistics in Table 2.2 the authors propose using iHS [Voight et al., 2006].

Peter et al. [2012] conclude that their implementation of ABC has little power to detect the true value of $\sigma \leq 100$. When $\sigma > 100$ the authors show that they can estimate the true value of σ , however like the results in this chapter they show that the estimates are highly variable. Unfortunately, the quality of their inference method is difficult to compare with ours as the results that they present in relation to accuracy of the parameter estimates have no scale. Moreover the authors only present the results of their parameter inference when $\sigma = 400$.

This comparison helps to highlight one of the most critical aspects in an ABC inference; choice of summary statistics. As we mentioned in the introduction, poor choice of summaries can severely impact the quality and accuracy of the method, and deciding on the right set of summary statistics has been an important area of research over the past decade. A number of methods, e.g. Fearnhead and Prangle [2010] have been developed to help reduce the impact of poor choices by model fitting. These methods are very promising, however they also introduce additional tuning parameters, e.g. model choice, which adds additional complexity.

An interesting feature of the Peter et al. [2012] method is its ability to distinguish between models of selection on standing variation and *de novo* mutations. Whilst

this is a useful feature their method only has power to distinguish between the two models of selection when $\sigma \geq 1000$, and when the frequency of the allele at the point when it became beneficial is sufficiently segregated in population, i.e. $> 10\%$. Moreover these results were constructed when the allele had reached near fixation in the population, i.e. $f = 95\%$.

Taking a step back we see that there is a level of agreement between our results and the approach by Peter et al. [2012], however as discussed direct comparisons are difficult. These results suggest that while ABC has been shown to be effective in various population genetic settings, its application to selection has many caveats; σ needs to be high, and the beneficial allele needs to be highly segregating in the population; $f > 75\%$.

One could make the argument that alternative summary statistics to those detailed in Table 2.2 could alter the effectiveness of the inference approach. Indeed one might also argue that alternative methods by which we assess proximity, or perform the inference, i.e. alternative ABC algorithms, could make the inference more efficient. Whilst these are fair criticisms it doesn't remove the fact that to make accurate inference using these methods requires knowledge of aspects of the evolutionary past which can be difficult to establish, i.e. simulating from the true model is only possible in simulated experiments.

Chapter 3

A Bayesian Chromosome Painting Approach To Detecting Genomic Regions Under Selection

“Essentially, all models are wrong, but some are useful.”

George Box

3.1 Introduction

In the previous chapter we discussed the challenges associated with using the coalescent likelihood to perform inference on the strength of selection. Whilst this is a well established challenge in modern population genomics, there are alternative methods that have been shown to be sensitive to the presence of classical selective sweeps. Many of these methods focus on finding regions of the genome with extended haplotype homozygosity (EHH) [Sabeti et al., 2002b, Voight et al., 2006, Sabeti et al., 2007, Albrechtsen et al., 2010]. As we saw in Chapter 1, the classical

model of a selective sweep describes the process by which a mutation arises and gets swept up to a high frequency significantly faster than one might expect under genetic drift [Smith and Haigh, 1974]. Haplotypes carrying the selected allele will tend to exhibit increased linkage disequilibrium (LD) at surrounding sites i.e. the ‘hitchhiking’ effect [Smith and Haigh, 1974], therefore increasing haplotype homozygosity further than would be expected by chance [Sabeti et al., 2002b].

Results in Sabeti et al. [2002b], Voight et al. [2006], and Sabeti et al. [2007], suggest that methods based on haplotype homozygosity decay can be a powerful approach to detecting signals of selection. These approaches use a range of non-parametric methods to detect regions of the genome that have extreme values of a statistic that is sensitive to patterns of variation associated with a classical sweep [Akey, 2009]. As a result they may not have power to detect alternative modes of positive selection, i.e. selection on standing variation. In this chapter we introduce a novel model based approach to determine regions of genome that have patterns of variation associated with both classical sweeps and soft sweeps.

We begin this chapter by introducing an approximate model that captures the main features of the coalescent, that has become popular for the analysis of genome wide variation data. We then show how this model can be used to infer recent ancestry among a sample of chromosomes. By discussing the impact of selection on the recent ancestry we show how it is possible to use this model to identify regions of a genome that have recently, or are currently experiencing positive selection. Finally, we introduce a simple Bayesian model that quantifies the evidence in support of a selectively non-neutral hypothesis.

3.2 HMMs in Population Genetics

The challenge of using the coalescent likelihood in Equation 1.2 to make inference on ancestral processes using genomic data is a well known problem in population genomics. The added challenge of recombination increases the complexity of Equation 1.2 further by allowing the ancestry between haplotypes to depend on the position in the genome [Griffiths, 1981, Hudson, 1983].

We showed in the previous chapter that using a Monte Carlo approach also has a number of significant challenges even in simple settings. To help alleviate some of these challenges and those surrounding using Equation 1.2, to directly perform inference, a number of methods have been proposed to tackle specific problems, e.g. recombination rate estimation [Hudson, 2001]. Within the past decade one particular class of models, Hidden Markov Models (HMMs) have risen to prominence within the field. The Markovian assumption makes this class of model particularly useful in population genomics as it allows full likelihood inference to be conducted using chromosome, or genome-wide variation data.

3.2.1 LSHMM

One of the first uses of HMMs in population genetics was a model proposed by Li and Stephens [2003] that still forms the basis of many modern tools in the field, e.g. Scheet and Stephens [2006], Delaneau et al. [2008], and Delaneau et al. [2013]

Rather than approaching the inference problem by approximating Equation 1.2 directly, Li and Stephens [2003] proposed a method that approximated the elements in the decomposition of the likelihood. Unless otherwise stated we assume that the haplotypes that we observe are haploid (or phased diploid), such that they take values in $\{0, 1\}$.

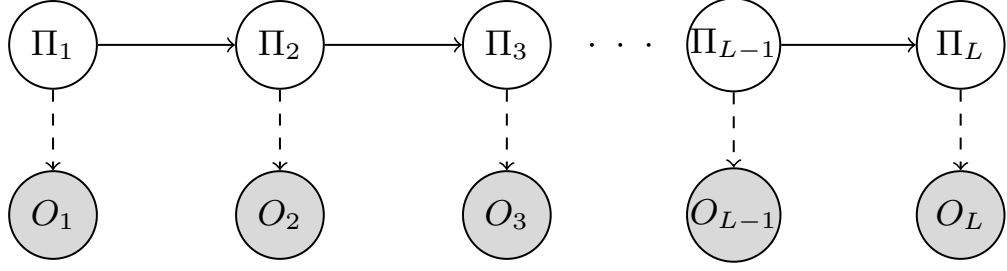


FIGURE 3.1: Schematic showing the general set up in a Hidden Markov Model. The hidden states are denoted by the Π_i random variable, and transitions between these states, i.e. solid arrows, are governed by the transition probability distribution, $A_{i,j}^{(l+1)} = \mathbb{P}(\Pi_{l+1} = j | \Pi_l = i)$. The observed states, denoted by O_i random variables, are emitted from Π_i by the emission distribution, $E^{(l)}(o_l) = \mathbb{P}(O_l = o_l | \Pi_l)$, i.e. dashed arrows.

For a sample, $\{H_j^{(i)}\}_{i,j=1}^{n,L} \in \{0,1\}^{n \times L}$, of n haplotypes at L segregating sites the likelihood decomposition can be expressed as follows

$$\mathcal{L}(\rho, \theta | H) = \mathbb{P}(H^{(1)} | \rho, \theta) \mathbb{P}(H^{(2)} | H^{(1)}, \rho, \theta) \dots \mathbb{P}(H^{(n)} | H^{(1)}, \dots, H^{(n-1)}, \rho, \theta). \quad (3.1)$$

To perform inference Li and Stephens [2003] estimate each element in the product using a HMM, which they then combine to form an estimate of the likelihood in Equation 3.1. The authors term this model, the product of approximate conditionals (PAC).

An HMM has two key components; the observed states $\{O_1, \dots, O_L\}$, and the hidden states $\{\Pi_1, \dots, \Pi_L\}$. The hidden states are modelled as a finite state, discrete time Markov chain, with transitions governed by transition distribution, $A_{i,j}^{(l+1)} = \mathbb{P}(\Pi_{l+1} = j | \Pi_l = i)$, with the observed states emitted from hidden states via an emission distribution, $E^{(l)}(o_l) = \mathbb{P}(O_l = o_l | \Pi_l)$ [Rabiner and Juang, 1986]. A schematic for a simple HMM can be seen in Figure 3.1.

The emission and transition densities in the $k+1^{th}$ Li and Stephens Hidden Markov Model (LSHMM), used to estimate $\mathbb{P}(H^{(k+1)} | H^{(1)}, \dots, H^{(k)}, \rho, \theta)$, can be expressed as follows. Between loci l and $l+1$, define d_l to be the distance, measured in base

pairs (BP), and r_l to be the average recombination rate, measured in Morgans per BP, and define N_e to be the effective population size. θ is defined as the mutation rate, which is often set to Watterson's estimate [Li and Stephens, 2003]. The transition and emission densities for $k + 1^{th}$ LSHMM can then be written follows;

$$a_{i,j}^{(l)} = \begin{cases} \exp(-r_l d_l \frac{4N_e}{k}) + (1 - \exp(-r_l d_l \frac{4N_e}{k})) / k & \text{if } i = j \\ (1 - \exp(-r_l d_l \frac{4N_e}{k})) / k & \text{o/w,} \end{cases} \quad (3.2)$$

$$e_{h_l^{(k+1)}}^{(\theta)}(h_l^{(i)}) = \begin{cases} k/(k + \theta) + \frac{1}{2}\theta/(k + \theta) & \text{if } h_l^{(i)} = h_l^{(k+1)} \\ \frac{1}{2}\theta/(k + \theta) & \text{o/w.} \end{cases} \quad (3.3)$$

Note, in this example that the hidden states are the previously observed k haplotypes, i.e. $H^{(1)}, \dots, H^{(k)}$. As argued by Li and Stephens [2003] we can consider the $k + 1^{th}$ haplotype as being a mosaic of the previous k haplotypes. As a result, these distributions imply that as the recombination rate r_l increases the probability that the $k + 1^{th}$ haplotype will stay in the current state decreases. This reflects the ARG where an increase in the recombination rate increases the number of trees and decreases the dependency between adjacent loci. Moreover, an increase in the mutation rate θ implies that the probability the $k + 1^{th}$ haplotype copies from one of the previous k haplotypes without error will decrease.

The accuracy of this method, as argued by Li and Stephens [2003], depends significantly on the accuracy with which it can approximate each element of the product. By building on previous work, e.g. Stephens and Donnelly [2000] and Fearnhead and Donnelly [2001], the authors constructed the transition (Equation 3.2) and emission densities (Equation 3.3), which they argue capture general characteristics regarding the relationship between haplotypes [Li and Stephens, 2003].

3.2.2 HMM Inference

Using HMMs to make inference can be very efficient owing to the development of algorithms that exploit the Markov structure of the model. Li and Stephens [2003] describe a particular application of one method, but depending on the type of inference there are three commonly used algorithms:

- The **Forward / Backward** [Rabiner and Juang, 1986] procedure allows calculation of the likelihood and the posterior probability of the hidden Markov chain occupying any state at any time point.
- The **Viterbi** [Viterbi, 1967] algorithm infers the most likely path through the hidden states given the observations.
- The **Baum / Welch** [Baum, 1972] method infers the transition and emission probabilities using an algorithm based on the expectation maximisation principle.

All three algorithms have linear complexity in the number of observations, e.g. positions in a genome [Durbin, 1998]. This allows HMMs to be applied to very dense datasets (i.e. genome-wide data) [Scheet and Stephens, 2006, Delaneau et al., 2008, Paul and Song, 2012, Zhang, 2013, Delaneau et al., 2013]. Unfortunately, this linear complexity does not extend to the number of states which, in general, has been shown to have quadratic complexity [Durbin, 1998].

3.2.2.1 Forward / Backward Algorithm

The Forward / Backward algorithms are used to determine the likelihood of an HMM given the hidden states and the parameters and either method can be used to calculate the likelihood [Rabiner and Juang, 1986]. These algorithms use the

Markov structure of a HMM to construct a simple dynamic programming statement to sum over all possible paths [Rabiner and Juang, 1986]. Often one method is used in conjunction with the other to sample paths, or find the expected holding times as in Lawson et al. [2012]. The Forward algorithm with notation from before is detailed in Algorithm 2

Data: $\left\{H_l^{(k+1)}\right\}_{l=1}^L, \left\{H_l^{(i)}\right\}_{l=1, i=1}^{L, k}$.
Result: $\pi^{(k+1)} \in \{1, \dots, k\}^L$.
Initialisation ($l = 0$): $f_0(0) = 1, f_i(0)$ for $i > 0$;
while $l \leq L$ **do**
 $f_i(l) = e_{h_l^{(k+1)}}^{(\theta)}\left(h_l^{(i)}\right) \sum_j f_j(l-1) a_{j,i}^{(l-1)}$;
end
Termination: $\mathbb{P}\left(H^{(k+1)}|H^{(1:k)}, \mathbf{r}, \theta\right) = \sum_j f_j(L)$;

Algorithm 2: Forward Algorithm adopted from Durbin [1998] with notation from Li and Stephens [2003].

The Backward algorithm with notation as before is described in Algorithm 3.

Data: $\left\{H_l^{(k+1)}\right\}_{l=1}^L, \left\{H_l^{(i)}\right\}_{l=1, i=1}^{L, k}$.
Result: $\pi^{(k+1)} \in \{1, \dots, k\}^L$.
Initialisation ($l = L$): $b_i(L) = 1$;
while $l \geq 1$ **do**
 $b_i(l) = \sum_j e_{h_{l+1}^{(k+1)}}^{(\theta)}\left(h_{l+1}^{(j)}\right) a_{i,j}^{(l)} b_j(l+1)$;
end
Termination: $\mathbb{P}\left(H^{(k+1)}|H^{(1:k)}, \mathbf{r}, \theta\right) = \frac{1}{k} \sum_j e_{h_1^{(k+1)}}^{(\theta)}\left(h_1^{(j)}\right) b_j(1)$;

Algorithm 3: Backward Algorithm adopted from Durbin [1998] with notation from Li and Stephens [2003].

As noted above, in general, the Backward and Forward algorithms have linear complexity in the number of positions and quadratic complexity in the number of states, i.e. $\mathcal{O}(k^2 L)$. In the LSHMM, however, it is possible to use the structure of the transition probabilities for reduce the complexity of both algorithms to linear

in the number of states, i.e. $\mathcal{O}(kL)$. Define $p_l = \exp(-r_l d_l \frac{4N_e}{k})$, then

$$f_i(l) = e_{h_l^{(k+1)}}^{(\theta)}(h_l^{(i)}) \sum_j f_j(l-1) a_{j,i}^{(l-1)} \quad (3.4)$$

$$= e_{h_l^{(k+1)}}^{(\theta)}(h_l^{(i)}) \sum_j f_j(l-1) \left[p_{l-1} \mathbb{I}_i(j) + \frac{(1-p_{l-1})}{k} \right] \quad (3.5)$$

$$= e_{h_l^{(k+1)}}^{(\theta)}(h_l^{(i)}) \left[p_{l-1} f_i(l-1) + (1-p_{l-1}) \sum_j f_j(l-1) \right] \quad (3.6)$$

Since only the first term in the square parentheses is dependent on i , the second term can be pre-computed at each iteration. As a result, the number of loops through the states is limited to two, i.e. it has linear complexity in states.

In addition to using these algorithms to calculate the likelihood, or sample paths, one can combine the values of $f_i(l)$ in Algorithm 2 and $b_i(l)$ in Algorithm 3 to determine the posterior probability at each position of the observed state being emitted from one of the hidden states [Rabiner and Juang, 1986], i.e.

$$\mathbb{P}(\Pi_l = i | H^{(1:k+1)}) = \frac{f_i(l) b_i(l)}{\hat{\mathbb{P}}(H^{(1:k+1)})}. \quad (3.7)$$

Evaluating Equation 3.7 is a procedure known as posterior decoding and can be an alternative inference approach when there are many paths through the hidden states have similar probability to the most probable path [Rabiner and Juang, 1986].

3.2.2.2 Viterbi Algorithm

Rather than calculating the value of the likelihood the Viterbi algorithm can be used to decode the HMM and determine the sequence of states through the state space that maximise the likelihood of the HMM given the emission and transition parameters [Durbin, 1998]. The Viterbi algorithm determines the most likely path using the Markovian structure of the HMM and a dynamic programming

argument [Rabiner and Juang, 1986]. This means that the Viterbi algorithm, shown in Algorithm 4, also has linear complexity in the number of positions and quadratic complexity in the number of states, i.e. $\mathcal{O}(k^2L)$. However, as in the Forward / Backward algorithms we can reduce the complexity to $\mathcal{O}(kL)$

Data: $\left\{H_l^{(k+1)}\right\}_{l=1}^L, \left\{H_l^{(i)}\right\}_{l=1, i=1}^{L, k}$.
Result: $\pi^{(k+1)} \in \{1, \dots, k\}^L$.
Initialisation ($l = 0$): $v_0(0) = 1, v_i(0)$ for $i > 0$;
while $l \leq L$ **do**
 $v_i(l) = e_{h_l^{(k+1)}}^{(\theta)}(h_l^{(i)}) \max_j (v_j(l-1) a_{j,i}^{(l-1)});$
 $\text{ptr}_l(i) = \text{argmax}_j (v_j(l-1) a_{j,i}^{(l-1)});$
end
Termination: $\pi_L^{(k+1)} = \text{argmax}_j (v_j(L));$
Traceback ($l = L, \dots, 1$): $\pi_{l-1}^{(k+1)} = \text{ptr}_l(\pi_l^{(k+1)});$

Algorithm 4: Viterbi Algorithm adopted from Durbin [1998] with notation from Li and Stephens [2003].

3.2.2.3 Parameter Inference

A particular challenge with using HMMs as an inference framework is that of learning the parameters for the transition and emission probability distributions [Durbin et al., 2010]. In their original implementation Li and Stephens [2003] propose using the golden bisection algorithm developed by Kiefer [1953]. It is also possible to use more complex, but potentially more efficient algorithms like Baum / Welch [Baum, 1972] method, which uses a combination of the Forward / Backward algorithms and the Expectation - Maximisation [Dempster et al., 1977]. These algorithms can be computationally expensive particularly when the likelihood surface is not log concave. In such settings, it may be necessary to run the algorithm multiple times, each from a different diffuse initialisation to help reduce the impact of local maxima.

3.2.3 Examples of HMMs in Population Genetics

The PAC model proposed by Li and Stephens [2003] provided a framework to describe the variation observed across a genome in terms of the recombination rate ρ , and mutation rate θ . Using simulated data Li and Stephens [2003] show that their PAC based estimates of ρ and θ were very accurate, however their original approach was shown to have systematic bias which they correct using empirical results. Even though this method is based on a HMM structure, the computational cost of making inference about the variation in ρ along a genome is significant. Later more efficient and accurate models were developed to infer recombination maps across the genome, e.g. McVean et al. [2004], however it still forms the basis of a number of different applications in modern genomics.

In addition to estimating the parameters, Li and Stephens [2003] describe a number of potential side applications of their model of LD. In particular they describe how their model could be used, in conjunction with PHASE [Stephens et al., 2001], to infer recombination rates, chromosomal phase, and impute missing genotypes [Stephens and Donnelly, 2003].

Since the introduction of the PAC model many other methods that use a HMM structure have been proposed. Some are based on the work by Li and Stephens [2003], like the model proposed by Delaneau et al. [2008] which helps to alleviate some of the challenges surrounding the quadratic complexity of HMMs in the size of the state space. By fixing the size of the state space this proposal, together with more recent modifications [Delaneau et al., 2013], allowed the phasing and imputation of very large sample sizes, and the same time maintaining the accuracy of the original implementations. Other HMM methods not based on the work by Li and Stephens [2003] have also been proposed. For example Zhang [2013] propose using a stochastic state space HMM to simultaneously impute, infer genotype, and phase raw data from next generation sequencing machines.

In addition to these applications, the model by Li and Stephens [2003] has recently been used to describe the relatedness between a sample of chromosomes [Lawson et al., 2012]. The method, termed chromosome painting, uses a version of the LSHMM, termed the pseudo likelihood (PL) model [Li and Stephens, 2003], to relate patterns of variation observed in haplotypes to the recombination rate, ρ .

3.3 Chromosome Painting

The pseudo likelihood (PL) model, proposed by Li and Stephens [2003], is constructed in a similar way to the PAC model, but rather than relating each new haplotype to the previously observed haplotypes it relates every haplotype to every other haplotype, i.e.

$$\mathcal{L}(\rho, \theta | H) \approx \prod_{i=1}^n \mathbb{P}(h^{(i)} | \{h^{(1)}, \dots, h^{(n)}\} \setminus \{h^{(i)}\}, \rho, \theta). \quad (3.8)$$

In their discussion, Li and Stephens [2003] argue that the PL model tended to result in poor estimates of the parameters. Their explanation for such a result is that the PL model contains information about recombination that occurs in the most recent parts of the tree and not the tree as a whole [Li and Stephens, 2003].

From Equation 3.8 we observe that the PL model makes inference about the parameters by relating each haplotype to the patterns observed in the set of all other haplotypes. All haplotypes have some degree of relatedness to each other and in any sample there will exist, for each haplotype, a most closely related haplotype. As a result, information learned about the parameters will tend to come from information relating each haplotype to its most related haplotype which tends to occur in the more recent parts of the tree. Li and Stephens [2003] propose that if making inference about the parameters is unimportant in an application then the PL model may be more applicable.

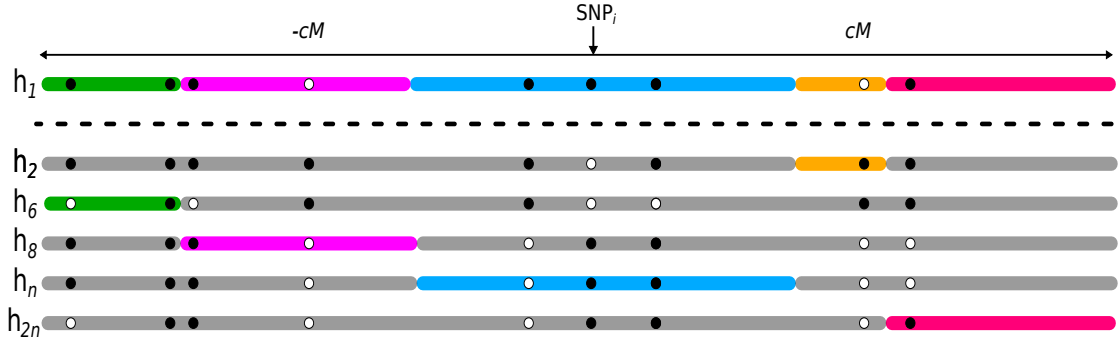


FIGURE 3.2: Illustration of chromosome painting adapted from Li and Stephens [2003] and Lawson et al. [2012]. The top haplotype is ‘painted’ in different colours representing the five other haplotype. Positions are denoted at circles, with white or black filling depending on the allelic type. We assume that these are phased chromosomes. At the first two positions, the h_1 haplotype ‘copies’ from the h_6 haplotype, this suggests that, in this realisation, over this section of the region, these two haplotype are more related to each other than the rest of the sample, i.e. they share a more recent ancestor. At the next block the h_1 haplotype switches and ‘copies’ from the h_8 haplotype.

Lawson et al. [2012] use the PL model as a way to describe the how the pattern of most closely related haplotypes varies across a genome, i.e. ‘paint’ or ‘colour’ a haplotype based on it’s most closely related ancestor in the sample. The colour of each painted chunk in the genome indicates the most closely related haplotype for the length of that chunk [Lawson et al., 2012]. Figure 3.2 shows a schematic of the painting procedure for one haplotype.

The painting algorithm estimates that at position one, haplotype one, h_1 , is most closely related to haplotype six, h_6 . As one moves along the genome, the painting algorithm switches to haplotype eight, h_8 . Lawson et al. [2012] argue that a switch in a painting corresponds to some ancestral recombination event which changes the ancestry at that position. From previously published results we know that, due to recombination, the structure of the tree relating a sample of haplotypes will change along the genome [Griffiths, 1981, Hudson, 1983].

An important result is that under these models the number of recombination events, in expectation, is proportional to the time for two chromosomes to find a common ancestor. Therefore one can use the number of recombination events

between two chromosomes as a proxy for the height of their phylogeny. Using the procedure in Figure 3.2 one can estimate the number of recombination events between two haplotypes by counting the number of switches in the sampled painting. By repeating this across sampled paintings and haplotype pairs one is able to construct a dissimilarity matrix that estimates the number of recombination events between any two haplotypes in the sample. Lawson et al. [2012] term this the ‘co-ancestry’ matrix which they use in conjunction with a Bayesian classification method to cluster individuals into population groups.

3.3.1 Local Ancestry

We decided to establish how well this chromosome painting procedure captures changes in recent ancestry. We simulated a neutrally evolving 2Mb region with 200 haplotypes, and at each position, for every haplotype, we used the simulated phylogeny from *msms* to determine the time to coalescent for each haplotype pair. For a full description of the parameters used refer to Methods at the beginning of Chapter 2.

As we discussed above, *chromopainter* [Lawson et al., 2012] uses sampled paintings to estimate the amount of genetic information two haplotypes share across the genome. This is a useful genome wide summary, however it is not possible to use this expectation to locally paint haplotypes as the painting is unlikely to be a valid path through the state space of the HMM. Therefore to find a painting that is a summary of the distribution we can determine the probability of each painting and choose the one that maximises it across the whole region. An efficient way of doing this is to use the Viterbi algorithm [Viterbi, 1967].

We applied both of these painting techniques to the simulated dataset. We used the latest implementation of *chromopainter* to sample 100 paintings for every haplotype, and our own implementation of the Viterbi algorithm to determine

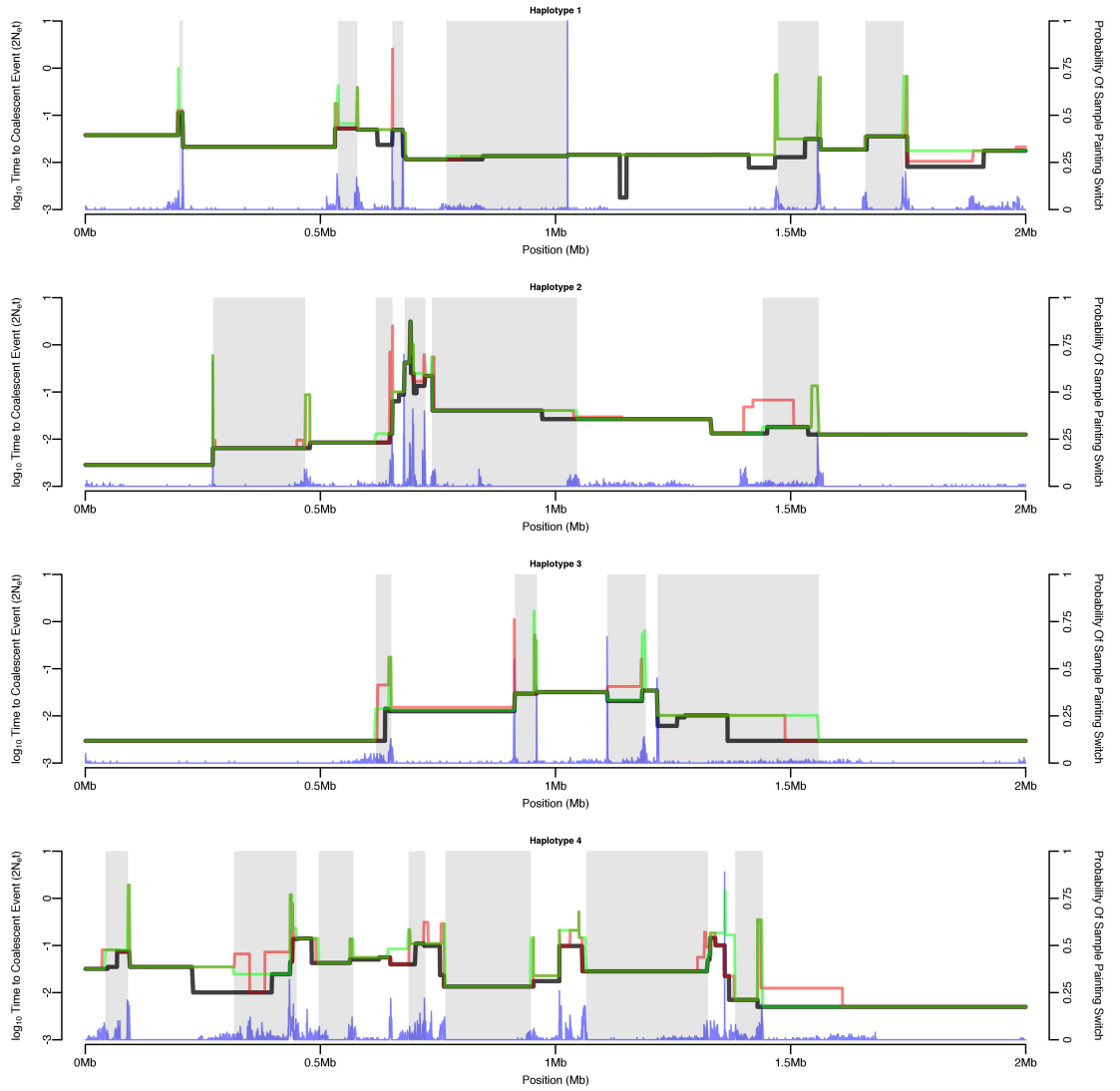


FIGURE 3.3: Each plot shows, for a given haplotype, the changes in the time to coalesce with its most closely related ancestor across this region. The solid black line shows how the true time to first coalescent event changes across the region. The green line shows the time to coalescence for the pair of haplotypes identified by the Viterbi algorithm. The red line shows the time to coalescence for the pair of haplotypes identified by the mode of the sampled paintings, taken at each position. The blue vertical lines shows the proportion of sampled paintings that switch at that position. The alternate shading identifies the boundaries of switches in the most likely painting.

the most likely painting. Figure 3.3 shows how these two painting procedures are related to changes in the time for a haplotype to find common ancestor across the simulated region.

We observe from Figure 3.3 that in our simulations, both the most likely painting

and sampled paintings appear to identify the majority of changes in the ancestry; on average sampled paths identify the closest ancestor, across all haplotypes, at 73.0% of SNPs, and at 77.4% of SNPs for the most likely painting. While the sampled painting method is more accurate at determining the number of recombination events, we observe that for 75.5% of haplotypes the most likely painting will correctly identify the closest ancestor at more SNPs than the sample paintings.

As one might expect, there also appears to be a trend for longer painted lengths to lie on sections of a haplotypes DNA that have shorter time to common ancestor. This follows from the discussion above; under the coalescent, the number of recombination events that can occur is governed by a Poisson process with rate proportional to the height of the phylogeny, therefore reduced time to common ancestor implies that, in expectation, fewer recombination events can occur [Hein et al., 2005]. As a result the ancestry will remain constant for much longer if two haplotypes have a more recent common ancestor.

We decided to explore this further by looking at the lengths of the most likely painted chunks related to the time for a haplotype to coalesce with its painted pair. Figure 3.4 shows the length of a particular painting described in terms of recombination distance, i.e. $X_u = 2N_e r_u x_u$ upstream, and $X_d = 2N_e r_d x_d$ downstream, where x_* the distance in number of positions and r_* is the average recombination rate over this region.

The total length of the chunk is expressed as $D = X_d + X_u$. Wiuf and Hein [1999] showed that given the height of the ancestry, t , at a position, the distance, in units of the recombination rate, until the first recombination event is governed by an exponential distribution, e.g.

$$\mathbb{P}(X_u > x|t) = \exp(-xt). \quad (3.9)$$

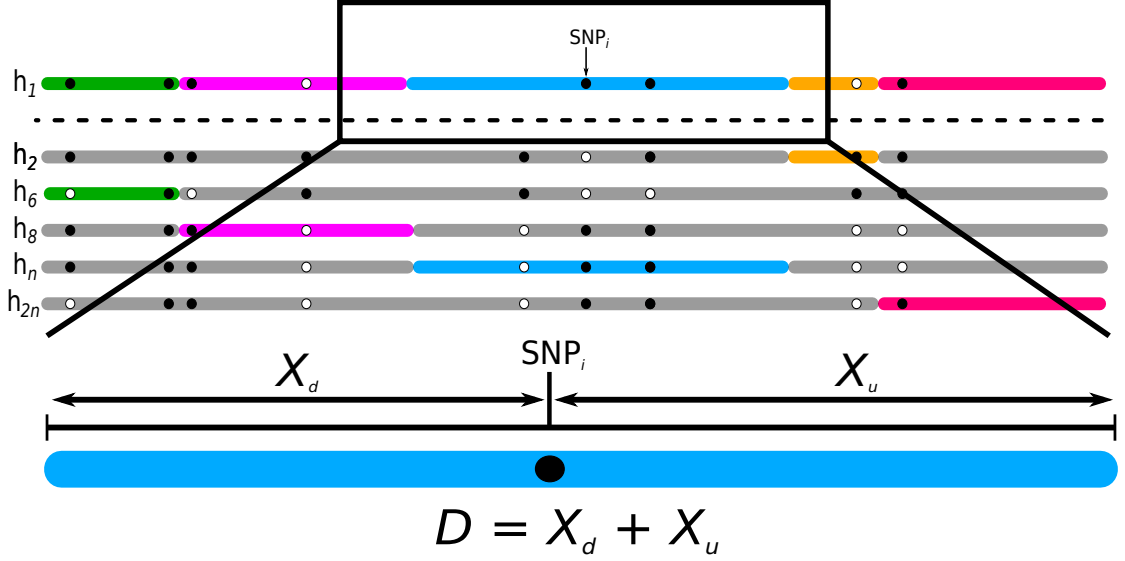


FIGURE 3.4: Illustration of how the lengths of a contiguous painting relates to the underlying recombination events. Using the same chromosome painting schematic from above, we focus on one contiguous painted segment at SNP i , of length D . From position i it is X_u recombination units long upstream, and X_d downstream.

Using this description, and assuming that the probability of a recombination event is independent of the direction from the position of interest, we can describe the distribution of D

$$\mathbb{P}(D \in (d, d + \delta d) | t) = td \exp(-td) \delta d \quad (3.10)$$

Therefore, in expectation we find that,

$$\mathbb{E}(D|t) = \frac{2}{t}. \quad (3.11)$$

We then produced a scatter plot of the empirical values of most likely painting and the corresponding time to common ancestor, and overlaid it with Equation 3.11, Figure 3.5.

As expected there is a strong relationship between the theoretical derivation of Equation 3.11 and the empirical observations. We observe that the fit appears to be closer to the empirical observations at more recent ancestries. Above a

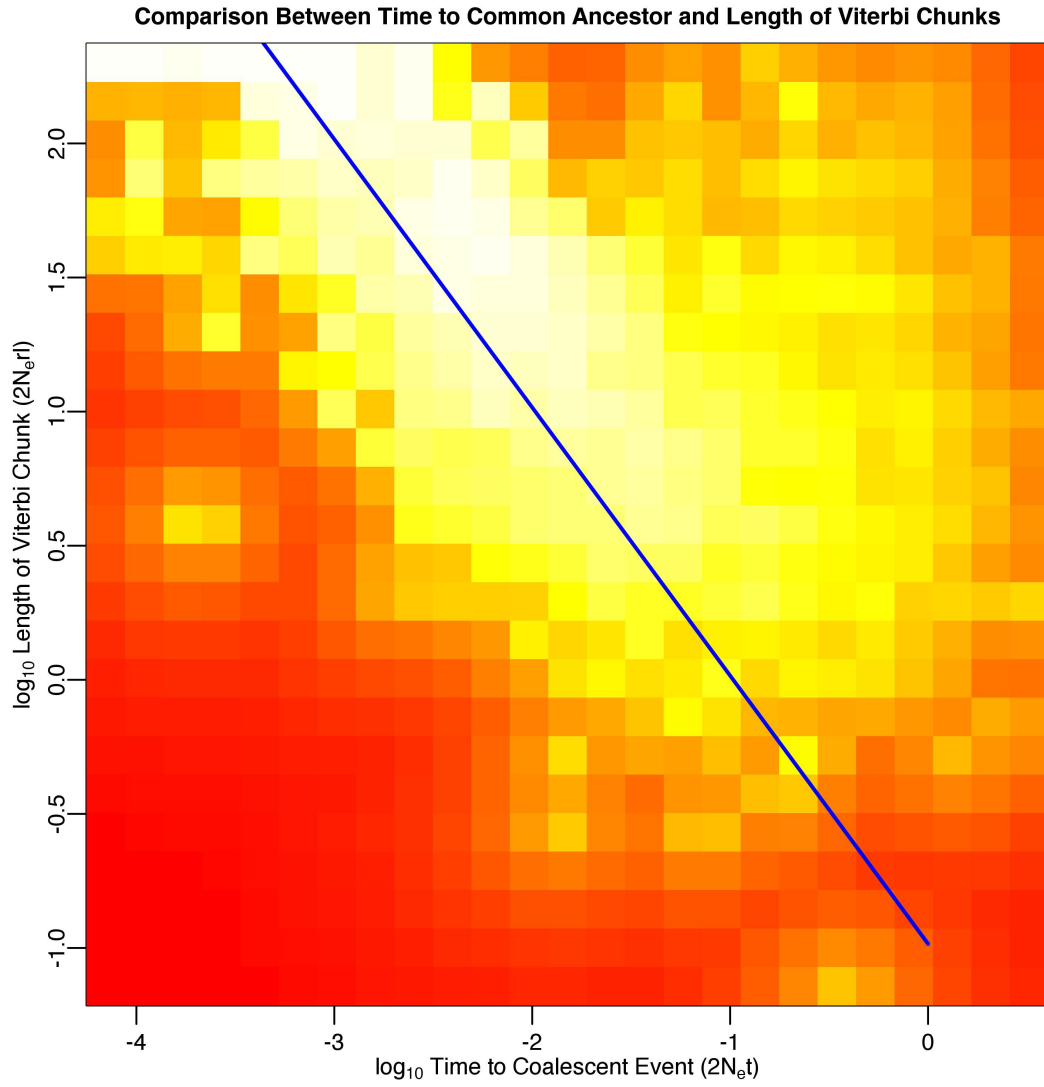


FIGURE 3.5: Heatmap showing a two dimensional kernel density estimate of a scatterplot between the time to common ancestor for a painted pair and the corresponding length of the painted segment. Bright yellow indicates high density, and dark red indicates low density. The blue line shows the theoretical relationship $\mathbb{E}(D|t)$ between the two quantities.

$\log_{10}$ coalescent time of -2 we observe that there is a bias towards longer painted lengths. This might be as a result of the challenge of using the PL model to make inference. As noted by Li and Stephens [2003] the PL model performs poorly when trying to estimate parameters as it uses information contained in the tips of the tree.

The results presented in Figure 3.5 appear to suggest that we can use the information contained in the distance between breaks in the painting as a proxy

for distance between recombination breaks. This combined with Equation 3.11 implies that the lengths between breaks in the painting can be interpreted as a function of the time to common ancestor.

3.3.2 Natural Selection

From theoretical descriptions of coalescent models under selection we know that the effect of a positive selective sweep is to reduce the time to common ancestor among those haplotypes carrying the beneficial allele [Coop and Griffiths, 2004, Prezeworski et al., 2005]. Therefore, for those haplotypes carrying the selected allele, we would expect to observe longer distances between recombination events at the selected loci. From the results presented above we would therefore expect to see, at the selected loci, longer painted lengths for those haplotypes carrying the selected allele, and shorter lengths for those haplotypes carrying the alternative allele.

Using the same approach as in previous sections we used the simulation tool *msms* [Ewing and Hermisson, 2010] to simulate genetic variation from trees that experience both classical sweeps and soft sweeps at different strengths of selection σ . For a full description of the parameters used refer to Methods at the beginning of Chapter 2.

As depicted in Figure 3.6, we calculated the lengths of most likely painting for those segments that cross the selected position. Using the selected allele we bi-partitioned the haplotypes and calculated the ratio of the median lengths of the most likely painted segments under the derived and ancestral allele. We would expect to see significantly longer lengths on the derived background relative to the ancestral background, i.e. in the distribution of the log ratios we should see a skew towards positive values. The distributions of ratios can be observed in Figure 3.7.

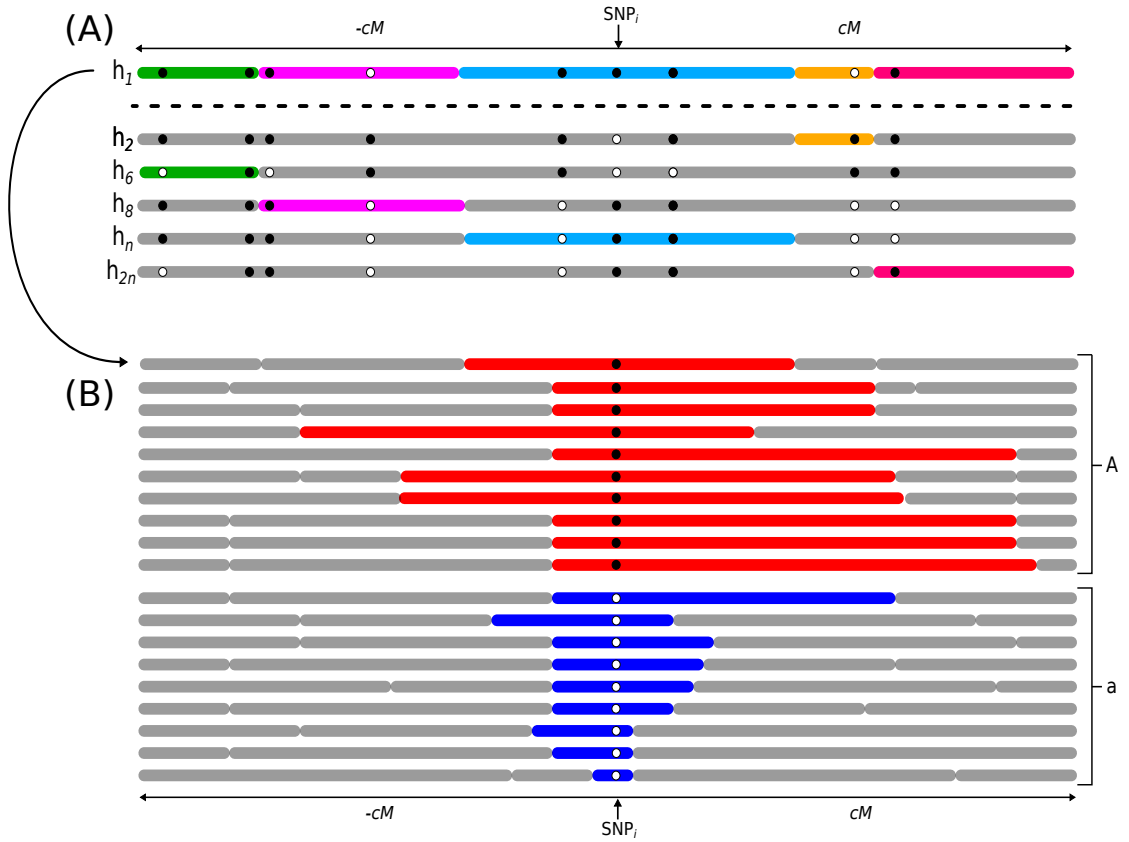


FIGURE 3.6: Illustration of how chromosome painting can be used to detect positive selection. **(A)** A haplotype is painted in according to who it is most closely related to at each position in the genome. **(B)** Using the lengths of the painting for each haplotype in the sample at SNP_i we bipartition them according to the allele that that haplotype carries at that position. To assess the evidence for selection we calculate the degree of association between long lengths and carrying a particular allele. Positions are denoted by circles with variants identified by colouring them black and white according to the allele that they carry.

As expected we observe a significant skew towards positive values on the derived allelic background relative to the ancestral background as the strength of selection increases. Reflecting the fact that individuals carrying the selected allele, in expectation, will coalesce much more quickly relative to those evolving neutrally. As a result they will tend to have fewer recombination breakpoints and longer painted segments. It is interesting to observe that the change across σ in larger frequencies appears to be slightly more pronounced than those observed at lower frequencies. This is because at the ends of the frequency spectrum there will be fewer observations in one of the alleles therefore the variance will increase.

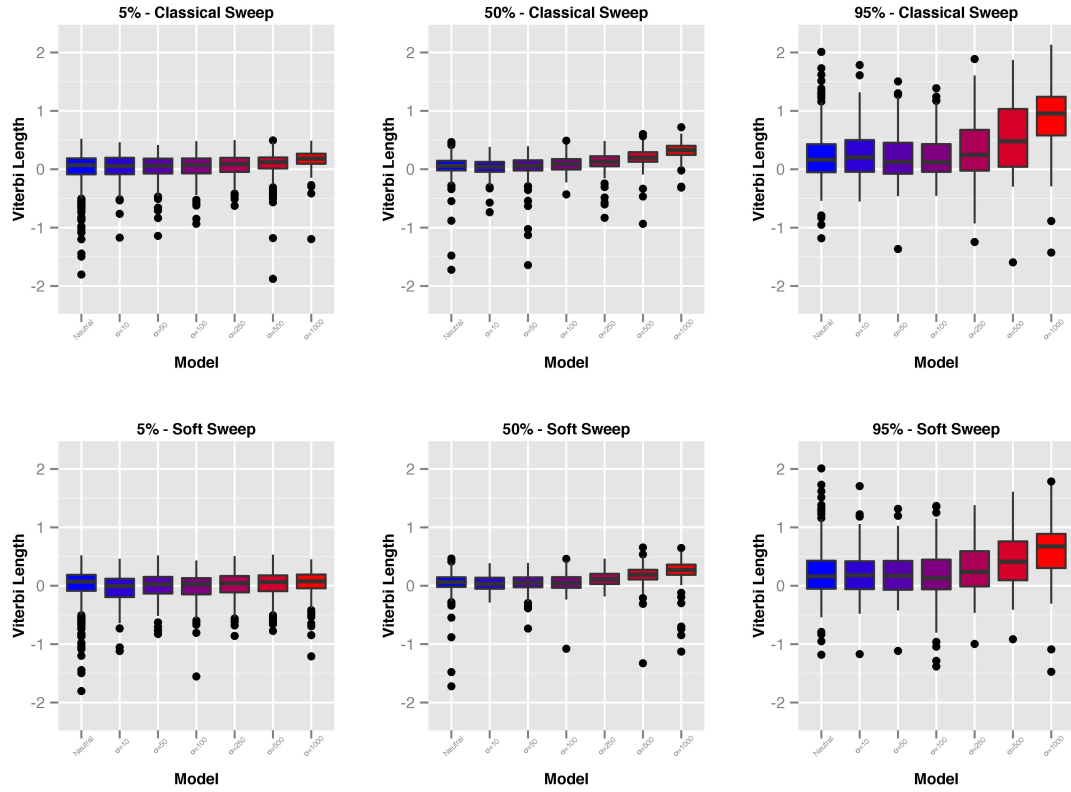


FIGURE 3.7: Each plot shows, for a given population frequency, the changes in the distribution of painted chunk length between the derived and ancestral alleles for different values of $\sigma = 2N_e s$. The top row shows the effect of classical selective sweep, and the bottom row shows the effect of a soft sweep. For $n = 200$ haplotypes, and 250 repetitions.

We also observe, in Figure 3.7, that the distribution of ratios for neutral evolution are slightly biased towards positive values. One explanation for this is that the time in the ancestral tree under the derived allele will be shorter than the time under the ancestral allele. Following the logic from above the lengths for those individuals carrying any derived allele will likely be marginally longer than those who carry the ancestral allele. Indeed this will likely change depending on the frequency of the allele, as younger alleles will tend to have lower frequency under neutral genetic drift.

The results above appear to suggest that it might be possible to use these painting techniques to detect regions in which a subset of the individuals have shorter coalescent time relative to the average coalescent time. Such patterns are indicative

of positive selection however we have to be careful as such features could be observed by population structure or alternative models of demography [Sabeti et al., 2002b].

In summary, we have shown in this section that;

1. Chromosome painting appears to be able to reconstruct the most recent parts of the underlying ancestral tree.
2. Empirical and theoretical results suggest that lengths between switches in the painting appear to be linked to the time to common ancestor between that pair.
3. Comparisons between the lengths of painted segments at a selected site, in simulated data, appear to be sensitive to both classical sweeps, and soft sweeps.

3.4 A Bayesian Test of Selection

Using the distribution of lengths of painted segments between bi-partitions of the haplotypes based on allelic type is similar to the approach that methods like iHS use to detect selection [Voight et al., 2006]. This similarity extends to the fact that long chunk lengths in one allele suggested increased linkage disequilibrium, or extended haplotype homozygosity.

The ratio of median lengths between the alleles might be sensitive to the presence of selective sweeps, however it will not be able to take into account the variation in lengths on either background, nor will it be robust to heterogeneous relatedness among samples. Therefore, rather than looking directly at the differences in

lengths we construct a Bayesian parametric model to test, at every locus, for association between chromosomes carrying a particular allele and having long painted segment.

3.4.1 Long Lengths

The length between changes in a painted chromosome appears to be a good approximation to the changes in the underlying recent ancestry, however as we observed above it not an exact relationship. Therefore we require that the test of association be sufficiently calibrated to reduce false positives whilst remaining sensitive to signals of a selective sweep.

There are a number of factors that might increase the false positive rate:

- As observed above, under neutrality the lengths of the painted chunks will be longer under the derived allele compared with the ancestral allele, as the derived allele will be younger.
- Variable recombination rates, will mean that switches in a painting will be more likely in certain parts of the genome than others.
- Alternative demographic histories, or cryptic relatedness; e.g. exponential expansion and bottlenecks will also change the distribution of lengths. For example an exponential expansion will increase the rate of coalescent events. This will reduce the average time to common ancestor across the genome which will increase the average distance between switches in the painting.
- Population structure will increase painting switches, between haplotypes from different populations, and decreases those within.

The last point is what Lawson et al. [2012] rely on to perform inference on population structure, however it will change the distribution of lengths within haplotype

pairs, which is a challenge for using the lengths to detect selection. To help alleviate this effect, and alternative ancestries we propose calculating a posterior probability that any painted segment is long when compared to its average length across the genome. This will enable us to make direct comparisons between the lengths across haplotypes at a single position.

To alleviate confounding effects of variable recombination rates we use genetic maps to calculate the length of the painted chunks in units of genetic distance. This has the effect that long painted segments that cross a recombination hotspot will be up-weighted and a long chunk that occurs in a recombination cold-spot will be down-weighted due to the lack of recombination events. We would expect haplotypes to recombine more quickly in a recombination hotspot, therefore the fact that we observe painted lengths that cross this region might indicate that the ancestry for these haplotypes is much shorter. Note that we assume that genetic maps are accurate, and not biased by selection.

To construct the posterior probability that a painted length is long, we assume that the lengths are exponentially distributed with particular genome-wide average. The justification for this assumption relies on the fact that we know that in a discrete state, discrete time, Markov chain the holding times are geometrically distributed [Norris, 1998]. Since lengths are measured in terms of the genetic map then a continuous approximation to the geometric distribution, i.e. exponential distribution may be appropriate.

To check whether the exponential assumption holds in practice we constructed most likely paintings for every haplotype in a sample of 170 CEU haplotype using phased sequence data from the 1000 Genome Project [1000 Genomes Project Consortium et al., 2012]. For a full discussion about this data please refer to Data in Chapter 5.

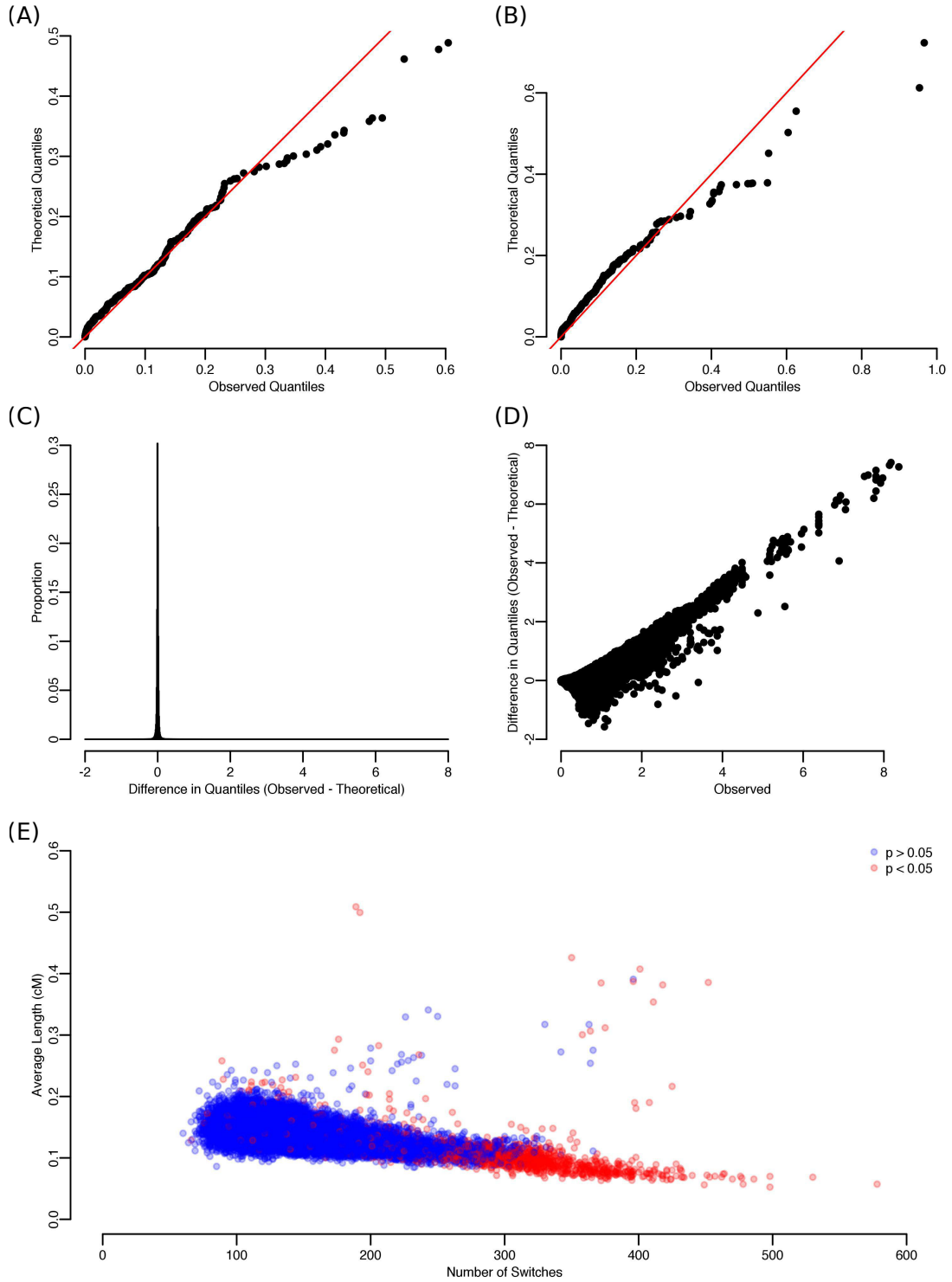


FIGURE 3.8: Results from analysis of painted lengths from a sample of 170 CEU haplotype using phased data from the 1000 Genome Project [1000 Genomes Project Consortium et al., 2012]. (A) and (B) show two quantile-quantile plots for two random pairs of haplotypes. The theoretical quantiles are determined by simulating lengths from exponential distribution with mean equal to the empirical mean of the chunks. (C) and (D) show the difference between the quantiles across all pairs and lengths. (E) plots the average length of a chunk to the number of switches within each pair of haplotypes. The points are coloured according to whether we would reject the null hypothesis of being exponentially distributed.

We ran our painting algorithm on all polymorphic loci in this sample and determined the distribution of chunk lengths for each pair of chromosomes. For every pair, using Kolmogorov-Smirnov (KS), we compared the empirical cumulative density function with the expected, and tested whether the chunk lengths are exponentially distributed with mean equal to the empirical mean of the chunks.

Figure 3.8 shows the results of the analysis. Any distribution that we try to fit to these lengths will be incorrect, however the results suggest that the exponential distribution appears to work in practice, i.e. for smaller lengths the quantiles appear to agree. We observe from the top two panels in Figure 3.8 that there appears to be a skew at the upper tail of the distribution, i.e. the true empirical distribution appears to be heavier tailed than the exponential distribution. Whilst not satisfactory, the fact that difference between the quantiles appears to be greater at larger values of painted lengths isn't a significant problem. The effect of this difference will be to inflate the probabilities for very long lengths.

Overall, these results suggest, with a 5% false positive rate, we do not reject the null hypothesis for 94.6% for pairs. For those pairs that do not pass this criteria, i.e. the red observations in Figure 3.8(E), we observe that they have on average 20% shorter chunks, and 1.9 times as many observations. Therefore it would appear pairs with short, but frequent, chunks do not behave exponentially. Since the chunks have low variation about their mean they will unlikely to be a problem as they will not contribute significantly to the test of association.

3.4.2 Description of Posterior Distribution

Under the exponential assumption it becomes possible to analytically calculate the posterior probability of a chunk being long. Define $v_i(k)$ to be the value of the most likely painting for the i^{th} chromosome at the k^{th} position, and suppose that

the length of the chunk that overlaps this position is $l_k^{(i)}$. Assuming that the distributional approximation holds, we can model $l_k^{(i)}$, as an exponential distribution with expectation α .

$$l_k^{(i)} \sim \text{Exp} \left(l_k^{(i)} | \alpha \right). \quad (3.12)$$

Define $\mu_{i,v_i(k)}$ to be the expected value of $l_k^{(i)}$ across the genome. Then, for every length, we wish to compare the following hypothesis;

$$H_0^{(i)} : \alpha = \mu_{i,v_i(k)} \quad (3.13)$$

$$H_1^{(i)} : \alpha < \mu_{i,v_i(k)}. \quad (3.14)$$

Define $p_{H_0} = \mathbb{P}(\alpha = \mu_{i,v_i(k)})$ to be the prior probability that a length is normal then under this exponential assumption we can calculate the corresponding posterior probability,

$$1 - \pi_i = \mathbb{P} \left(\alpha = \mu_{i,v_i(k)} | l_k^{(i)} \right) \propto p_{H_0} \mathbb{P} \left(l_k^{(i)} | \alpha = \mu_{i,v_i(k)} \right) \quad (3.15)$$

$$= \frac{p_{H_0}}{\mu_{i,v_i(k)}} \exp \left(\frac{-l_k^{(i)}}{\mu_{i,v_i(k)}} \right). \quad (3.16)$$

To calculate the probability that a length is long, i.e. $\frac{\mu_{i,v_i(k)}}{\alpha} < 1$, we introduce a dummy variable λ such that $\lambda\alpha = \mu_{i,v_i(k)}$. We set the prior of λ to be Uniform $(0, 1)$ so that it is diffuse and conjugate. This assumption stems from the fact that we wish $l_k^{(i)}$ to be the principle factor in determining whether it is long relative to its mean $\mu_{i,v_i(k)}$. If we define $p_{H_1} = \mathbb{P}(\alpha > \mu_{i,v_i(k)})$ then we can calculate the posterior

probability under H_1 as

$$\pi_i = \mathbb{P} \left(\alpha > \mu_{i,v_i(k)} | l_k^{(i)} \right) \propto p_{H_1} \mathbb{P} \left(l_k^{(i)} | \alpha > \mu_{i,v_i(k)} \right) \quad (3.17)$$

$$= p_{H_1} \int_0^1 \text{Exp} \left(l_k^{(i)} | \frac{\lambda}{\mu_{i,v_i(k)}} \right) \mathbb{P}(d\lambda) \quad (3.18)$$

$$\begin{aligned} &= p_{H_1} \int_0^1 \text{Exp} \left(l_k^{(i)} | \frac{\lambda}{\mu_{i,v_i(k)}} \right) d\lambda \\ &= p_{H_1} \frac{1}{\mu_{i,v_i(k)}} \int_0^1 \lambda \exp \left\{ -\lambda \frac{l_k^{(i)}}{\mu_{i,v_i(k)}} \right\} d\lambda \\ &= p_{H_1} \frac{\mu_{i,v_i(k)}}{(l_k^{(i)})^2} \int_0^1 \text{Gamma} \left(\lambda | 2, \frac{l_k^{(i)}}{\mu_{i,v_i(k)}} \right) d\lambda. \end{aligned} \quad (3.19)$$

Together these distributions allow one to determine the posterior probability that any length is long, by accounting for its expected length across the entire genome.

To see the effect of assuming this distribution in practice, we looked at the posterior probability as a function of the observed length $l_k^{(i)}$. By setting the prior probabilities p_{H_0} , and p_{H_1} to be $\frac{1}{2}$, we can write down π_i in terms of $\hat{\lambda}$.

$$\pi_i \left(\hat{\lambda} \right) = \frac{\hat{\lambda}^2 \int_0^1 \text{Gamma} \left(\tilde{\lambda} | 2, \frac{1}{\hat{\lambda}} \right) d\tilde{\lambda}}{\hat{\lambda}^2 \int_0^1 \text{Gamma} \left(\tilde{\lambda} | 2, \frac{1}{\hat{\lambda}} \right) d\tilde{\lambda} + \exp \left(-\frac{1}{\hat{\lambda}} \right)} \quad (3.20)$$

As you can see Equation 3.20 is not depend directly on the average length only on the observed inflation $\hat{\lambda}$. Figure 3.9 shows the values of Equation 3.20 across a range of inflation levels.

As expected, Figure 3.9 shows that as the inflation factor increases the posterior probability increases, and is invariant to the mean length. This is a crucial feature of the model which, as we will show in Chapter 4, makes the method more robust to alternative ancestries like population structure, and demographic histories.

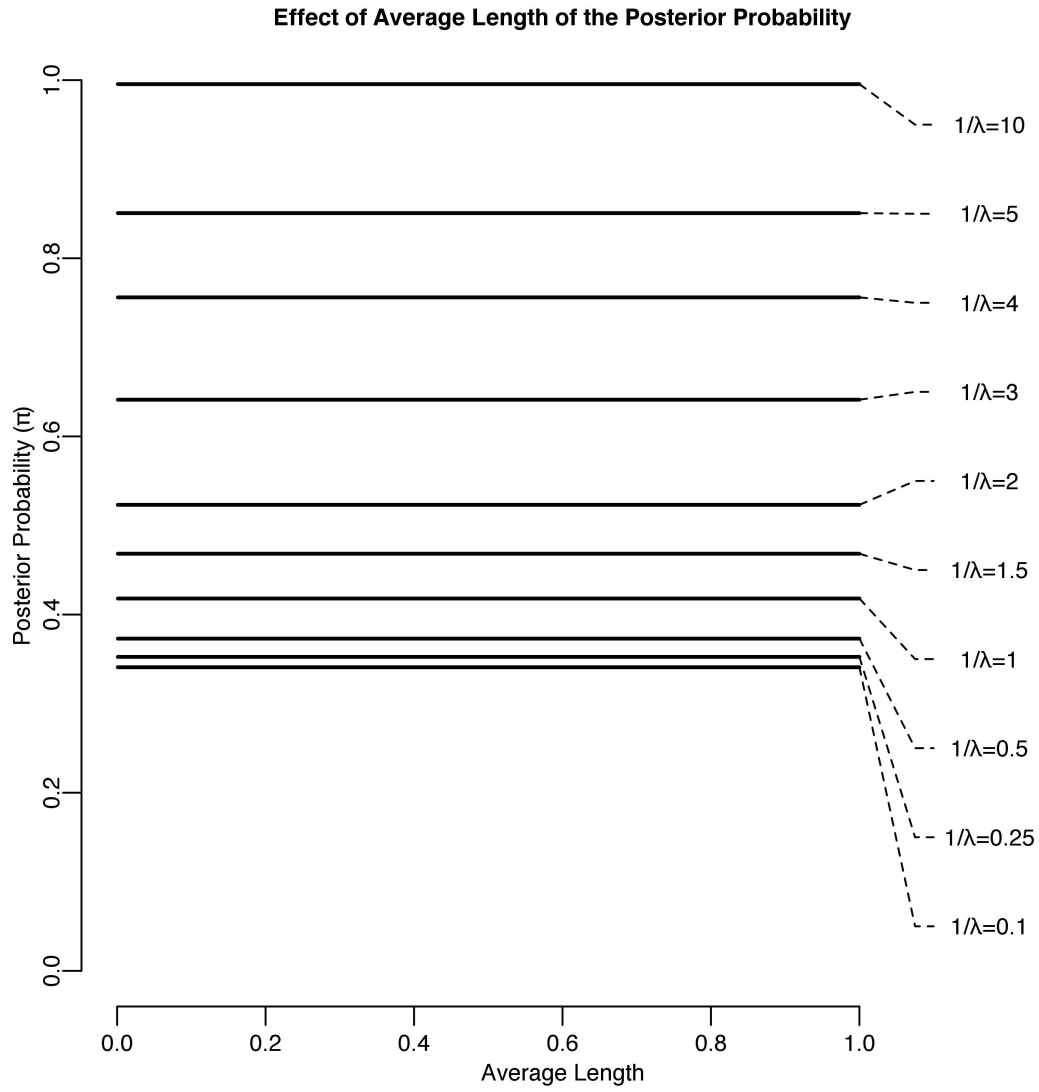


FIGURE 3.9: The posterior probability that a length is long as a function of the inflation factor λ , and average length $\mu_{i,v_i(k)}$. This function is calculated using Equation 3.20.

3.4.3 A Test of Association

Using posterior distributions from before we can establish whether there is association between long lengths in one allele and normal lengths in the other. One way to do this is to construct a logistic model where the outcomes are the allelic type, $g_i \in \{0, 1\}$ and the predictor is normal or long chunks, $\theta_i \in \{0, 1\}$ which we

observe with posterior probability $(\pi_i, 1 - \pi_i)$ respectively, i.e.

$$\text{logit}(g_i) = \mu + \beta\theta_i. \quad (3.21)$$

Therefore to perform this Bayesian association test, we compare the following hypothesis,

$$H_0 : \beta = 0 \quad (3.22)$$

$$H_1 : \beta \neq 0. \quad (3.23)$$

Assuming conditional independence between the observations the likelihood for this model can be expressed

$$L(\mu, \beta | \Theta, G) = \prod_{i=1}^n \mathbb{P}(g_i | \theta_i, \mu, \beta), \quad (3.24)$$

where $\mathbb{P}(g_i | \theta_i, \mu, \beta) = p(\theta_i)^{g_i} (1 - p(\theta_i))^{(1-g_i)}$, and

$$p(\theta_i) = \text{logit}^{-1}(\mu + \beta\theta_i). \quad (3.25)$$

Since we do not observe θ_i , we must marginalise over the uncertainty in θ_i , i.e. the posterior probability, π_i , that an individual is long which we calculated before. The likelihood can therefore be rewritten as

$$L(\mu, \beta | \Pi, G) = \prod_{i=1}^n [\mathbb{P}(g_i | \theta_i = 1, \mu, \beta) \pi_i + \mathbb{P}(g_i | \theta_i = 0, \mu, \beta) (1 - \pi_i)]. \quad (3.26)$$

To calculate the Bayes factor for the above test of association we must marginalise Equation 3.26 over μ and β using their distributions according the hypothesis, i.e.

$$BF_{H_1 \text{ v.s. } H_0} = \frac{\mathbb{P}(\Pi, G | H_1)}{\mathbb{P}(\Pi, G | H_0)}. \quad (3.27)$$

where, assuming independence between μ and β ,

$$\mathbb{P}(\Pi, G|H_i) = \int_{\beta} \int_{\mu} L(\mu, \beta|\Pi, G) \mathbb{P}(\mu|H_i) \mathbb{P}(\beta|H_i) \delta\mu \delta\beta. \quad (3.28)$$

We follow the argument in Marchini et al. [2007] and set the priors for μ and β to be Gaussian with mean 0 and variance 1. We chose these based on the fact that they contain limited prior information and allow the likelihood to dominate the posterior inference.

In Equation 3.21, β is the size of the effect of the association. In simple settings the model described by Equation 3.21 is analogous to a contingency table of counts [Casella and Berger, 1990]. Considering the contingency table, at low frequencies under one allele, there will reduced counts in one bin compared with the others. In these settings, therefore, the power to detect presence of patterns of variation associated with selection will be lower. We observe this effect in simulated data in Chapter 4, and it is an inherent feature of these types of association tests.

3.4.4 Laplace Approximation of Marginal Likelihood

Calculating Equation 3.28 analytically is non-trivial due to the form of the likelihood in Equation 3.26; it is difficult to find conjugate priors for the parameters. There are a number of potential Monte Carlo solutions to this problem, however they can be computationally expensive. If the likelihood is well behaved, an alternative approach is to use a Laplace Approximation.

The Laplace Approximation [Barber, 2012] is a method to approximate a marginal likelihood by performing a Taylor expansion, which can often be expressed as Gaussian kernel. This approach can be very efficient particularly when the likelihood is unimodal and log concave. To use a Laplace approximation it is necessary to

calculate the score and information matrix for the likelihood in Equation 3.28. These are detailed in explicitly Appendix A.

Using these together with their null model counterparts, we can estimate the maximum a-posteriori (MAP) estimate for $\gamma = (\mu, \beta)$ under both models; $\hat{\gamma}_{H_1}$, and $\hat{\gamma}_{H_0}$. This can be done a number of ways but we chose a Newton-Raphson based hill climbing approach [Barber, 2012].

Our results suggest that this method converges very rapidly (< 5 iterations) to a stable estimate of the parameter. Following the discussion in Marchini et al. [2007] we use the Laplace approximation to approximate the marginal likelihoods under the null and alternative

$$\log [\mathbb{P}(\Pi, G|H_0)] \approx l(\hat{\gamma}_{H_0}) + \log \mathbb{P}(\hat{\mu}_{H_0}|H_0) - \frac{1}{2} \log |I_{H_0}(\hat{\gamma}_{H_0})| \quad (3.29)$$

$$\log [\mathbb{P}(\Pi, G|H_1)] \approx l(\hat{\gamma}_{H_1}) + \log \mathbb{P}(\hat{\beta}_{H_1}|H_1) + \log \mathbb{P}(\hat{\mu}_{H_1}|H_1) - \frac{1}{2} \log |I_{H_1}(\hat{\gamma}_{H_1})| \quad (3.30)$$

Combining the posterior probabilities and the test of association allows one to accurately estimate a value of the Bayes factor of association at every position in the genome.

In conclusion, we have set out, in this section, a model which we believe to be sensitive to positive selective sweeps, and robust to alternative ancestries. Over the course of the next two chapters we shall demonstrate this model's utility in a variety of simulated and real situations. Finally, methods in genetics often have a simple acronym that describes the key feature of the model, e.g. extended haplotype homozygosity (EHH) [Sabeti et al., 2002b], or integrated haplotype score (iHS) [Voight et al., 2006]. The construction of this model has motivated us to refer to it as ABACUS - **A** **B**ayesian **C**hromosome painting approach to detecting regions **U**nder **S**election.

3.5 Discussion

We set out at the beginning of this chapter to describe the development of this model and discuss the more technical aspects that make it capable of detecting classical sweeps and soft sweeps. We began by describing the background to the model, e.g. Li and Stephens [2003] and how it has been used in a number of other applications in populations genomics. We then showed how one application of the method, namely chromosome painting can be used as a way of making inference on how the relationships between the haplotypes change across the genome by looking at the most closely related ancestors.

We proposed that the lengths of this painting procedure could be used to detect regions of the genome where a subset of the haplotypes are more related over a longer continuous segment of genetic sequence than one might expect given the rest of the genome. Moreover we showed that this approach was sensitive to both classical sweeps and soft sweeps. Using this information we proposed a novel and efficient Bayesian model which calculates the probability that one allele is associated with increased relatedness relative to an alternative allele.

At the beginning of this chapter we detailed three of the algorithms that we have explored in relation to HMMs. We argued that in some cases it may not be appropriate to use the Viterbi algorithm as a summary of the HMM especially when the distribution over paths through the hidden states is flat. To alleviate this challenge we explored a range of alternative inference techniques that would allow us to accurately detect variation associated with selection. Two of the models that we explored were also sensitive to classical sweeps and soft sweeps, however like ABC it proved to be a very computationally expensive method which would not scale to current datasets. In the final chapter of this thesis we detail these methods under the future work section.

The chromosome painting model is an approximation to the underlying recent ancestry, therefore it is necessary to construct a test of association that is sufficiently calibrated to reduce false positives whilst remaining sensitive to signals of a selective sweep. For this reason we decided to not look at the raw lengths but construct a posterior distribution that established the probability that a chunk length was long. We shall show in the next section that this makes the model more robust to alternative demographies, and population structure, whilst still being sensitive to selective sweeps. Another effect of this model is that it will give each haplotype equal weight in determining whether there is evidence of selection.

An alternative approach to using the empirical mean of lengths within a haplotype pair to construct the genome wide average would be to treat it as a hyper parameter with a prior centred on the empirical mean. Equally, rather than calculating the posterior probability that each haplotype has a long length one might consider calculating the posterior probability that all haplotypes carrying one allele have long lengths. These approaches might increase the power of the model to detect selection, particularly on rare variants.

The test of association is an essential part of the model as it allows us, together with the posterior distributions on lengths, to probabilistically establish the evidence that a region exhibits patterns of variation associated with a selective sweep. Unlike alternative methods, e.g. iHS [Voight et al., 2006] we make no distinction between the ancestral and derived alleles. Whilst this increases our applicability to new datasets it is also a disadvantage as derived alleles are by definition ‘younger’. As we observed, this has the effect that low frequency derived alleles have longer lengths relative to the ancestral allele. We show in the next chapter that this doesn’t appear to be a significant issue, however it is something to bare in mind for future work.

Chapter 4

Model Validation and Power to Detect Selective Sweeps

“Everything that can be counted does not necessarily count; everything that counts cannot necessarily be counted.”

Albert Einstein

In the previous chapter we introduced a novel Bayesian approach to detect regions of the genome where a subset of the haplotypes exhibit patterns of variation associated with positive selection. In this chapter we focus on applying ABACUS to simulated data and calculating its power to detect selective sweeps.

We begin by looking at the distribution of $\log_{10}\text{BF}$ values calculated when ABACUS is applied to neutrally evolving populations, and comparing it to the distribution of $\log_{10}\text{BF}$ when ABACUS is applied to samples from some models of structured populations, and alternate demographies. Using existing techniques; iHS [Voight et al., 2006] and H_{12} [Garud et al., 2013], we compare the power of all three approaches to detect both classical sweeps, and soft sweeps across a wide range of population allele frequencies, f , and strengths of selection, σ .

We conclude that, in the scenarios we consider, the distribution of ABACUS appear to be largely stable when it is applied to samples from alternate demographic histories, and structured populations. Moreover, when comparing its power to detect selective sweeps we show that it tends to perform as well as iHS and H_{12} in classical sweeps across a wide range of allele frequencies, and has considerably more power to detect soft sweeps.

4.1 Methods

In this section we describe the methods that were used to perform the simulation experiments. We begin by detailing the coalescent simulator and the parameters used to generate the data. We then carefully describe the procedure by which we calculate power for the three methods.

4.1.1 Data Generation

As with previous chapters we used the coalescent simulator, *msms* [Ewing and Hermisson, 2010] to efficiently simulate genetic variation datasets in parallel. Unless otherwise stated, each dataset in this chapter is simulated on a 2Mb region for 200 haplotypes in conjunction with the estimates of effective population size N_e , mutation rate θ , and recombination rate r from Methods in Chapter 2. These parameter values were chosen to reflect the average variation observed in humans. Note that we assume that the recombination rate is constant.

When looking at the effect of selection on the statistics we construct regions as before except now there is a single selected locus which we assume occurs at 1Mb (i.e. in the middle of the region). The software allows us to change the strength, current day frequency, and mode of selection.

As described in Chapter 1, we simulate soft sweeps by following Pennings and Hermisson [2006], Messer and Petrov [2013], and Garud et al. [2013] and allow recurrent mutations from the ancestral allele to the derived allele at the selected site. By allowing this type of mutation it helps us to understand the effect of soft-sweeps on the power of the statistics to detect selection using the same parameters that were used to simulate classical sweeps.

It is argued by Messer and Petrov [2013] that the patterns of variation observed by simulating soft sweeps in this way, i.e. recurrent mutations, are similar to those obtained under standing variation. Since it is difficult to establish whether this holds across all possible scenarios, we decided to simulate regions under standing variation in addition to the recurrent mutation simulations from above. The challenge with this, however, is that the simulation software *msms* [Ewing and Hermisson, 2010] is not capable of conditioning on the current day frequency. Therefore determining the power of the methods within frequencies is difficult. We decided to focus on $\sigma > 250$, and fix the frequency at the start of the selective pressure to be 5%. By varying the age of the selective pressure we are able to simulate regions so that the distribution of current day frequencies, at the selective allele, in 5% bins, is approximately even.

In addition to looking at the sensitivity of the model to detect selection relative to other methods we investigated the effect of alternative demographic histories on the distribution of the metric. As with Chapter 2 we focus on two types of variable population size models described in Figure 1.6; a population experiencing a bottleneck lasting 800 generations, with instantaneous current day recovery, and a population experiencing exponential expansion 1000 generations ago. For the bottleneck we looked at the effect of reducing the effective population size during the bottleneck: {50%, 25%, 10%, 5%}. For the exponential expansion we looked at the effect of changing the growth rate $\alpha = \{27.7, 55.5, 92.1, 184.2\}$, which is equivalent to the size of the population at the start of the expansion of {50%, 25%, 10%, 1%},

respectively. Separately, we also simulated samples which are taken from a structured, two population, model.

Finally, we decided to look at the effect of recombination hotspots on the distribution of the Bayes factor of association. *msms* [Ewing and Hermisson, 2010] is unable to simulate regions experiencing variable recombination. Therefore we used the program *msHot* [Hellenthal and Stephens, 2007] to simulate recombination hotspots of varying strength, λ : the ratio of the average recombination rate inside and outside the hotspot. Due to the computational complexity in simulating these regions we limited our attention to simulating a 200Kb region with a single recombination hotspot with $\lambda = \{1, 2, 5, 10\}$ spanning 10Kb in the centre of the region. We repeated the simulations 100 times.

4.1.2 Estimating power

Define X to be the statistic that we wish to calculate the power for. Then for a given pair of hypotheses, and false positive rate, α , power is formally defined as $p = \mathbb{P}(X > H_0^\alpha | H_1)$, where H_0^α is the α quantile of the null distribution, i.e. $\mathbb{P}(X > H_0^\alpha | H_0) = \alpha$. Note that we are assuming that larger values of X suggest departures from the null. Analogous versions can be constructed for two-sided and other one-sided tests. Figure 4.1 provides a helpful schematic showing the relationship between the null and alternative distributions, and power.

While power is a frequentist principle it is possible to use it in a Bayesian setting by treating the $\log_{10}\text{BF}$ as a statistic. Suppose want to test the hypothesis that a SNP is selectively neutral, or under positive selection. Parameterising the strength of selection as $\sigma = 2N_e s$, i.e. the population scaled selection rate, we test at each

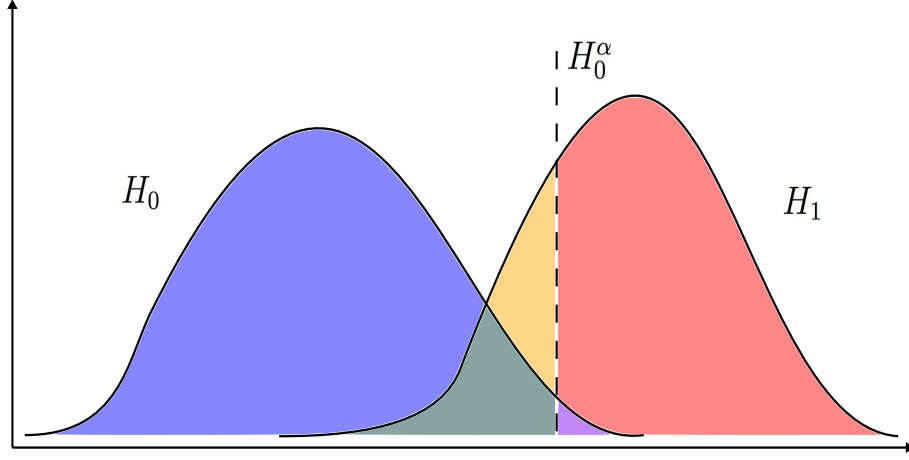


FIGURE 4.1: Schematic diagram showing how power is related to the difference between the distributions of H_0 and H_1 . The blue area represents the distribution of H_0 , and the yellow area represents the distribution of H_1 . The vertical dashed line represents the upper α quantile of H_0 , i.e. H_0^α . Power is then specified by the ‘red’ area, i.e. the area under the alternative distribution greater than the null critical value.

SNP

$$H_0 : \sigma = 0$$

$$H_1 : \sigma > 0.$$

If the distribution of the statistic, X , is known under both the null and alternative hypotheses then calculating power is theoretically straightforward. In the case of the models under investigation it is not possible to analytically determine the distribution of the statistics. Therefore we use Monte Carlo simulations to estimate the power.

To estimate power using a Monte Carlo approach, we obtain samples from the distribution of the given metric under the null hypothesis. By ordering these samples one can estimate H_0^α by taking the value of the top $\alpha\%$. Then for a fixed number of simulations under the alternative hypothesis one would determine the proportion of metrics that have a value greater than the estimated α critical value. Note in all the experiments in this thesis we chose $\alpha = 0.05$.

The distribution of some of the methods e.g. iHS and ABACUS are affected by the frequency of the derived allele. Therefore to avoid potential confounding factors we construct null and alternate distributions within frequency bins. One of the challenges with this approach, however is generating enough observations under the two models. To help alleviate this issue we decided to simulate the null by assuming that the selection coefficient of the derived allele was small. This allowed us to guarantee a fixed number of observations at specific frequencies under each null. As such we simulated null distributions setting $\sigma = 1 \times 10^{-3}$, which results in a selective advantage at each generation $s = 5 \times 10^{-9}$, which we assume to be negligible.

Note that we constructed separate null distributions under classical sweeps and soft sweeps. We could not do this for standing variation simulations for the reasons mentioned above, therefore we used the soft sweep null as a comparison. See Appendix B for examples of *msms* code for each simulation.

4.1.3 Calculating metrics in simulated data.

One of the challenges that we had in calculating ABACUS in simulated data was the sparsity of the region wide matrices that are used to determine the average length of a chunk for a pair of haplotypes. Since the regions over which we calculate the Viterbi paths are relatively short (~ 5000 segregating sites), the number of switches in painting is quite small therefore the resulting matrix can be quite sparse.

To resolve this we decided to determine the average length that two haplotypes share an ancestry across all pairs of haplotypes and simulations from the null. This value was then used as the ‘genome-wide’ mean length for all null, and selectively non-neutral evolutionary scenarios irrespective of chromosome labels. This is a

valid procedure as we know in simulated data the exact ancestry of the sample and in selectively neutral simulations the haplotype labels are interchangeable.

To calculate iHS we used the *R* package *rehh*, which in addition to calculating the unstandardised iHS score (uIHS), allows one to characterise the decay of EHH. We note that like ABACUS, iHS is based on bi-partitioning the alleles at each SNP and making relative comparisons between the two groups. However, unlike ABACUS, iHS makes a one way comparison between derived and ancestral alleles; an effect of standardising and taking the absolute value is that it will increase the clustering of the statistic at sites close to the selected allele [Voight et al., 2006]. We focus on the distribution of the statistic at a particular site, and construct it so that the derived allele is always the beneficial allele. Since a signature of a selective sweep is extended homozygosity on haplotypes carrying the beneficial allele we would expect to see large negative uIHS values. Therefore rather than calculating iHS we focus our attention, at least in simulated data, on comparing uIHS within frequencies.

To calculate H_{12} we follow the discussion by Garud et al. [2013] and centre a 401-SNP window over the selected site. Then using *R* we calculate the frequency of each haplotype, within this window, in descending order, i.e. p_i is the i^{th} most frequent haplotype. H_{12} is then calculated

$$H_{12} = (p_1 + p_2)^2 + \sum_{i>2} p_i^2. \quad (4.1)$$

4.2 Metric characteristics under neutrality

Before we look at the power of the model to detect selection it is important to determine the characteristics of each metric under selectively neutral evolution. The data for this section was generated as discussed in the Methods section. We

calculated the distribution of the three statistics at specific frequencies. Within each frequency bin we simulated 1,000 regions and calculated the three metrics, as discussed above. Figure 4.2 shows the resulting distributions as box-plots.

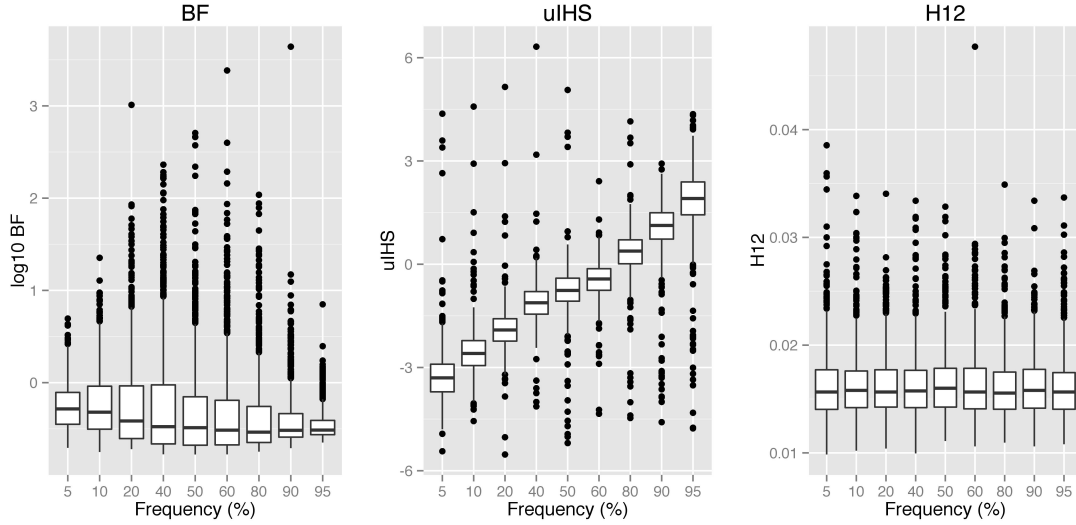


FIGURE 4.2: Box-plots showing the distribution of ABACUS, uHS , and H_{12} applied to populations experiencing neutral evolution, across a range of derived allele frequencies.

We observe in Figure 4.2 significantly different characteristics for each approach. The results from ABACUS appear to suggest a narrowing of $\log_{10}BF$ distribution at the edges of the frequency spectrum. Note also that there appears to be a noticeable trend from higher values of $\log_{10}BF$ when the derived allele has a low frequency moving towards lower $\log_{10}BF$ values at higher derived allele frequencies. The correlation between frequency and $\log_{10}BF$ is -0.03.

The narrowing of the distribution of $\log_{10}BF$ can be explained by the procedure with which the metric is calculated. Like iHS , ABACUS relies on bi-partitioning the alleles at each SNP. As a result, when the allele frequency is very low, or very high there will be fewer observations in one allelic group compared to the other. In this case there is likely to be a bias towards lower $\log_{10}BF$ as it is likely to require a very significant association for the $\log_{10}BF$ to increase beyond a critical threshold.

The downwards trend can be explained by the age of the derived allele. When looking at the differences between the raw lengths in Chapter 3 we noted that lower frequency derived alleles will tend to be younger, and therefore have longer lengths relative to the ancestral allele. This effect is reflected in the distribution of $\log_{10}BF$, we observe that low frequency derived alleles tend to have higher values than high frequency derived alleles. This feature is also present in other models that used homozygosity based metrics, e.g. *iHS*. Our analysis of *uHS* scores, Figure 4.2, suggests that this skew is significantly greater than that observed in ABACUS. Indeed this skew, as noted by Voight et al. [2006], means that it is necessary to perform standardisation within frequency bins.

In addition, whereas ABACUS has reduced variation at low minor allele frequencies (MAF), *uHS* appears have an increase. ABACUS has reduced power to detect associations because there are not enough observations in one of the comparison groups, however this affects *iHS* differently. EHH relies on the comparison of haplotypes within a group. When the number of haplotypes in a comparison group is small there will likely be fewer differences that occur between them [Voight et al., 2006]. This artificially inflates the decay of EHH causing the skew towards the allele with the lower frequency. In Figure 4.2 at 50% frequency *uHS* is a mean approximately equal to zero, with a slight bias towards negative values due to the derived and ancestral allele comparison. As a result, it is argued by the authors that *iHS* should not be constructed when the MAF falls below 5%.

In contrast to the results from ABACUS and *iHS*, the distribution of H_{12} does not appear to be affected by the frequency of the derived allele. One reason for this is that, where the other two models make inference using allele bi-partitioning, H_{12} looks at changes in frequencies of the two most likely haplotypes. Therefore in a neutrally evolving scenario it is unlikely to be affected by changes in the allele frequency. The fact that H_{12} appears to be invariant to the allele frequency is a distinct advantage over alternative methods, however like *iHS* it is not a statistic

that has direct interpretation in a probabilistic setting, i.e. it is not possible to observe a single statistic in any given scenario and assess support of one model over another.

Overall, the change in the distribution of $\log_{10}\text{BF}$ suggest that it may be necessary to perform some form of standardisation, like iHS. One problem, however with this approach is that it clouds the interpretation of the results, which is a significant advantage of a probabilistic model. The results that we present in this thesis suggest that the patterns we observe under neutral evolution have little effect on the quality of the results, however it is important to remember that there will be slight differences in the values of ABACUS depending on the frequency of the derived allele.

4.2.1 Effect of Demographic Changes

The assumption of a neutrally evolving population of constant size will not always hold in real data, and as we have seen in Chapter 2 an incorrect assumption regarding demographic history can lead to misleading conclusions. As a result we decided to look at the effect of different demographic histories on the distribution of $\log_{10}\text{BF}$ constructed by ABACUS.

To compare the distributions we looked at the effect on the false positive rate (FPR) as a function of the frequency of the derived allele. Using *msms* it is not possible to simulate alleles with fixed current day frequencies when the population size is changing over time. Therefore we focus on the distribution of $\log_{10}\text{BF}$ within 5% frequency bins. We estimated the FPR by finding the critical value of the distribution within a frequency bin under a population evolving under constant size and then counting the proportion of SNPs in that frequency bin, under a demographic scenario, that had a value greater than it. Figure 4.3 shows how the FPR of 5% varies as a function of frequency of the allele. In addition we used

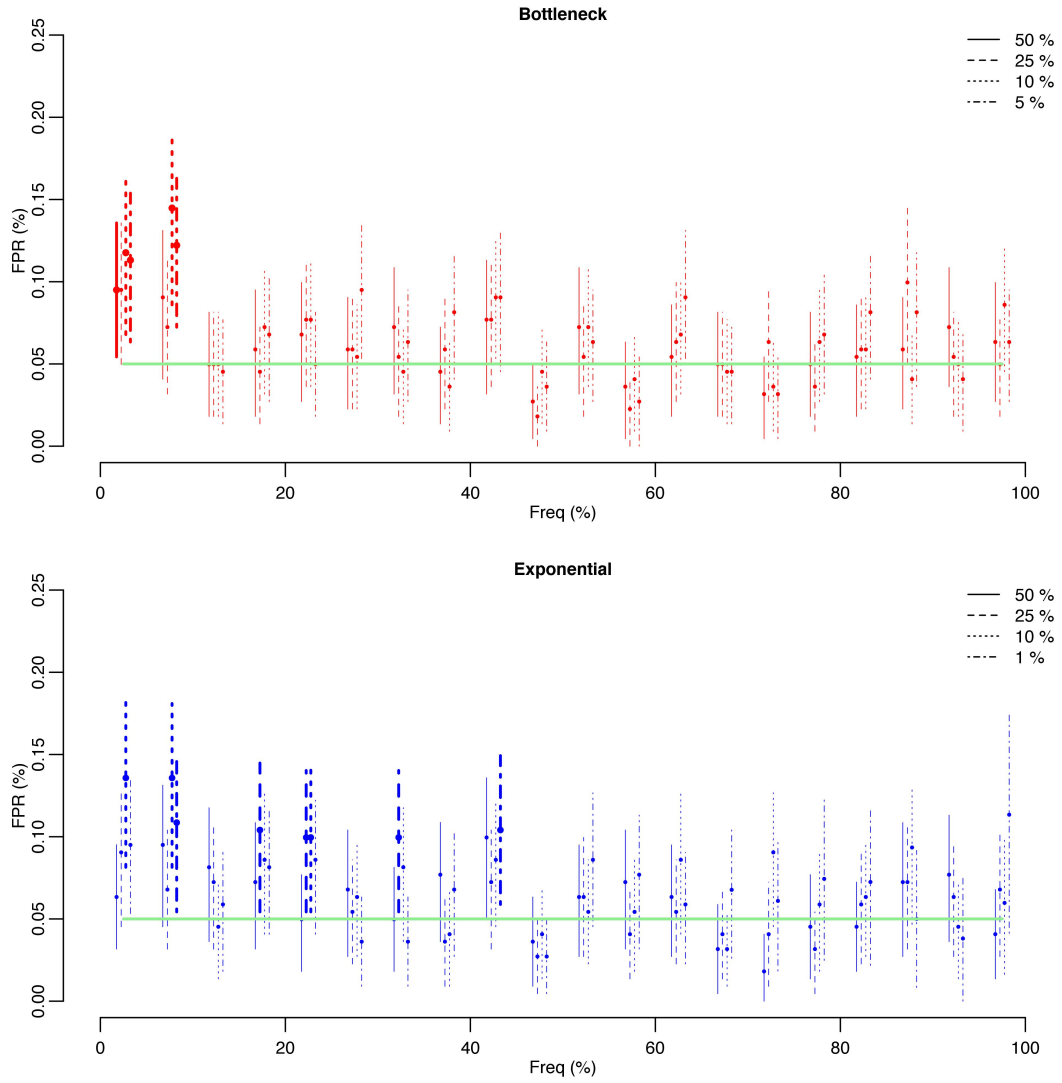


FIGURE 4.3: The effect of different demographies on the false positive rate compared to the $\log_{10} BF$ distribution under neutral evolution across a the frequency spectrum of the derived allele. The plot shows each estimated false positive rate, and the 95% bootstrap estimated confidence intervals using 1,000 resamples. Bold face identifies confidence intervals that do not included the 5% FPR.

Bootstrap resampling to estimate the 95% confidence intervals [Efron, 1979]. Note that the sample sizes used to estimate each confidence interval and FPR are the same.

In the demographic scenarios we observe that nearly all scenarios (94% in exponential growth, and 88% in bottlenecks), have estimated confidence intervals that

include or are less than the FPR of 5%, across all frequencies. There is some evidence to suggest that those frequencies whose confidence interval is greater than the FPR of 5% seem to be slightly biased towards lower frequencies. As we discussed above, this could be a consequence of the effect of younger mutations on the $\log_{10}\text{BF}$.

In these scenarios, we observe that there is little evidence to suggest that ABACUS is strongly affected by alternative demographic scenarios explored here. As we mentioned in the previous chapter, one reason for this is that ABACUS relies on a relative comparison between haplotypes, therefore alternate evolutionary scenarios that affect the whole genome should not increase the false positive rate.

4.2.2 Effect of Population Structure

Next we decided to explore the impact of multiple populations on the distribution of $\log_{10}\text{BF}$. As discussed in Methods we simulated genetic variation data on samples taken from a two population model. Using the painting model we calculated the total amount of shared painted segments within and across the two populations. Figure 4.4 shows the relatedness matrix across all 250 repetitions.

As we can see there is significantly greater sharing within populations, than across them. Note that this ‘co-ancestry’ matrix was calculated using the lengths of the matrix, rather than counting the number of lengths shared between haplotypes, as in Lawson et al. [2012]. To look at the effect of adjusting the lengths of a pair of haplotypes by their genome-wide average, we decided to perform two experiments. We calculated the $\log_{10}\text{BF}$ using the within and between population average matrix, shown in Figure 4.4, and compare the result to calculating $\log_{10}\text{BF}$ when we ignore the population structure and average the copying lengths across haplotype pairs. The distributions of the ABACUS results, under these two scenarios, is shown in Figure 4.5.

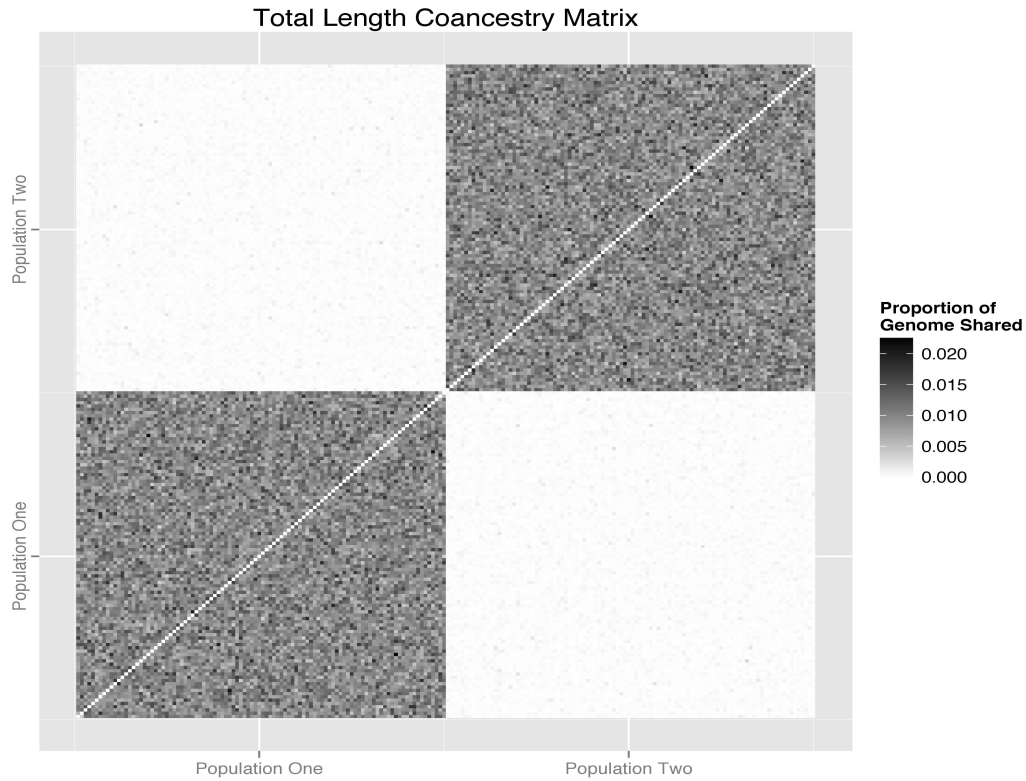


FIGURE 4.4: A row-normalised, ‘co-ancestry’ matrix calculated by pair-wise summing the total lengths that two haplotypes share across all repetitions.

Evidence from Figure 4.5 appears to suggest that, in this experiment, values of ABACUS, when we adjust for heterogeneous relatedness, appear to be unaffected by population structure. This is reinforced by the distribution of FPR in the top panel of Figure 4.5. Note that when we do not adjust for relatedness in the sample we observe a very different distribution of $\log_{10}\text{BF}$ values. We see an inflation of values at the ends of the frequency spectrum, and suppression of values at intermediate frequencies.

In conclusion, we observe, in the simulations that we consider, that ABACUS appears to have relatively stable false positive rates across the frequency spectrum, when there is population structure, demographic pressures, or neutral evolution. This is a key feature of the model, that is achieved by performing a relative comparison between haplotypes, based on their relatedness across a genome. There is evidence to suggest that at low frequencies we observe an increase in the FPR, this,

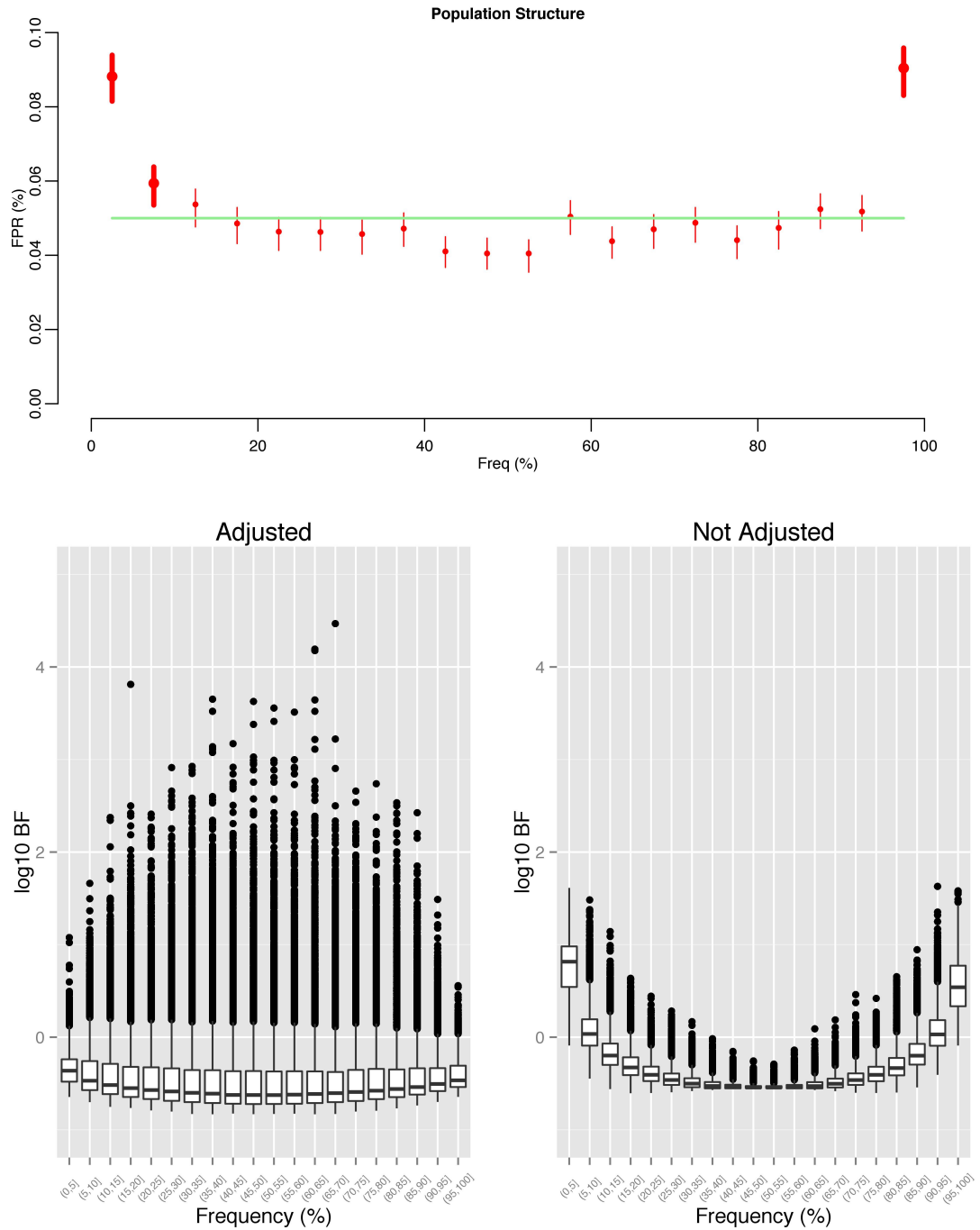


FIGURE 4.5: The top plot shows estimated false positive rate for ‘adjusted’ ABACUS values, and the 95% bootstrap estimated confidence intervals using 1,000 resamples. Bold face identified confidence intervals that do not include a 5% FPR. The bottom two plots show the distribution of $\log_{10} \text{BF}$ across the frequency spectrum for the derived allele. The left hand plots shows ABACUS results when we correct for heterogeneous relatedness, and the right hand plots shows the results when we assume that everyone has the same relatedness.

as we have discussed above, is due to the derived allele having more homogeneity, possibly due to its age.

4.2.3 Effect of Recombination Hot Spots

Finally we decided to explore the effect of recombination hotspots on the distribution of $\log_{10}BF$ values. As discussed in Methods we simulated genetic variation data on samples taken from regions with a single recombination hotspot with a recombination rate λ times the background recombination rate, where $\lambda = \{1, 2, 5, 10\}$. Figure 4.6 shows the false positive rates from SNPs with an allele frequency between 5% and 10%. The false positive rates were calculated as before only now comparing the distribution of $\log_{10}BF$ values inside to those outside the hotspot. Note that the results from other frequencies are similar, however because we cannot fix the frequencies of the alleles there is less data at higher frequencies.

From Figure 4.6 we observe that the false positive rates appear to be unaffected by recombination hotspots. Referring back to Equation 3.11 and Figure 3.5 one might argue that this observation is unsurprising. We observed that the lengths of the painting are highly correlated with the total genetic distance between recombination events. As a result in the presence of higher recombination rates the lengths will be reduced due to the increase in the probability of recombination. Assuming a neutral population this increase in recombination will affect all individuals. Therefore one might expect that by looking at a relative comparison between individuals there shouldn't be an increase in the false positive rates.

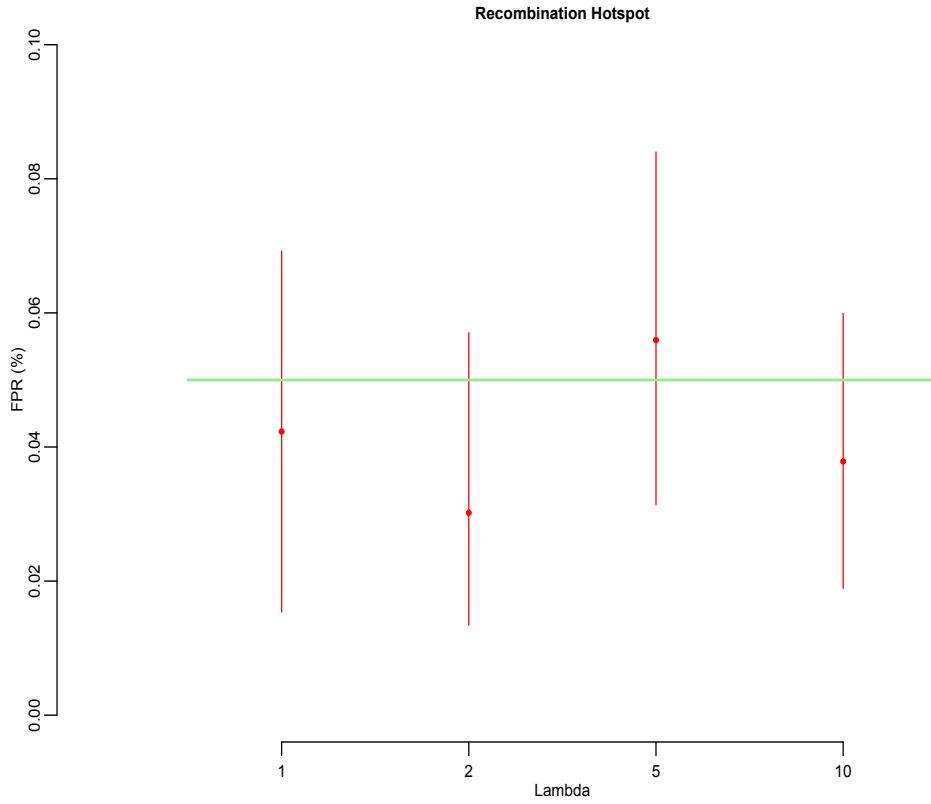


FIGURE 4.6: The effect of recombination hotspots of different strength on the false positive rate. The strength of the recombination hotspot is measured by λ which represents the recombination rate in the hotspot divided by the background recombination rate. The plot shows each estimated false positive rate, and the 95% bootstrap estimated confidence intervals using 1,000 resamples. The results were calculated by analysing SNPs with an allele frequency between 5% and 10%.

4.3 Power to detect ongoing selective sweeps

In this section we determine the power of ABACUS to detect signals of positive selection across a wide range of strengths of selection, and frequencies. In addition to classical selective sweeps we look at the power of ABACUS in soft sweep settings, specifically recurrent mutations, and standing variation. Note that unless otherwise stated we look at panmictic populations with a constant recombination landscape and constant population size.

Before we look at the results we show a comparative example that illustrates why ABACUS will have better power than alternative methods in soft sweep scenarios.

Figure 4.7 shows the results of the painting procedure in a region experiencing a classical selective sweep, and a region experiencing a soft sweep. In both scenarios we fix the strengths of selection, $\sigma = 500$, and the population frequency of the beneficial allele, $f = 50\%$. Note that these are the same regions that we looked at in Figure 1.10.

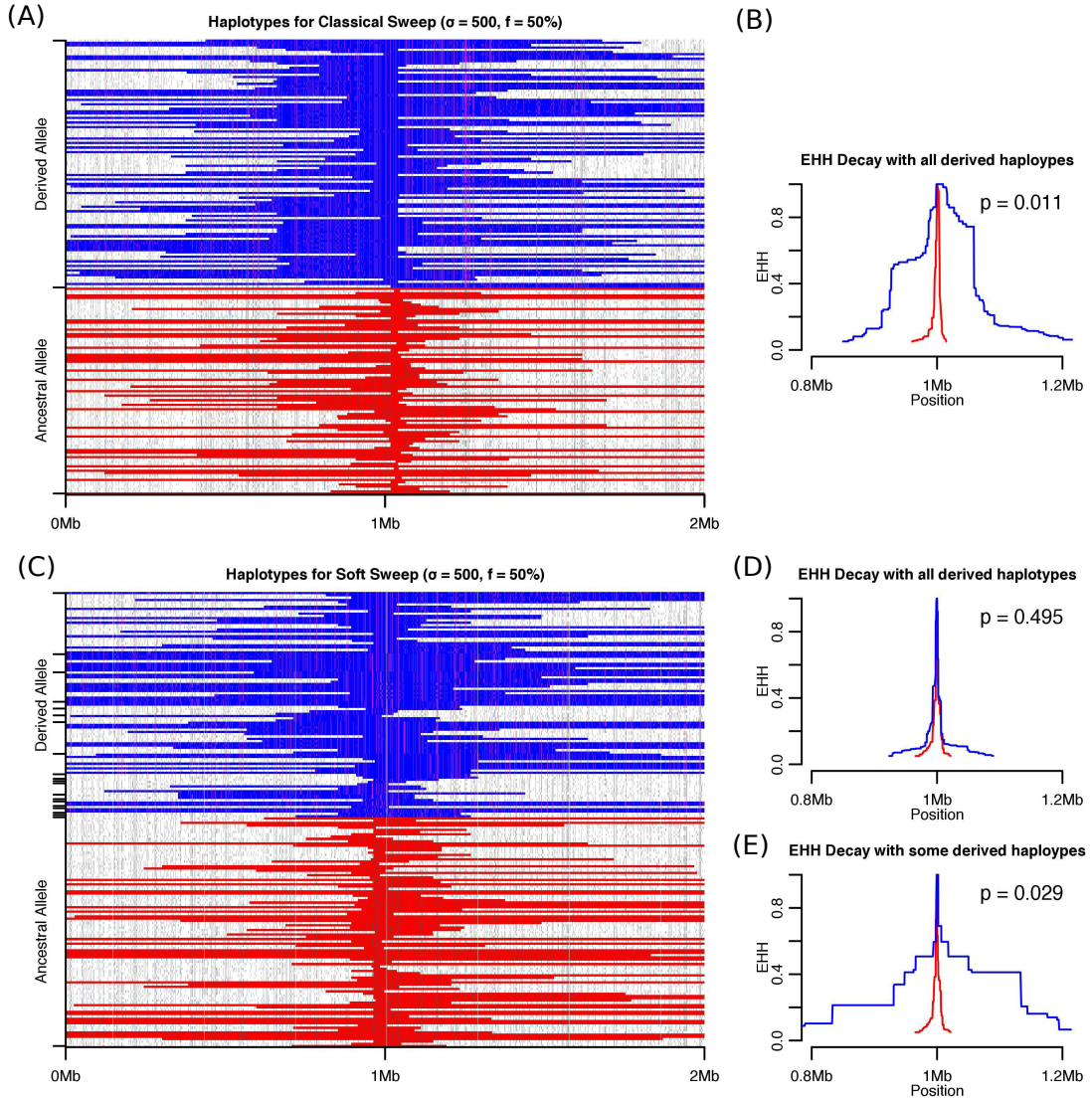


FIGURE 4.7: (A) A haplotype plot for a sample experiencing a soft sweep ($\sigma = 500, f = 50\%$), overlaid with ABACUS painted paths. (B) EHH decay at the selected site for the soft sweep using all derived haplotypes. (C) EHH decay at the selected site for the soft sweep using derived haplotypes obtained from the largest sample from a single mutation source. (D) A haplotype plot for a sample experiencing a classical sweep ($\sigma = 500, f = 50\%$), overlaid with ABACUS painted paths. (E) EHH decay at the selected site for the classical sweep using all derived haplotypes. Painted paths are coloured according to the allele that they carry at the selected SNP, blue for derived, red for ancestral.

As you can see in Figure 1.10 and Figure 4.7, under the classical sweep there is one haplotype on the derived mutation, and under the soft sweep there are multiple. The painted segments on the derived alleles, in both scenarios, are significantly longer than those on the ancestral background. We observe that the ABACUS value of association under the classical sweep was $\log_{10} BF = 6.62$, and under the soft sweep, $\log_{10} BF = 3.17$. Both of these have an empirical p-value $< 0.1\%$, under their corresponding nulls.

We calculated the unstandardised iHS score, and compared it to the corresponding null in both scenarios. We observe an empirical p-value of 0.011 in the classical sweep, and a p-value of 0.495 in the soft sweep. Therefore with a false positive rate of 5% we would correctly reject the null hypothesis in the classical sweep, and incorrectly accept the null in the soft sweep.

As we discussed above, the reason for this is that a soft sweep will likely create multiple haplotypes carrying the selected allele [Messer and Petrov, 2013], therefore haplotype homozygosity will decrease much faster and iHS will be suppressed. ABACUS will be robust to this because it relies on identifying regions of the genome where a subset of individuals have reduced time to common ancestor relative to the alternative group. By using the painting, rather than the raw haplotypes, ABACUS will make no distinction between a single long haplotype or multiple long haplotypes.

To illustrate this point we removed all, but the most common haplotype on the derived background in the soft sweep scenario, reducing the frequency of the allele to 20%. We then re-calculated the unstandardised iHS score and compared it to the appropriate null. We observe that the uHS value for this dataset was -2.943, which at 20% frequency has a p-value of 0.029. Therefore at a 5% false positive rate we would correctly reject the null hypothesis.

We shall see that this conclusion is not limited to this small experiment. Indeed we observe that both ABACUS and iHS have similar power to detect classical sweeps, but in the case of soft sweeps ABACUS has considerably greater power.

4.3.1 Classical sweeps

We begin by looking at the power of each method to detect classical selective sweeps. We simulated variation data using a range of derived allele frequencies and strengths, and the parameters that we discussed in Methods in Section 4.1 above. For each combination of parameters we simulated 250 regions. As described in Methods we constructed ABACUS, uHS, and H_{12} at the selected locus, and calculated, for $\alpha = 5\%$, the power maps shown in Figure 4.8.

From Figure 4.8 we observe that all three methods appear to have similar power (> 0.8) to detect strong selection ($\sigma \geq 500$) across intermediate frequencies ($10\% < f < 90\%$), and low power (< 0.2) to detect selection when the derived allele has low frequency ($f \leq 5\%$). We note that uHS has greater power than both ABACUS, and H_{12} to detect nearly complete sweeps ($f > 90\%$), and ABACUS has greater power than H_{12} to detect new sweeps ($f \leq 20\%$).

The corresponding receiver operator characteristic (ROC) curves are shown in Figure 4.9. Here we observe three frequencies $f = \{10\%, 50\%, 90\%\}$ and $\sigma = \{250, 500, 1000\}$. An optimal ROC curve will have an area under the curve equal to 1.

We observe that uHS has a greater area under its ROC curves for all σ at 10% frequency, however the differences between the methods seem to disappear as frequency of the derive allele and σ increases. ABACUS surpasses the H_{12} statistic nearly all but the ($\sigma = 250, f = 90\%$) ROC curve.

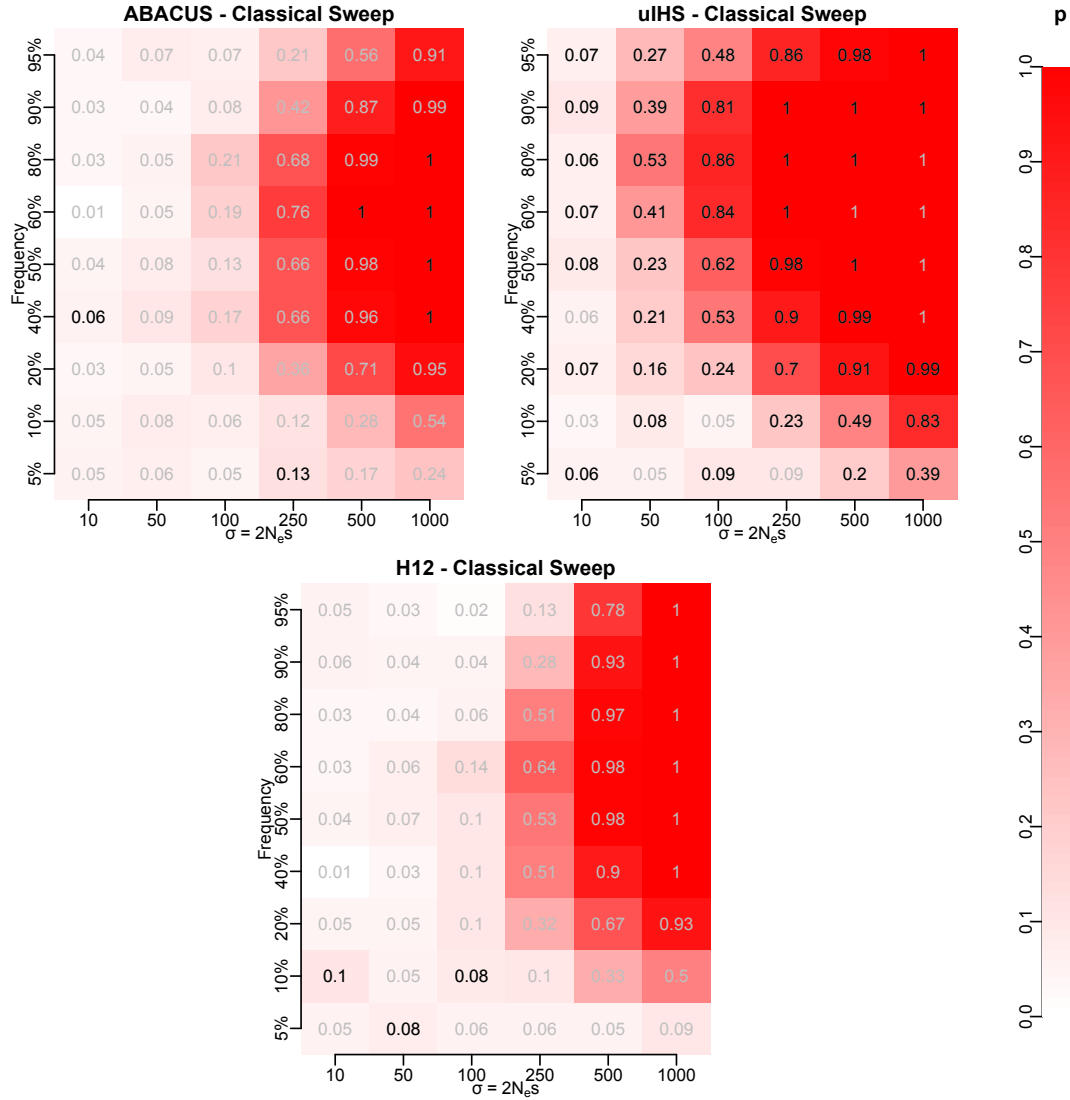


FIGURE 4.8: Three heat maps showing the power of uIHS, H_{12} , and ABACUS to detect a classical selective sweep, across $f = \{5\%, 10\%, 20\%, 40\%, 50\%, 60\%, 80\%, 90\%, 95\%\}$, and $\sigma = \{10, 50, 100, 250, 500, 1000\}$. Power is calculated as discussed in Methods in Section 4.1. The right hand panel provides the legend for all three maps.

In conclusion for classical sweeps we observe, in these simulations, that ABACUS appears to do as well as H_{12} in nearly every derived allele frequencies, and all strengths of selection that we considered. As we have mentioned before, the weak power of ABACUS at the edges of the frequency spectrum can be explained by the reduction in the number of observations in one allele group relative to the other. The upward bias of power towards higher frequencies in all methods is symptomatic of increased homozygosity on the derived background relative to the

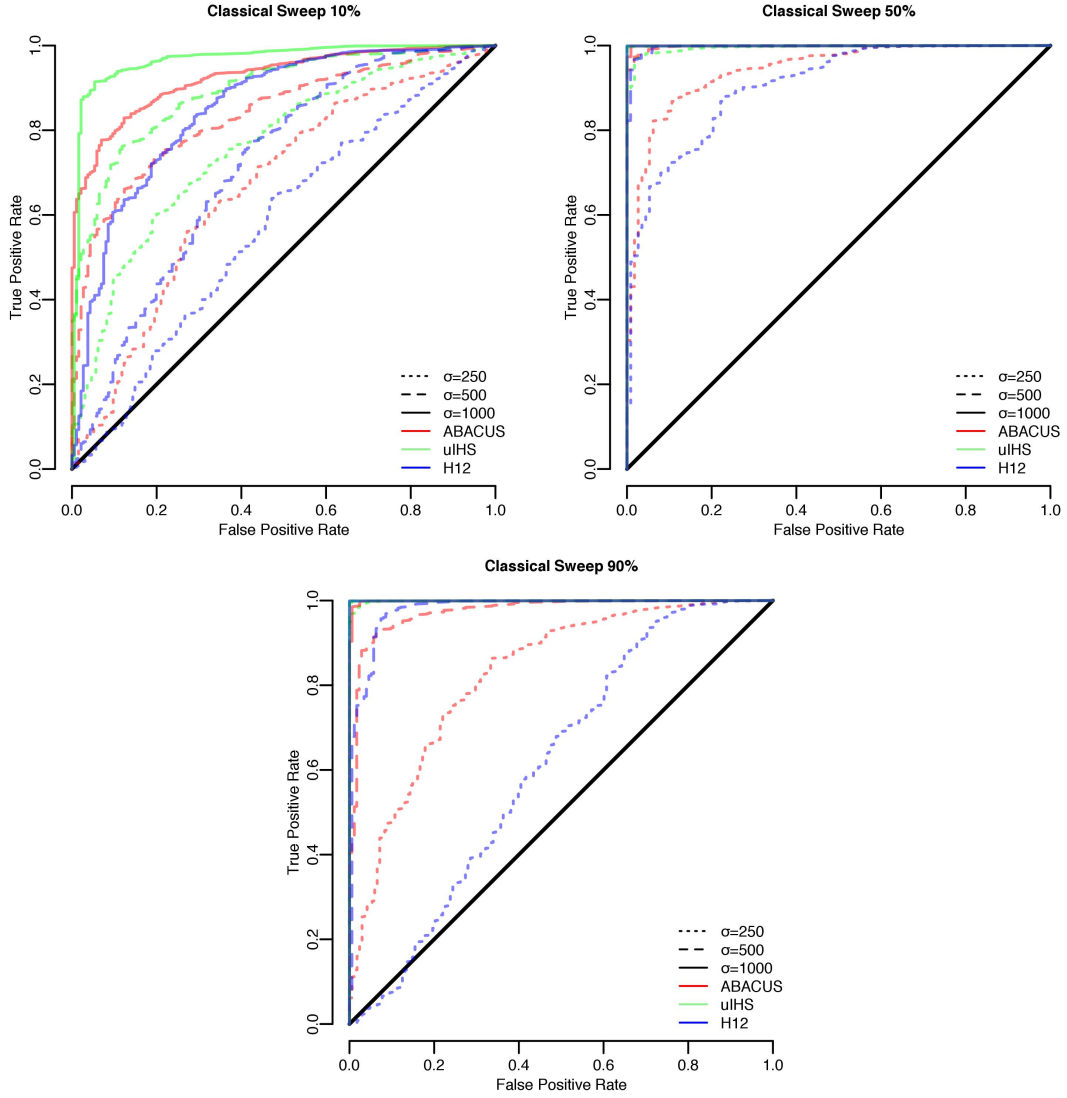


FIGURE 4.9: Receiver Operator Characteristic (ROC) curves comparing the relationship between the true negative and true positive rate for selection on a *de novo* mutation. Here we show the ROC curves for uIHS, H_{12} , and ABACUS for $f = \{10\%, 50\%, 90\%\}$ and $\sigma = \{250, 500, 1000\}$.

ancestral.

4.3.2 Soft Sweep - Recurrent Mutations

To investigate power to detect soft sweeps by recurrent mutations, we simulated regions with the same parameters as before, only now we relax the condition that the beneficial allele that arose from a single mutation. See the discussion in Methods in Section 4.1, and Figure 1.10 for a full discussion. At the selected

site we calculated ABACUS, uIHS, and H_{12} and using the discussion above we estimated, for $\alpha = 5\%$ the power maps shown in Figure 4.10.

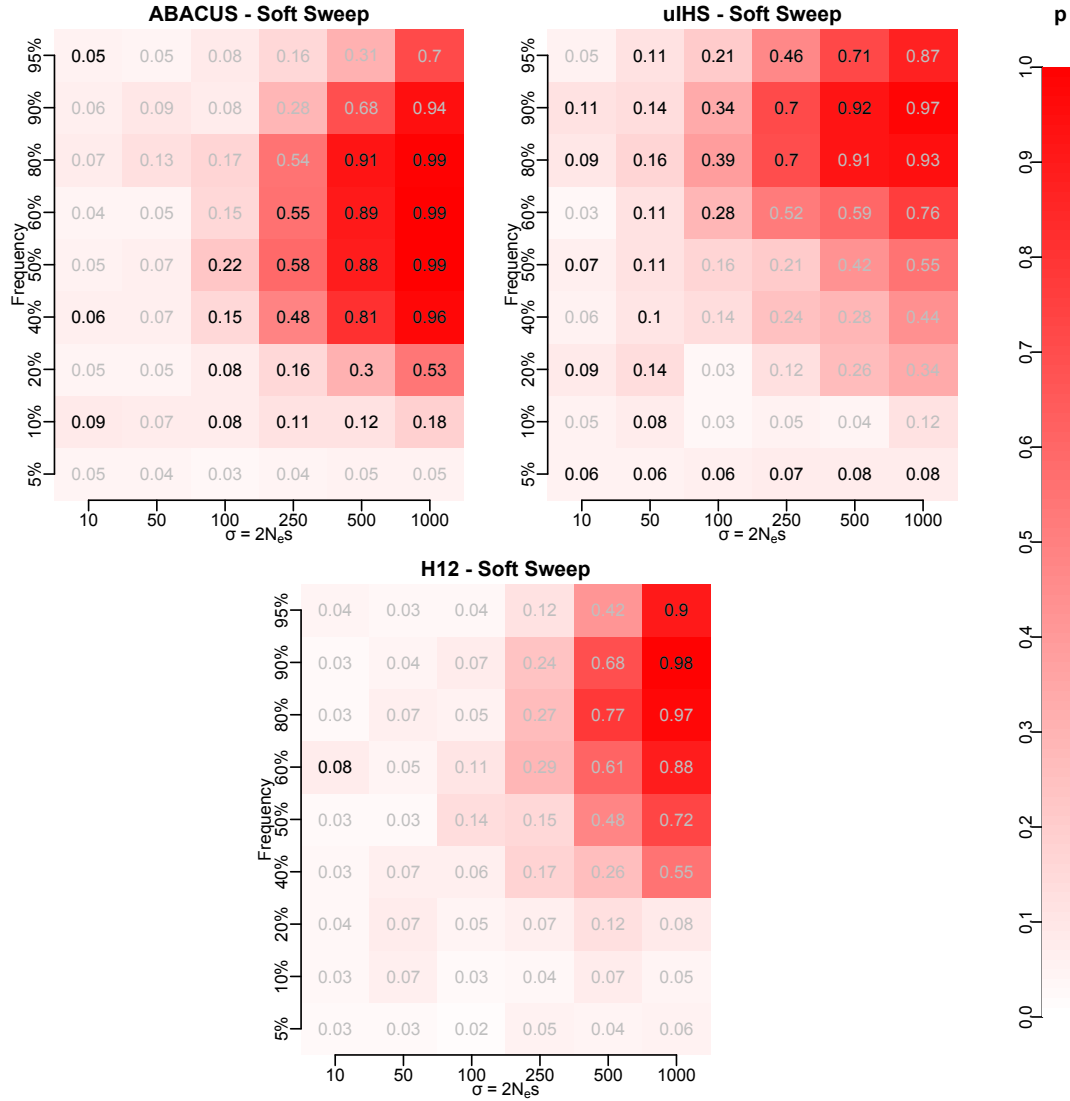


FIGURE 4.10: Three heat maps showing the power of uIHS, H_{12} , and ABACUS to detect a soft selective sweep, across $f = \{5\%, 10\%, 20\%, 40\%, 50\%, 60\%, 80\%, 90\%, 95\%\}$, and $\sigma = \{10, 50, 100, 250, 500, 1000\}$. Power is calculated as discussed in Methods in Section 4.1. The right hand panel provides the legend for all three maps.

From Figure 4.10 we observe that all three methods have similar power (> 0.8) to detect soft sweeps that have a current day population $f > 80\%$. Below $f = 80\%$, however we observe that power drops off significantly in uIHS. As we discussed above, EHH decay is downwardly biased in soft sweeps, because the multiple origins of the beneficial allele increases the chances that multiple ancestral haplotypes

appear on the derived background. We observe that ABACUS has better power than H_{12} to detect weaker selection ($250 < \sigma < 500$) at intermediate frequencies ($20\% < f < 80\%$) and all strengths of selection at lower frequencies ($f < 20\%$). Looking at the ROC curves in Figure 4.11 we see a similar pattern.

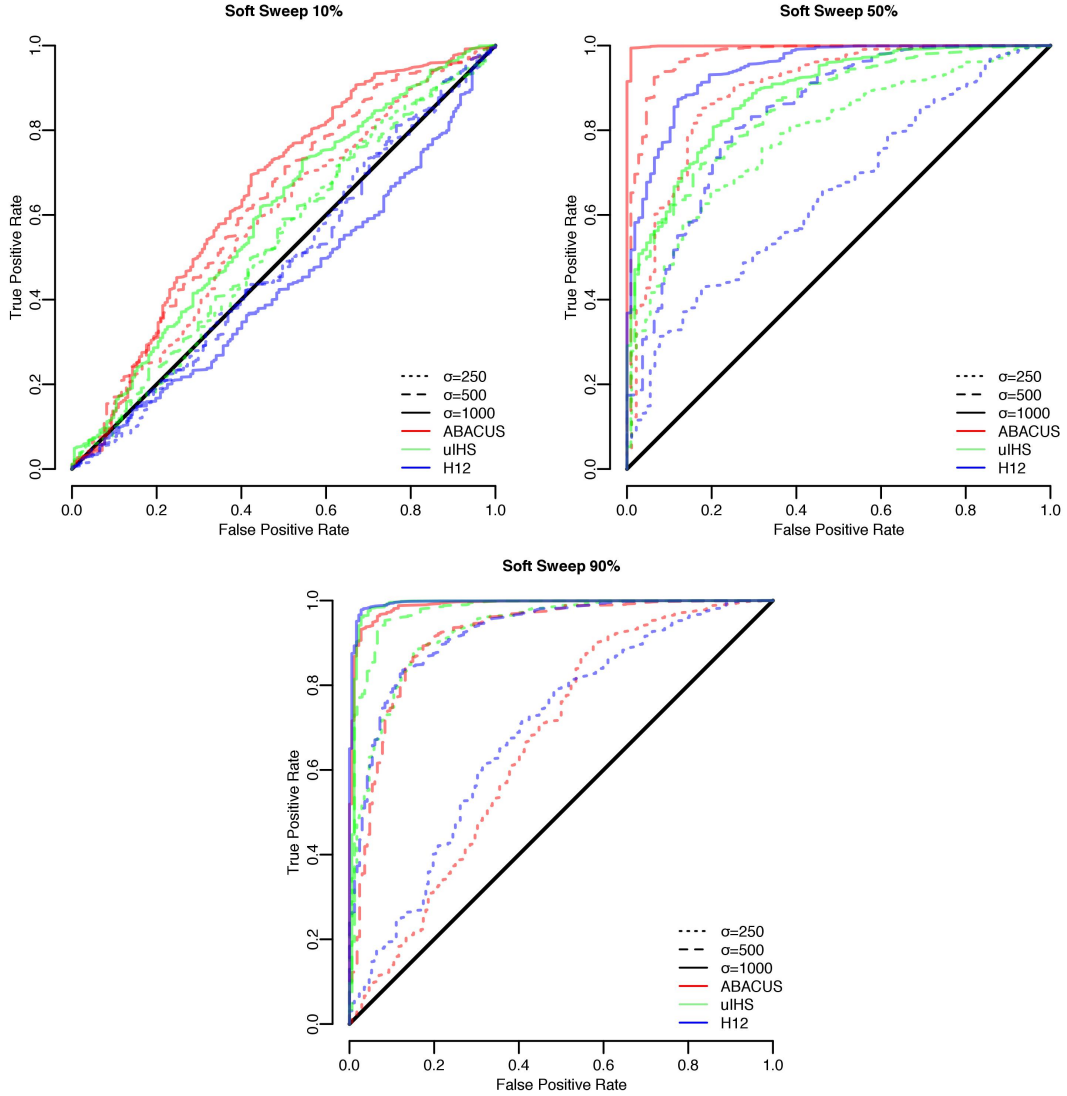


FIGURE 4.11: Receiver Operator Characteristic (ROC) curves comparing the relationship between the true negative and true positive rate for soft selective sweeps. Here we show the ROC curves for uIHS, H_{12} , and ABACUS for $f = \{10\%, 50\%, 90\%\}$ and $\sigma = \{250, 500, 1000\}$.

The ROC curves for ABACUS H_{12} , uIHS, at 10% frequency suggest that both methods struggle to detect any signal of selection, however because the curves lie,

approximately, along the diagonal suggests that the null and alternative distributions are nearly identical. At 50% we observe ROC curves for both H_{12} and ABACUS that have greater area under the curve, however it appears that ABACUS surpasses H_{12} . At 90% frequency the ROC curves suggest that methods have performed well.

In conclusion, H_{12} , in these simulations, performs significantly better than uIHS under soft selective sweeps, but not as well as ABACUS. We observe that the Bayesian approach has greater power to detect soft sweeps at lower frequencies. One reason for this could be that H_{12} requires a window in order to calculate haplotype frequencies [Garud et al., 2013]. Regions of older selection will likely experience recombination that results in the selected haplotype losing homozygosity at the edges of the window. A window based approach does have advantages in that it one can calculate simple statistics, however as we discussed in Chapter 2 changing the size of the window can drastically alter the quality of the inference.

4.3.3 Soft Sweep - Standing Variation

The results from the analysis of the soft sweep data generated by recurrent mutations are very encouraging. Unfortunately, it is not clear whether we are able to extend them to the selection on standing variation scenario, despite assertions in published literature that the patterns of variation are similar to those observed under recurrent mutations [Messer and Petrov, 2013].

To establish whether the power of ABACUS holds under these scenarios we decided to simulate regions experiencing selection on standing variation. As we mentioned above, we could not condition on the current day frequency of the selected allele, therefore we had to do a large number of simulations in order to ensure that there were enough observations in each frequency bin. Using the same null distribution

as the soft sweeps we constructed the power curves for ABACUS, uIHS, and H_{12} in Figure 4.12.

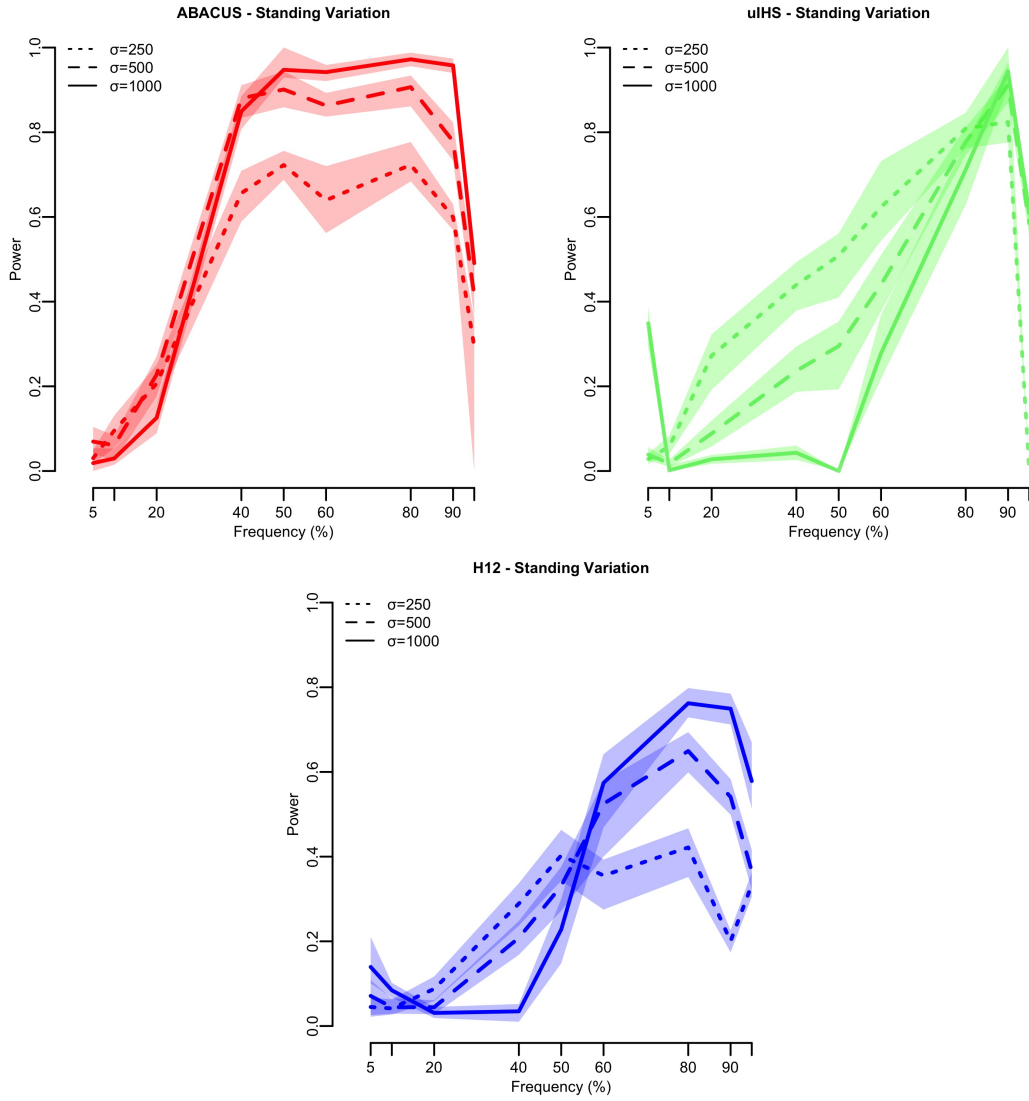


FIGURE 4.12: Power curves comparing the relationship between the true negative and true positive rate for selection on standing variation. Here we show the power curves for uIHS, H_{12} , and ABACUS at $\sigma = \{250, 500, 1000\}$. The shaded regions are calculated by performing a bootstrap sampling, with replacement, and estimating the 95% confidence intervals.

Comparing Figure 4.10 with Figure 4.12, we observe ABACUS has similar power to detect selection on standing variation and soft sweeps via recurrent mutations. Note that the power curves of uIHS and H_{12} do not always appear to increase with the strength of selection, this is likely due to stochastic variation in the

power calculations. That being said, overall we observe similar power to detect standing variation and soft sweeps by recurrent mutations.

Finally we compared the ROC curves for the three methods on the selection by standing variation datasets. As expected we observe similar patterns as in Figure 4.11, and ABACUS appears to out perform the two alternative methods across all frequencies.

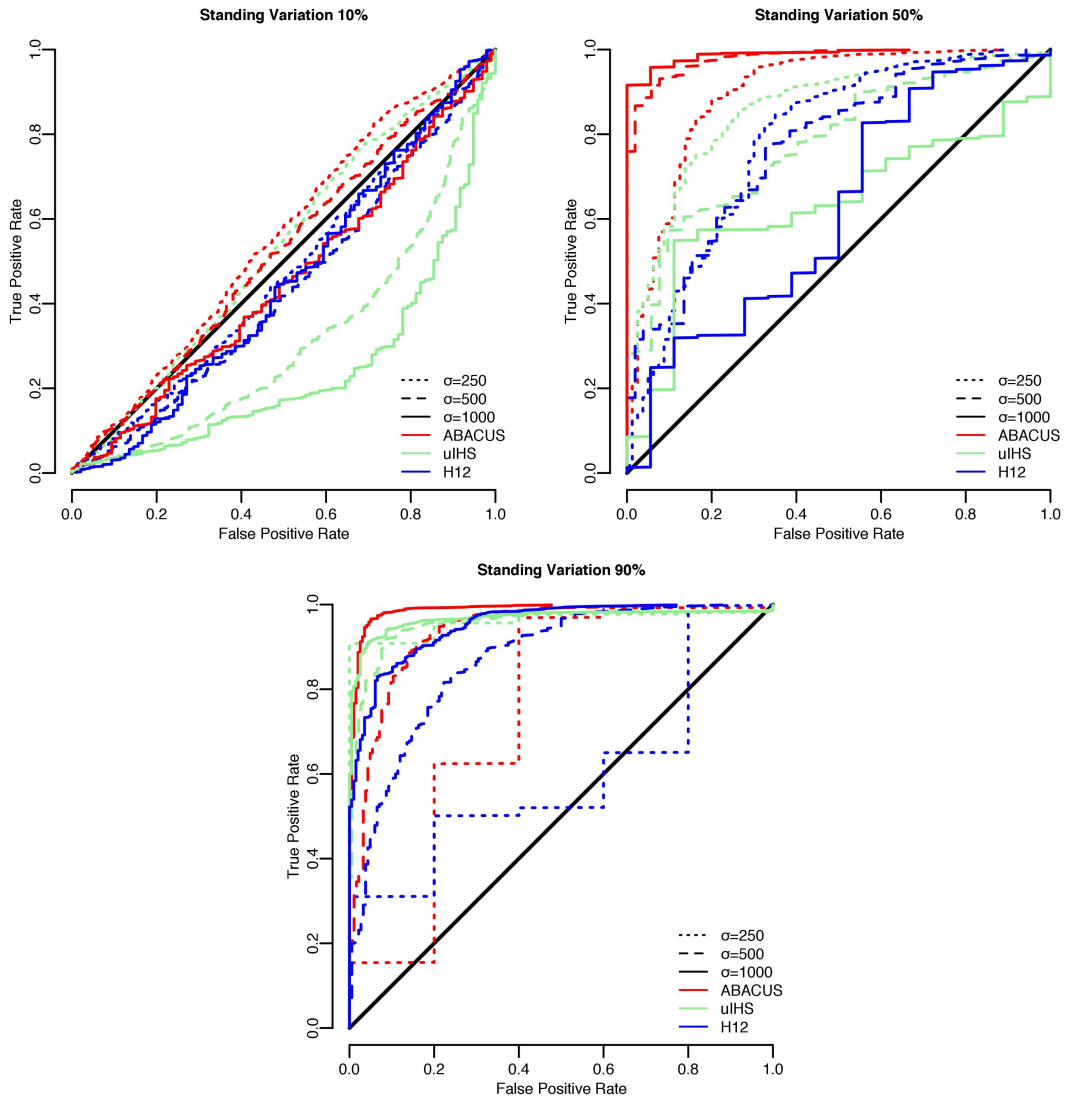


FIGURE 4.13: Receiver Operator Characteristic (ROC) curves comparing the relationship between the true negative and true positive rate for selection on standing variation. Here we show the ROC curves for uHS, H_{12} , and ABACUS for $f = \{10\%, 50\%, 90\%\}$ and $\sigma = \{250, 500, 1000\}$.

Observe, in Figure 4.13, that the ROC curves for uIHS at 10% frequency appear to suggest that the false positive rate increases faster than the true positive rate. This observation is due to the fact that under the null and alternative distributions are not simulated from the sample model. The null in this scenario is simulated by setting $\sigma = 1 \times 10^{-3}$ and allowing recurrent mutations. Therefore the observation in this ROC curves suggests that there is more diversity on the derived allele under the alternative model (i.e. higher uIHS values) than one observes under the null. Since we are performing a one sided test, i.e. a decrease in uIHS values relative to the null suggesting more extended haplotype homozygosity on the derived background the power calculations in Figure 4.12 will not be sensitive to this feature.

We have tried simulating the null from the same model, but the fact that we cannot condition on the frequency implies that it would take an intractable number of simulations to get enough observations at higher frequencies. Other simulation packages offer this feature, however we found that they were not able to scale using the parameters that we chose.

Despite these challenges, the results presented in this section suggest that ABACUS is capable of detecting regions of a genome that experience classical selection, soft sweeps by recurrent mutation, and soft sweeps by standing variation. Furthermore, ABACUS has, in general, considerably more power to detect non-classical selective sweeps than either of the two incumbent models that we considered. Moreover, we do not need to change the method according to the type of selection that we wish to identify.

4.4 Effect of Sample Size

ABACUS makes inference on local relatedness by looking at the tips of the underlying coalescent tree. To establish whether a SNP has evolved under selective pressures ABACUS compares a proxy for the time to common ancestor between those haplotypes carrying one allele to those haplotypes carrying the other. One challenge with this approach, however is that as sample sizes increase, the differences between the time to common ancestor for those individuals carrying the selected allele to those who don't will decrease.

To see this we use the model of a coalescent tree under selection proposed by Coop and Griffiths [2004]. Suppose that there are $\lfloor nf \rfloor$ chromosomes carrying the selected allele and $n - \lfloor nf \rfloor$ chromosomes carrying the non-selected allele, and the trajectory of the A allele is governed by the function $\mathcal{X}(\cdot)$; where f is the current day frequency of the selected allele. Using Figure 1.9 as a guide we find that for those haplotypes carrying the selected allele the time of first coalescent event is governed by

$$t_j^c \sim^{iid} \text{Exp} \left(t_j^c \left| \frac{1}{\mathcal{X}(\nu)} \binom{\lfloor nf \rfloor}{2} \right. \right). \quad (4.2)$$

Equally for those chromosomes not carrying the beneficial allele their time to first coalescent event is governed by

$$t_j^c \sim^{iid} \text{Exp} \left(t_j^c \left[\frac{1}{1 - \mathcal{X}(\nu)} \right] \binom{n - \lfloor nf \rfloor}{2} \right). \quad (4.3)$$

Simplifying the above distributions to remove the integer notation, $\lfloor \cdot \rfloor$, we can write down the expectation of the difference between coalescent times under these

two distributions. Moreover we can calculate its limit as $n \rightarrow \infty$.

$$d_{\nu,f}(n) = \frac{2\mathcal{X}(\nu)}{nf(nf-1)} - \frac{2(1-\mathcal{X}(\nu))}{(n-nf)(n-nf-1)} \quad (4.4)$$

$$= \frac{2}{n^2} \left[\frac{\mathcal{X}(\nu)}{f(f-\frac{1}{n})} - \frac{1-\mathcal{X}(\nu)}{(1-f)(1-f-\frac{1}{n})} \right]$$

$$\Rightarrow \lim_{n \rightarrow \infty} d_{\nu,f}(n) < \lim_{n \rightarrow \infty} \frac{2}{n^2} K = 0 : |K| < \infty \quad (4.5)$$

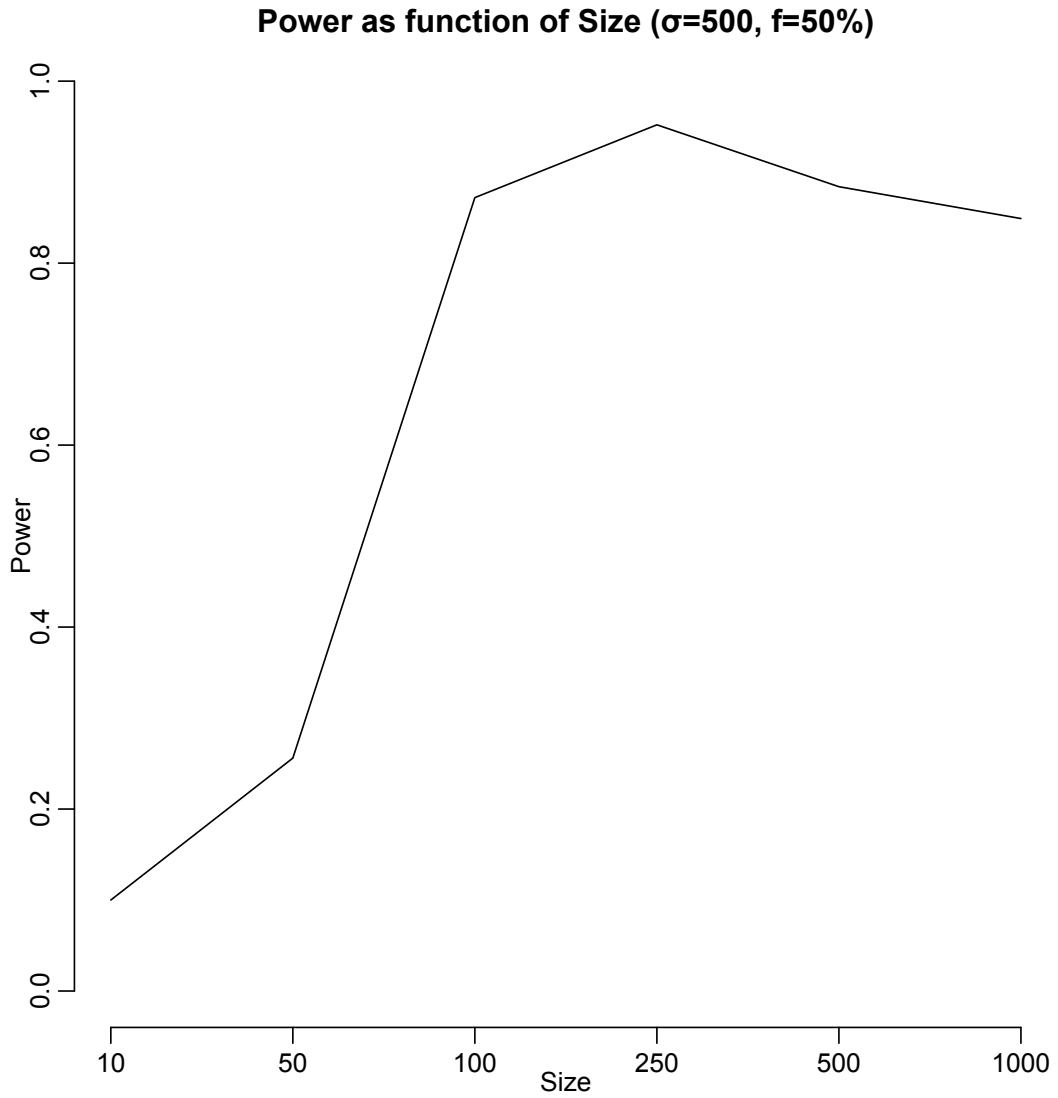


FIGURE 4.14: Relationship between number of haplotypes and the power of ABACUS to detect selection. Data simulated using a classical sweep model of selection with $\sigma = 500$, and $f = 50\%$. Each sample size was repeated 250 times.

The results in Equation 4.5 implies that as the number of samples grows the differences between the two sub-groups will be more difficult to establish. To

establish whether changes in realistic sample sizes have an effect on the ability of ABACUS to detect selection we constructed a small simulation experiment. The results can be seen in Figure 4.14.

We observe that for small sample sizes ($n \leq 50$) we have little power to detect strong selection ($\sigma = 500$) at an intermediate frequency ($f = 50\%$). As sample sizes increase the power also increases sharply, to a maximum of 0.97 at $n = 250$. We see that as sample sizes increase further the power tails off and begins to decrease.

4.5 Discussion

Building on the development of the ABACUS model in Chapter 3, we demonstrated, in this chapter, the power of the model to detect the presence of selection across a wide range of simulated scenarios.

We show that the distribution of $\log_{10}\text{BF}$ in neutrally evolving populations is slightly correlated ($\rho = -0.03$) with the frequency of the derived allele, and appears to be suppressed at the edges of the frequency spectrum. One reason for this pattern is the reduction in the number of samples in one of the comparison groups at the ends of the frequency spectrum, and less frequent derived alleles will have longer lengths due to their age. Contrasting these observations with iHH, we observe, in the tails of the frequency distribution, that there is increased variation potentially leading to an increase in the false positive rate.

Rather than looking at the raw values of $\log_{10}\text{BF}$ an alternative approach might involve a simple standardisation step not unlike that which iHS performs. The problem, however, with approach is that it would alter the interpretation of the results from ABACUS. $\log_{10}\text{BF}$ represents the level of evidence in support of a putatively ‘non-neutral’ hypothesis across all frequencies. If we were to perform

standardisation it would distort this definition and we would no longer be able to ascertain the evidence in support of either model. A more useful alternative would be to calculate empirical p-values using the observed distribution of $\log_{10} \text{BF}$.

In our simulations of demographic and structured populations, we showed that false positive rates for ABACUS remained largely stable relative to neutrally evolving samples. We explained that is was due to the fact that we make comparisons between haplotypes based on their observed level of relatedness across the genome.

We simulated structured populations to demonstrate the impact of not adjusting for within haplotype pair relatedness, however detecting selective sweeps in such settings is likely to be challenging. Suppose that the samples in our study come from two distinct populations. If an beneficial allele is present before a population split, then the frequency in the current day population will depend on the within population selective pressures, i.e. it may be more efficacious to carry that allele in one population relative to the other. Another potential scenario is that the beneficial allele is present in only one of the populations. Therefore when we bi-partition the alleles we have reduced power as it is only present in one population.

A solution to this problem would be to determine the values of ABACUS within defined populations. This problem is not unique to ABACUS, for example EHH [Sabeti et al., 2002b] suffers from similar challenges [Sabeti et al., 2007]. Indeed, XP-EHH by Sabeti et al. [2007] is designed to look for regions of the genome where a selected allele has risen to near fixation in one population by basing its inference of within population comparison of EHH scores. A potential direction of future work might be to design a version of ABACUS that incorporates information from samples from multiple populations.

In addition to population structure and demography we showed in the false positive rates do not appear to be affected by a number of recombination hotspots scenarios. This conclusion appears to hold across the frequency spectrum, however this is

more difficult to ascertain as we cannot fix the frequency of the alleles. A more detailed study will need to be undertaken in the future to empirically estimate the effect of variable recombination rates on the power to detect positive selection.

One undesirable feature of the model is that the power of ABACUS does not appear to increase monotonically with an increase in sample size. Using the model by Coop and Griffiths [2004] we explained this feature by showing that as sample sizes get very large the differences between those individuals who carry the selected allele and those who don't will tend to zero. From our experiments this doesn't appear to be a major issue with the current datasets at our disposal. For reasonably large ($n = 1000$) datasets of samples from a single population ABACUS still had good power ($p > 0.8$) to detect strong selection, however due to the complexity of the underlying inference algorithm it is not obvious that these datasets will be computationally feasible with the current implementation.

A stopgap solution to this phenomena might be to construct multiple subsamples and measure the Bayes factor of association using an approach like the bag of many bootstraps method by Kleiner et al. [2012]. Another approach may be to explore the effect of alternative HMM decoding algorithms e.g. k-segment paths (*Chris Yau - personal correspondance*), where rather than looking at the most likely path one would look at the k-most likely path. This may have the effect of looking higher in the tree therefore learning the overall structure and being less affected by increased relatedness among the sample.

One of the major challenges with performing the experiments in this chapter was that, using the simulation software *msms* [Ewing and Hermisson, 2010] we were unable to condition on the allele frequency of the beneficial allele in the current population. Other software packages, e.g. *mbf* [Teshima and Innan, 2009], can be re-engineered to have this feature, however we found that they were considerably less efficient at simulating datasets with the necessary parameters and it would

be infeasible to use them for the size of experiments that we require. From our own experience with developing coalescent simulation software, it would likely be straightforward to add this functionality to the *msms* program.

Whilst we can show under classical sweeps and soft sweeps by recurrent mutation that ABACUS performs well, we cannot conclude with the same level of assuredness that we will observe similar results in soft sweeps by standing variation. From the results presented in this chapter we have showed that ABACUS still remains powerful in these settings, however we cannot test all frequencies and strengths of selection.

Note that these results were constructed assuming that the phase of the chromosomes was known. In reality this won't be the case as phasing algorithms will introduce errors. As observed by Lawson et al. [2012] one effect of this error will be to incorrectly break or extend a painted length. This shouldn't have a significant impact on the results as we look at the relative size of the painted lengths and use that to compare between individuals at a single loci. This should mean that ABACUS is more robust to such inaccuracies, however a more extensive investigation would need to be carried out to ensure this.

In conclusion the results from this chapter show that under classical selective sweeps, ABACUS has similar power to H_{12} across all values of σ and f and less power, compared with iHS, to detect strong selection $\sigma > 250$. We observe that a significant advantage of our model is that in scenarios where the samples come from a population experiencing a soft sweep, either by recurrent mutation, or standing variation, we had equal or better power than the other two methods. These results were constructed assuming that we know the causal variant, however with sequencing technology become more readily available we will always type the causal variant. Furthermore, ABACUS is based on an HMM framework that scales linearly with positions, which makes it readily applicable to these datasets.

The evidence presented in this chapter suggests that ABACUS has good power to detect recent classical sweeps and recent soft sweeps in simulated data. Moreover, these results were constructed without the need to perform non-parametric standardisation, or windowing, and not using ancestral allelic information.

Chapter 5

An Application to 1000 Genome

“If we knew what it was we were doing, it would not be called research, would it?”

Albert Einstein

5.1 Introduction

Building on the development of ABACUS in Chapter 3, and its application to simulated data in Chapter 4, we focus in this chapter on genome wide sequence data. Using the latest release of human genome wide variation data from the 1000 Genome Project (1KG) [1000 Genomes Project Consortium et al., 2012] and ABACUS, we scan the genomes of samples from three different populations for evidence of positive selection. We also apply iHS and investigate the discrepancies between the two models in previously identified candidate regions of positive selection.

We show that both methods identify many of the well established regions of classical selective sweeps, e.g. *LCT* [Hollox et al., 2001]. In regions where there is evidence to suggest alternative methods of positive selection, e.g selection on

standing variation, we show that ABACUS appears to more readily identify these regions having patterns of variation associated with selection. Across the genome, we observe that results from ABACUS show good agreement ($> 80\%$) with previously constructed maps of selection in the human genome [Akey, 2009]. We then highlight a number of novel candidates of selection, some of which show patterns of variation consistent with a soft selective sweep.

5.2 Data

In this chapter we use genome wide sequence data from phase one of the 1000 Genome Project (1KG). At the time of publication, this dataset represents the largest collection of global variation in humans; $> 95\%$ of all common SNPs ($f > 5\%$). [1000 Genomes Project Consortium et al., 2012]

From the IMPUTE2 [Howie et al., 2009] website (http://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_integrated.html) we downloaded phased haplotypes for 1,092 individuals from phase one of 1KG. The data was phased using BEAGLE [Browning and Browning, 2007] and MaCH/Thunder [Li et al., 2011] to obtain haplotypes.

Pop.	Number	Description	Polymorphic SNPs
CEU	85	Utah residents (CEPH) with Northern and Western European ancestry	10,842,168
YRI	88	Yoruba in Ibadan, Nigeria	17,783,794
CHB	97	Han Chinese in Beijing, China	10,133,550

TABLE 5.1: The number of individuals and polymorphic SNPs for three populations of 1000 Genome Project Phase 1 [1000 Genomes Project Consortium et al., 2012]

From this dataset, we removed all insertions, deletions, and structural variants which left 36,648,992 autosomal single nucleotide polymorphisms (SNPs). Building on the discussion regarding inference across population structure from the previous chapter, we split the 1,092 samples into their respective populations; Table 5.1. We ensured that all SNPs within each population sample were polymorphic, i.e. in CEU 10,842,168 SNPs (29.4%) were polymorphic.

For this analysis we focus on those samples from the CEU, CHB and YRI populations. We chose these populations as they have been used in a number of previous studies of positive selection on human variation data, e.g. Voight et al. [2006] and Akey [2009]. Figure 5.1 shows the allele frequency distribution across the three populations.

We observe that the samples from the CHB and CEU populations tend to have similar distributions as measured by the proportion of SNPs within each population, however samples from the YRI population appear to have significantly more rare derived variants ($f \leq 5\%$). This is not unique to the YRI population as similar excess of rare variants, compared with populations of non-african ancestry, have been observed from both the African American (ASW), and Luhya Kenian (LWK) populations [1000 Genomes Project Consortium et al., 2012]. The relative upward shift in allele frequency distributions for populations of non-african ancestry could be evidence of ancestral bottlenecks in those populations [Marth et al., 2003].

This excess of rare variants will have different effects on ABACUS and iHS. In Chapter 4 we showed that when a population from which we take a sample is evolving selectively neutral the distribution of $\log_{10} \text{BF}$ at SNPs with a minor allele frequency (MAF) $\leq 5\%$ will have considerably reduced power because of the low number of haplotypes in the group carrying one allele, relative to the other.

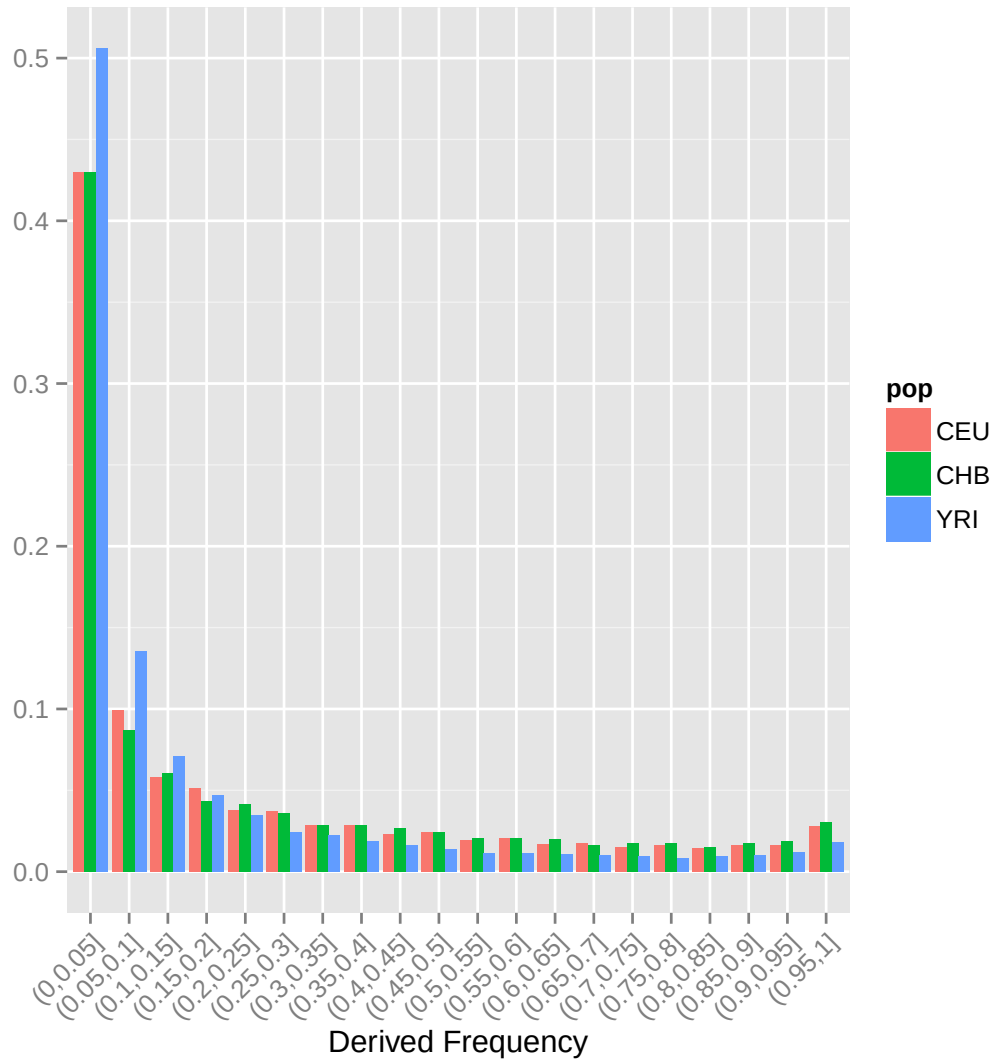


FIGURE 5.1: Barplot comparing the distribution of SNPs, stratified by derived allele frequency, across the three populations: CEU, CHB, and YRI. The height of each bar is the proportion of SNPs in that population that have that derived allele frequency.

The effect of this feature, as observed in Figure 4.2, will be to reduce the values of $\log_{10} \text{BF}$ in these frequencies relative to intermediate frequencies.

Compare this to iHS where, following the discussion by Voight et al. [2006], we do not calculate iHS at SNPs where the $\text{MAF} < 5\%$. The reason for this, as mentioned earlier, is that the reduction in observations in one of the comparison groups may artificially inflate the EHH decay. As a result, at SNPs where the derived allele frequency is very small this will decrease the value of the unstandardised iHS score

Pop.	Missing Ancestral (%)	Low MAF (%)	Overlap (%)
CEU	5.14	45.48	2.02
YRI	4.64	52.02	2.04
CHB	5.29	45.69	2.12

TABLE 5.2: The proportion of SNPs in each population that have missing ancestral information, minor allele frequencies $< 5\%$, or both.

potentially increasing the false positive rate. In addition, we do not calculate an iHS value at SNPs when it is not possible to establish which allele is derived, and which is ancestral. For this dataset we used the inferred Human Ancestral Genome obtained from Ensemble [Flicek et al., 2013]. Table 5.2 details the proportion of polymorphic SNPs in each population that have missing iHS values.

The implications of this level of missingness is that for samples from CEU and CHB it is only possible to calculate iHS scores on 51.4% and 51.14% SNPs respectively. In samples from YRI this falls to 45.38% SNPs. We have observed that the location of SNPs with $MAF < 5\%$ tend not to be clustered in any particular region of the genome. Unlike SNPs with missing ancestral information which tend to be clustered in highly polymorphic regions or regions of high recombination (data not shown).

After removing all monomorphic SNPs within a population we constructed a linearly interpolated recombination map using build 37 of the human Genome Reference Consortium (GRCh37) from the National Centre for Biotechnology Information (NCBI). Using the EM algorithm provided by *chromopainter* [Lawson et al., 2012] we estimated the effective population size for the samples from each population. Finally, we use the genome reference panel GRCh37 to identify locations of genes.

5.2.1 Method Implementations

We constructed our implementation of the ABACUS pipeline using a combination of concurrent *Java* and Linux. See Appendix C for a full description of the model, implementation, and input parameters. We used simulations to estimate the CPU run time for the three populations using the latest release of the software. The results are shown in Table 5.3.

Pop.	No. Individuals	No. SNPs	Processing Time (CPU Hrs)
CEU	85	10,842,168	85
YRI	88	17,783,794	149
CHB	97	10,133,550	103

TABLE 5.3: Detailing the processing time in CPU hours based on estimates from simulated data. (2.7×10^{-10} CPU Hrs number of haplotypes⁻² number of SNPs⁻¹).

ABACUS is designed with inbuilt parallelisation therefore using ten, twelve core nodes we can construct the painting, and calculate an estimate of the BF at every SNP in the CEU population in two hours.

To apply iHS to this dataset we used a scalable version of iHS (*Si Li Quang - personal correspondence*). It performs the same calculations as the original implementation by Voight et al. [2006], however it spreads the computation across multiple cores. Using this implementation we calculated the values of the unstandardised iHS score at every SNP in the genome and then standardised the values within 1% frequency bins.

5.2.2 Comparing ABACUS and iHS

Comparing the values of ABACUS and iHS directly will prove difficult due to the inherent confounding with allele frequencies and in how the statistics are calculated. To circumvent this we decided to compare the values by using the position

in their empirical distribution i.e. the empirical p-value. This allows us to determine how ‘significant’ a SNP is under either model. If both models are detecting similar phenomena one would expect that these values are correlated.

5.3 Existing Candidates for Positive Selection

We ran ABACUS and iHS on all polymorphic SNPs within each population group. Figure 5.2 shows the plots for the results of ABACUS applied to each population. Tables D.1 and D.2 in Appendix D provide a more detailed descriptions of the top ten regions for each population, for ABACUS and iHS respectively.

In the Manhattan plots and Tables D.1 & D.2, we observe that both models show elevated values at previously identified candidates of selection, e.g. *LCT* [Hollox et al., 2001]. Indeed, the top ten regions for ABACUS in all three populations include the *MHC* which is known to be an important region in terms of protecting an organism from pathogens [Borghans et al., 2004].

The results from ABACUS suggest, however, that the distribution of $\log_{10} BF$ in the YRI population is skewed more towards zero than the distribution of the other two populations. This might be the effect of increased rare variants, however if we compare the distributions of $\log_{10} BF$ across derived alleles in Figure 5.3(A) we see that this effect is observed at most frequencies. To understand why this is the case we compared the distribution of the average distance between breaks in the painted chromosomes for each population, Figure 5.3(B).

Compared with CEU and CHB we see that the distribution of average distance between breaks in the painting for the YRI population is considerably lower. Recalling the discussion from Chapter 3 we observed a close relationship between breaks in the painting procedure and the distance between recombination events. Using the argument that shorter distances between recombination events suggest

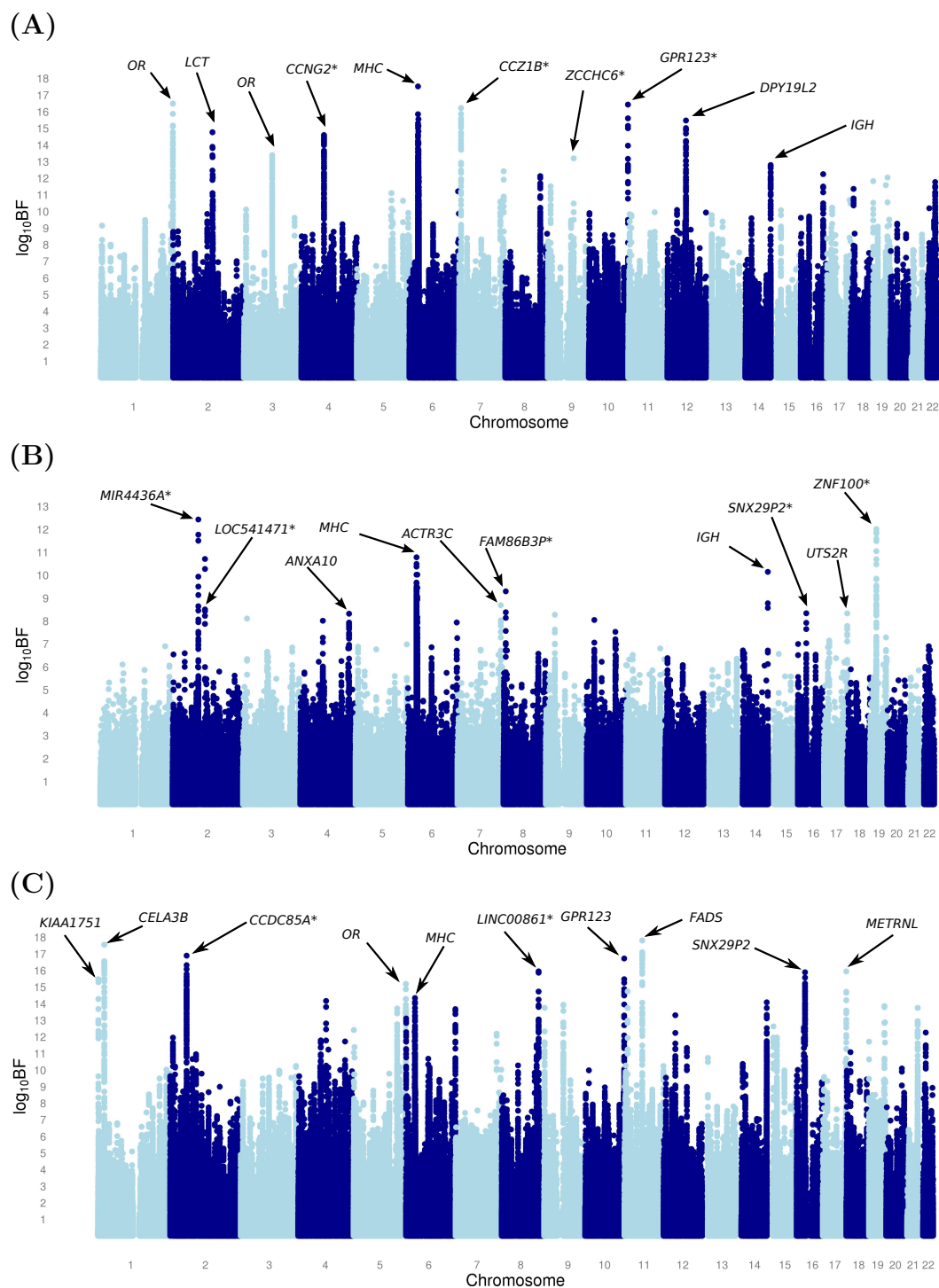


FIGURE 5.2: Evidence for selective sweeps in all autosomal chromosomes of samples from (A) CEU, (B) YRI, (C) CHB, populations obtained from 1000 Genomes Project Consortium et al. [2012]. Using *GRCh37* we have annotated the top ten peaks by the genes that the region overlaps. * are the closest genes to a highlighted region.

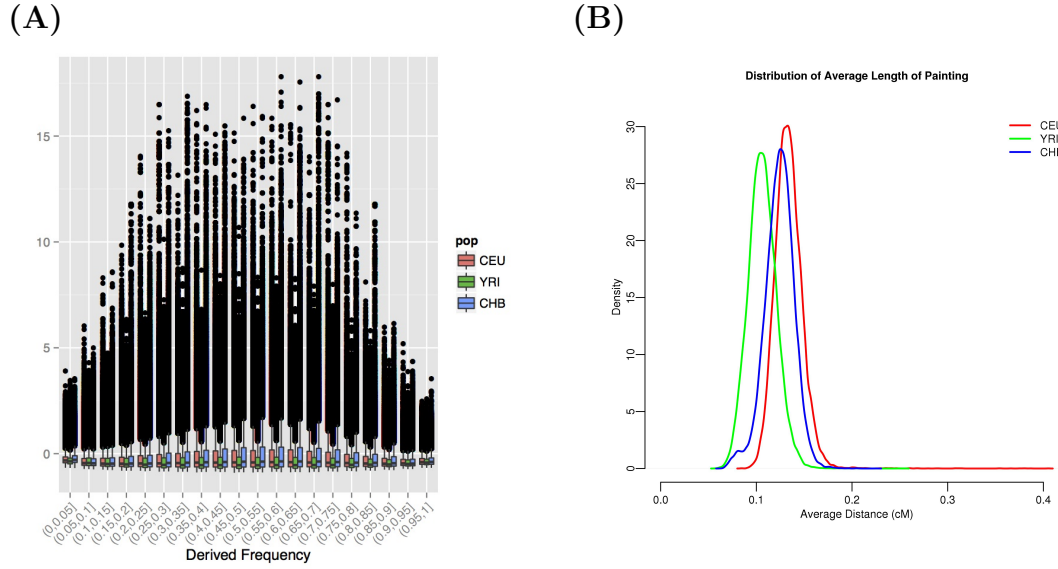


FIGURE 5.3: Analysis of results from genome wide scan of positive selection in CEU, YRI, and CHB populations. **(A)** compares the distribution of $\log_{10}$ BF stratified by derived allele frequency for each population. **(B)** compares the distribution of average lengths between breaks in a painted chromosome for each population.

longer time to common ancestor, it would appear that the YRI population, relative to CEU and CHB has a much older ancestry and the samples are less related. Recall that the effect of a bottleneck is to increase the rate of coalescence reducing the time to common ancestor among the sample. This suggests that direct numerical comparisons of $\log_{10}$ BF across different populations is not appropriate. Note, we observe that the results from iHS also exhibit an overall reduction in values of the score in the YRI population relative to the CEU and CHB populations.

5.3.1 Signals in Six Candidate Regions

To begin our analysis we focus on investigating regions of the genome that have been previously identified as having patterns of variation associated with positive selection. We used the list of classical and soft selective sweep candidate regions identified by Peter et al. [2012] to investigate the difference between the results from ABACUS and iHS. As we have discussed in previous chapters, the authors of

this paper use the candidate regions, detailed in Table 5.4, to assess the evidence for a classical selective sweep, or a sweep from standing variation.

Gene	Population	Function	Model (Peter et al. [2012])
LCT	FIN	Lactase persistence	Classical sweep
ASPM	GBR	Microcephalism	Standing variation
TRPV6	CEU	Calcium absorption	Classical / Standing variation
ADH1B	CHB	Alcohol metabolism	Classical sweep
EDAR	CHB	NF- κ B Activation	Classical sweep
PSCA	YRI	Involved in bladder & pancreas cancer	Standing variation

TABLE 5.4: The six autosomal candidate regions identified by Peter et al. [2012] to have patterns of variation associated with either classical sweeps, or soft sweeps by standing variation. Model identifies the mode of selection that Peter et al. [2012] infer from their method.

Note that throughout this section we used the population with the closest ancestry to those chosen by Peter et al. [2012]. For example, for the Finnish population (FIN) we looked at the signal of selection in the CEU population. We begin, this analysis by focusing on the evidence for the presence of a selective sweep at the LCT locus in populations of European ancestry.

5.3.1.1 CEU Candidates

LCT *LCT* encodes a protein that enables mammals to digest lactose carbohydrates (disaccharides) in milk [Hollox et al., 2001]. In the majority of mammals this ability decreases when they are weaned off milk in infancy, however variants identified in humans can lead to a phenotype known as lactase persistence, i.e. the ability to digest lactose into adulthood [Hollox et al., 2001, Bersaglieri et al., 2004].

The C/T - 13910 polymorphism (rs4988235) is observed at very high frequency in European populations, and has been shown by Enattah et al. [2002] to be 100% associated with lactase persistence in samples from Finnish populations. It was argued by Hollox et al. [2001] that the observed high frequency of lactose persistence

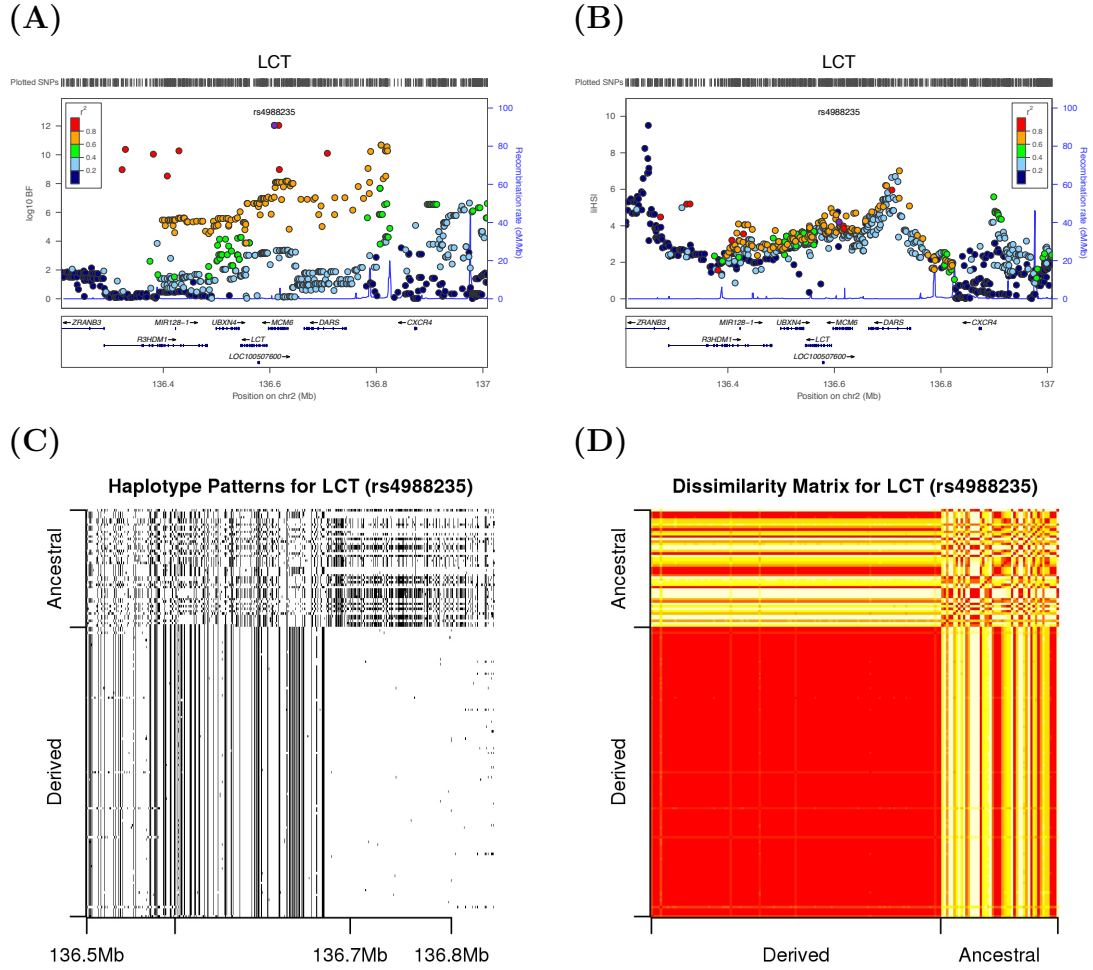


FIGURE 5.4: **(A)** and **(B)** show the regions surrounding the CEU locus, centered on rs4988235, for ABACUS and iHS respectively. SNPs are coloured according to their R^2 value with the highlighted variant. Overlaid with the GRCh37 reference genome, and recombination map. Constructed using the LocusZoom web application Pruim et al. [2010]. **(C)** Haplotype variation plot, with individuals as rows, and positions as columns. Coloured black or white depending on the allele that they carry. **(D)** Pairwise difference matrix between the haplotypes over this region. Dark red indicates low number of differences.

in European populations was as a result of the spread of agricultural techniques around the same time that modern humans were expanding into southern and eastern Europe. This variant has been observed at low frequency, or complete absence in other populations across the world, e.g. Figure 1.7 and Enattah et al. [2002].

Figure 5.4(A) shows the results from ABACUS in this region. We see that there are a considerable number of SNPs with $\log_{10} BF > 2$ in the CEU population. The

two highest values in this region corresponds to the C/T - 13910 polymorphism, and to another SNP that was also identified by Enattah et al. [2002], (rs182549), as being highly associated with lactase persistence.

Note that both of these SNPs have the same $\log_{10} \text{BF}$ value, because they both lie on the same painted segments and have an $R^2 = 1$, i.e. in complete linkage. Whilst direct numeric comparisons between populations is not appropriate we do observe considerably fewer SNPs with values of $\log_{10} \text{BF} > 2$ in both YRI and CHB (data not shown) supporting the argument by Hollox et al. [2001]. Furthermore, the causal SNP (rs4988235) is not polymorphic in either of the two alternative populations.

An equivalent analysis of the results from iHS calculated at the *LCT* locus, Figure 5.4(B), again shows a large number of SNPs with $|\text{iHS}| > 2$. We observe neither of the two proposed causal variants (rs4988235 & rs182549) have the largest value of $|\text{iHS}|$ in this region, despite having ancestral allele information and intermediate derived allele frequencies. Whilst the $|\text{iHS}|$ values for the two variants are not the highest, the iHS values are negative (-4.21, and -3.99, respectively). This suggests that the derived allele shows more evidence of extended haplotype homozygosity than the corresponding ancestral allele.

Analogous to the simulated data in Figure 1.11 we produced the haplotype plot and dissimilarity matrix for a 600 SNP window surrounding the C/T - 13910 polymorphism. Figures 5.4(C) & 5.4(D) show a large reduction in variation between haplotypes carrying the derived allele and those carrying the ancestral. On average, across this region, any two haplotypes who carry the ancestral allele at C/T - 13910 vary at 25.9% of SNPs, compared to 1.1% of SNPs for any two haplotypes that carry the derived allele.

Extending the analysis to a larger region; Figure 5.5 shows the output of a diagnostic tool that we developed to better understand the evidence of a selectively

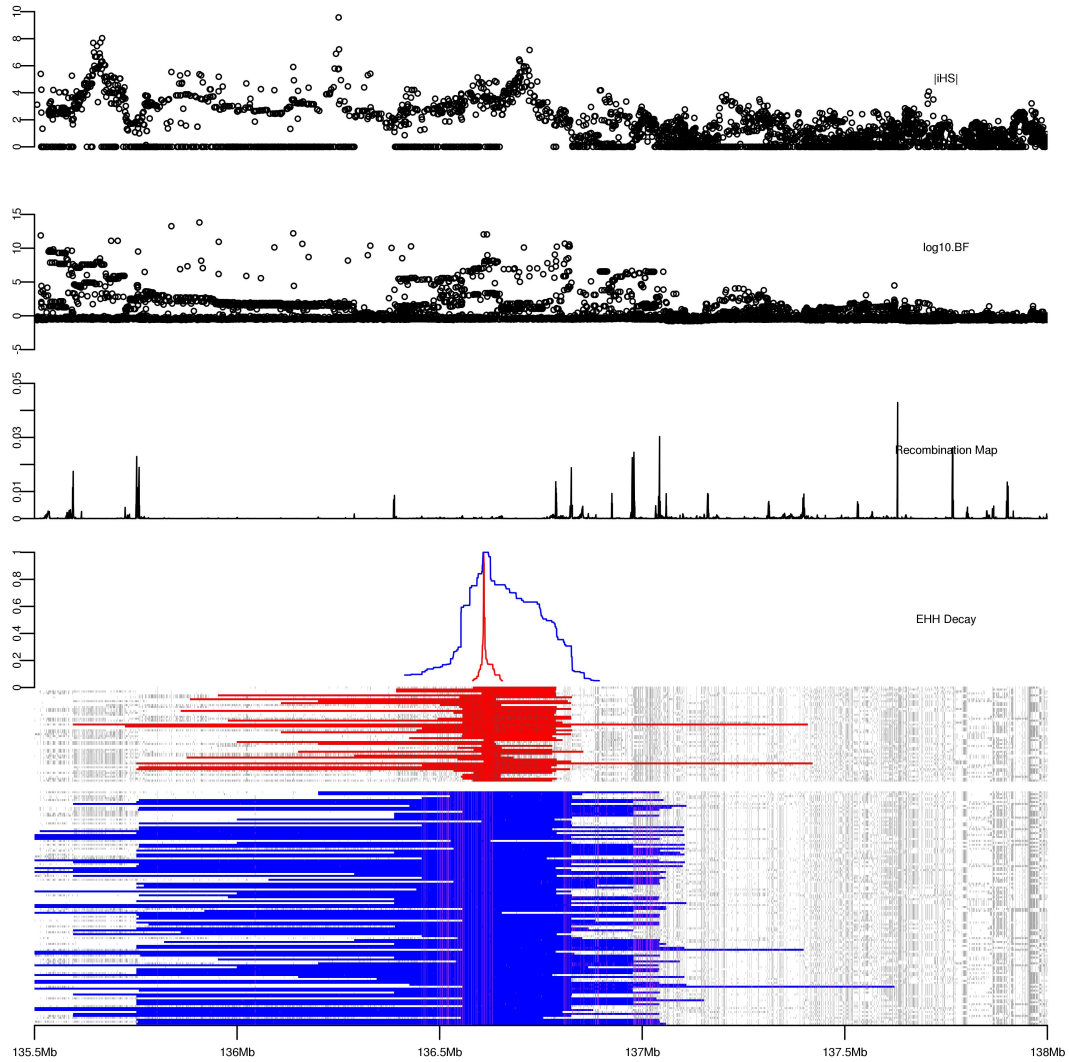


FIGURE 5.5: Diagnostic plot focusing on a 2Mb region surrounding the *LCT* locus on chromosome 2 for the CEU samples. The top panel two panels show the variation in $|iHS|$ and $\log_{10} BF$ values across this region. The middle plot shows the recombination map in cM / BP. The lower middle panel shows the EHH decay. The bottom plot shows the Viterbi lengths that overlap the SNP of interest, in this case *rs4988235*, overlaid on to the haplotype map.

non-neutral evolutionary hypothesis. It focuses on the causal C/T - 13910 polymorphism, which has a $\log_{10} \text{BF} > 12$. In the lower panel of this plot we see that those haplotypes coloured blue (derived) show distinct patterns of reduced variations relative to those haplotypes coloured red (ancestral). This is reflected in the EHH decay plots which show that the homozygosity decay is significantly slower in the blue haplotypes.

To directly compare the values of both methods at the causal SNP (rs4988235) we can calculate the empirical ranks for the corresponding metric. To calculate the rank we determine the number of SNPs with a value of that metric greater than the one we observed, i.e. an empirical p-value. Since ABACUS does have some dependency on derived allele frequency we performed the calculations for it in 5% frequency bins. For ABACUS we observe that the empirical rank of the causal SNP is 1.51×10^{-3} , compared with 2.25×10^{-3} for iHS. This suggests that both models consider this SNP to be in the tail of their corresponding distributions, and at a 1% false positive rate would reject the null hypothesis of selectively neutral evolution.

The large $\log_{10} \text{BF}$ values in this region, together with the patterns of variations suggest that the C/T - 13910 polymorphism is experiencing, or has recently experienced, a classical sweep. This conclusion is in agreement with previous studies, e.g. Peter et al. [2012] who showed that their model of a classical sweep had 98.7% posterior support compared with 1.2% for selection on standing variation [Peter et al., 2012].

ASPM It was shown by Bond et al. [2002] that *ASPM* is a major determinant factor in the size of brains. By quantifying the difference between chimpanzee and human DNA sequences at this locus, a number of studies have identified evidence of positive selection on the human lineage [Zhang, 2003, Kouprina et al., 2004]. Mekel-Bobrov et al. [2005] identified a particular SNP, A44871G (rs41310927)

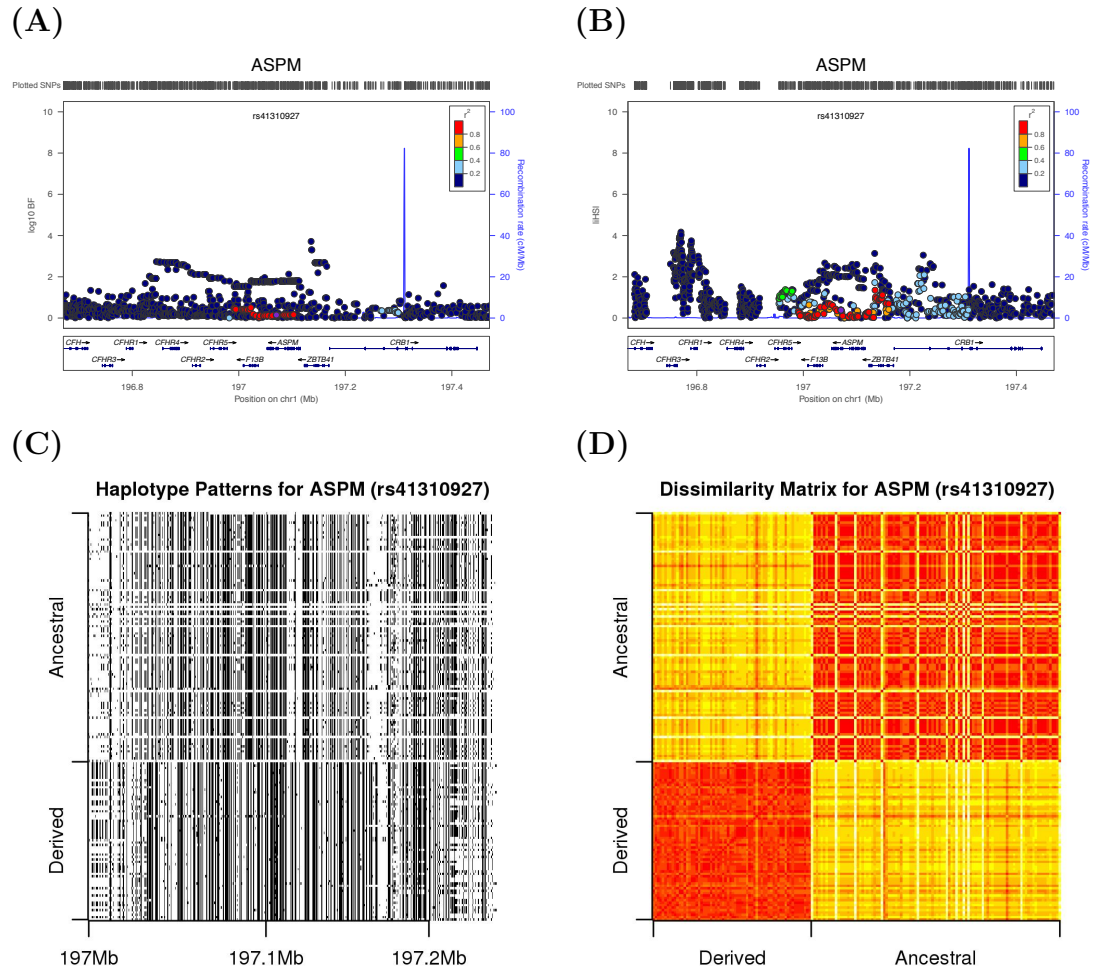


FIGURE 5.6: (A) and (B) show the regions surrounding the *ASPM* locus, centered on rs41310927, for ABACUS and iHS respectively. SNPs are coloured according to their R^2 value with the highlighted variant. Overlaid with the GRCh37 reference genome, and recombination map. Constructed using the LocusZoom web application [Pruim et al., 2010]. (C) Haplotype variation plot, with individuals as rows, and positions as columns. Coloured black or white depending on the allele that they carry. (D) Pairwise difference matrix between the haplotypes over this region. Dark red indicates low number of differences.

which they estimate arose 5,800 years ago and was swept up to high frequency. Focusing on this SNP, Figures 5.6(A) & 5.6(B) show the results for ABACUS and iHS in this region.

The number of SNPs with $\log_{10} \text{BF} > 2$, or $|iHS| > 2$ around the focal SNP in Figure 5.6 suggests that this region exhibits patterns of variation associated with positive selection. The haplotype variation plot in Figure 5.6(C) and diagnostic plot in Figure E.1 suggest that there is a considerable reduction in variation on

the derived background relative to those haplotypes carrying the ancestral allele. Furthermore, the dissimilarity matrix in Figure 5.6(D), suggests that there is a single haplotype on the derived background which appears to be very different to those on the ancestral. Nevertheless, the empirical ranks for the causal SNP for ABACUS and iHS are 0.238, and 0.672, respectively, suggesting that neither method consider this SNP to have patterns of variation associated with positive selection.

Despite the fact that neither method consider the proposed causal SNP to have patterns of variation associated with positive selection, this region does have a considerable number of SNPs with elevated values under both ABACUS and iHS. Moreover, the observation, at the causal SNP, that there appears to be a single haplotype on the derived background, does suggest that this region has patterns of variation consistent with a classical selective sweep. Note, this observation is at odds with Peter et al. [2012] who conclude that this region has a posterior probability of 0.87 of having patterns of variation associated with selective sweep on standing variation, 0.13 from a classical sweep, and 2×10^{-7} from a neutral model.

TRPV6 The next candidate in populations of European Ancestry, *TRPV6*, was shown by Akey et al. [2004] to reside in a 115Kb region of chromosome 7 that exhibited patterns of variation associated with positive selection. *TRPV6* encodes a protein that is responsible for calcium absorption [Akey, 2006], and whilst Peter et al. [2012] concluded that there was evidence of selection in this region their method was unable to distinguish between a classical sweep model or a soft sweep.

Using the first non-synonymous SNP (rs4987682), identified by Akey [2006], we centred a 1Mb window on the locus and looked at the results from ABACUS and iHS this region; Figure 5.7. The elevated values of both method, in Figures 5.7(A)

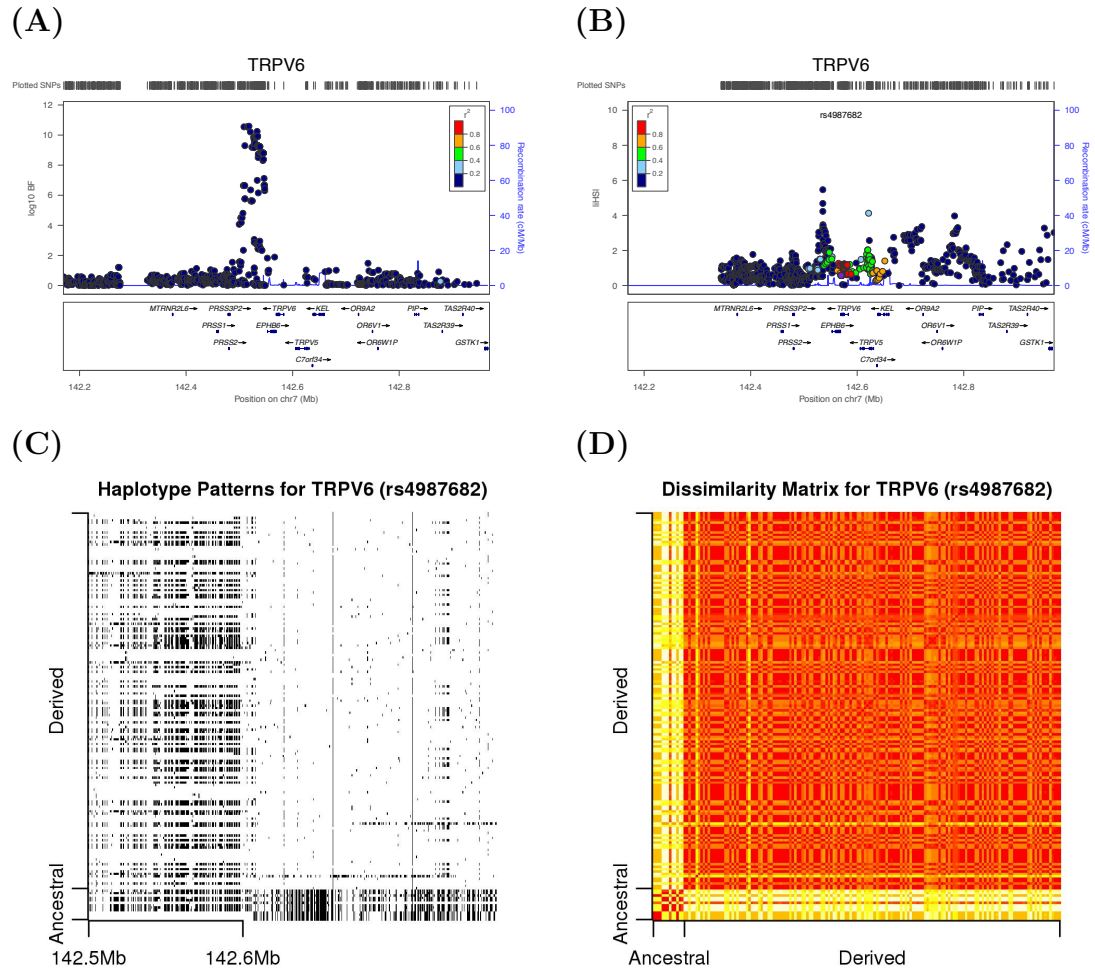


FIGURE 5.7: (A) and (B) show the regions surrounding the TRPV6 locus, centered on rs4987682, for ABACUS and iHS respectively. SNPs are coloured according to their R^2 value with the highlighted variant. Overlaid with the GRCh37 reference genome, and recombination map. Constructed using the LocusZoom web application [Pruim et al., 2010]. (C) Haplotype variation plot, with individuals as rows, and positions as columns. Coloured black or white depending on the allele that they carry. (D) Pairwise difference matrix between the haplotypes over this region. Dark red indicates low number of differences.

& 5.7(B), suggest that this region exhibits patterns of variation associated with positive selection. Note, however, that neither appear to indicate that the focal SNP is the target of selection. Observe in Figure 5.7(C), that haplotype variation suggests that there is considerably reduced variation on the derived background.

The reduction in differentiation in the dissimilarity matrix in Figure 5.7(D) together with the variation in Figure 5.7(C) suggests that there is a single, short haplotype on the derived background that has experienced recombination upstream

and downstream of the focal SNP increasing the variation. Figure E.2 indicates that the painted lengths do capture this feature, however a large number of segments switch at the recombination hotspot 50Kb upstream, as a result only a few of them will have longer lengths. Also, since the frequency of the derived allele is 92.4% ABACUS will have little power to detect selection at this SNP.

Under iHS, there does appear to be more SNPs with $|iHS| > 2$, relative to the number of SNPs with $\log_{10} BF > 2$ under ABACUS. Nevertheless, the patterns surrounding the focal SNP in Figure 5.7(B) appear to suggest that the target of selection is not the focal SNP (rs4987682). Indeed it is interesting to observe that both methods indicate that there is a region that has much stronger evidence of patterns of variation associated with positive selection 100Kb downstream of the focal SNP. The empirical ranks for the causal SNP for ABACUS and iHS is 0.225, and 0.529, respectively. This again suggests that despite the observed patterns of variation neither method have enough evidence to reject the null hypothesis of selectively neutral evolution.

The observation of a single core haplotype on the derived background might indicate that this region has experienced a classical selective sweep, however the increased variation on the derived background due to the increased recombination makes it difficult to rule out a soft sweep. This increased variation, may have lead to the inconclusive results by Peter et al. [2012]. Note, to perform their analysis, Peter et al. [2012] use an older recombination map from International HapMap Consortium [2005].

5.3.1.2 CHB Candidates

ADH1B The *ADH1B* encodes a enzyme involved in the alcohol degradation pathway, and it is suggested to be one of the major genetic factors influencing alcoholism risk [Li et al., 2007]. A mutation, Arg47His (rs1229984), is thought to

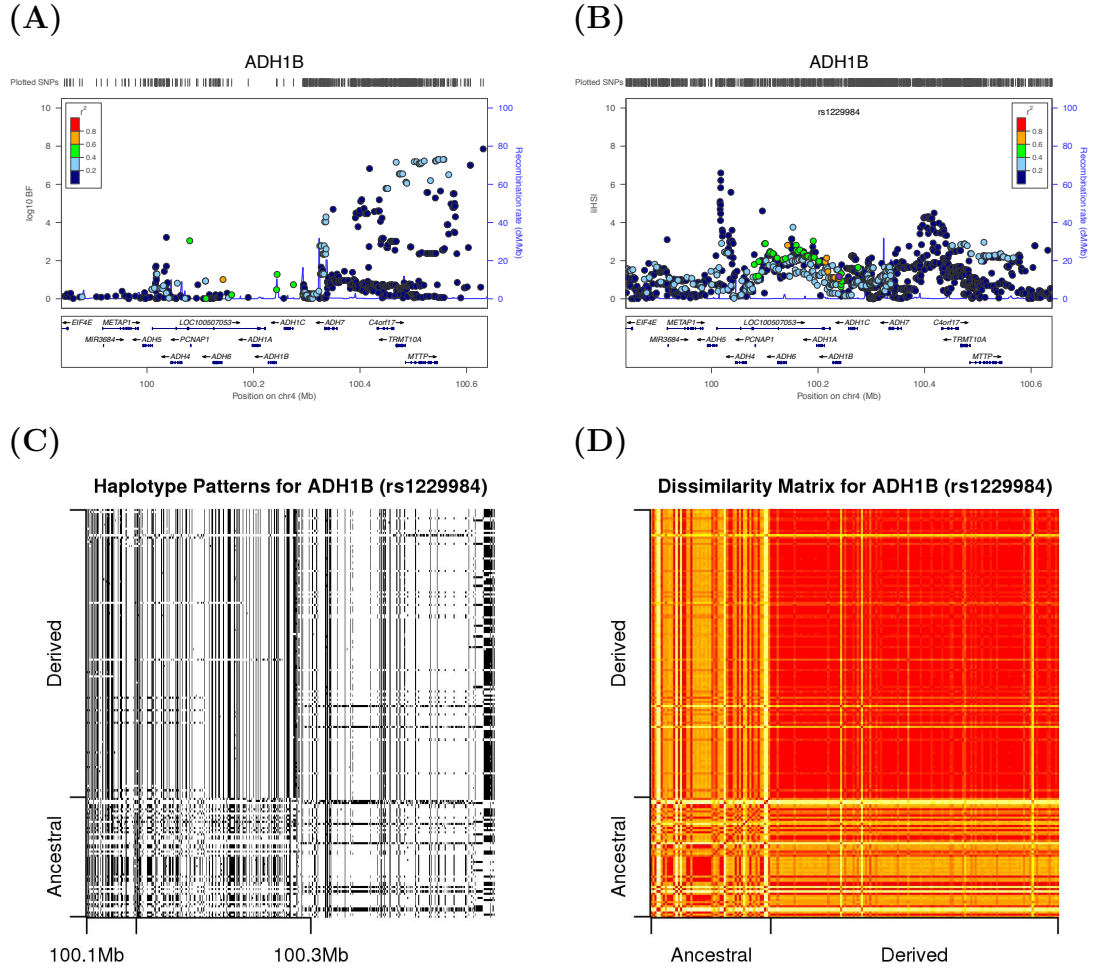


FIGURE 5.8: (A) and (B) show the regions surrounding the ADH1B locus, centered on rs1229984, for ABACUS and iHS respectively. SNPs are coloured according to their R^2 value with the highlighted variant. Overlaid with the GRCh37 reference genome, and recombination map. Constructed using the LocusZoom web application [Pruim et al., 2010]. (C) Haplotype variation plot, with individuals as rows, and positions as columns. Coloured black or white depending on the allele that they carry. (D) Pairwise difference matrix between the haplotypes over this region. Dark red indicates low number of differences.

be under selection due to its increased enzyme activity and estimates of its age relative to its frequency in the population [Li et al., 2007, Peter et al., 2012].

Using this SNP as a focal point we looked at the variation in ABACUS and iHS values surrounding this locus. Figures 5.8(A) & 5.8(B) show the results for ABACUS and iHS, respectively.

If we accept a selectively non-neutral evolutionary hypothesis at this SNP, we

would conclude from Figure 5.8 that, since there isn't a large number of SNPs with $\log_{10} BF > 2$, iHS, in this region, is better able to detect the patterns of variation consistent with positive selection. The recombination landscape in Figures 5.8(A) & 5.8(B), together with the variation on the derived background in Figures 5.8(C) & E.3 suggest that this is a region of high recombination. As a result, the dissimilarity matrix in Figure 5.8 (D), compared with previous examples, shows a decrease in the number of SNPs which differ between those haplotypes carrying the derived allele, and those carrying the ancestral allele. As before, the empirical ranks for the causal SNP (rs1229984) for ABACUS and iHS are 0.759, and 0.211 respectively, suggesting that neither method can reject the null hypothesis.

It is challenging to make a conclusion about whether this region is under selection, or not. The number of SNPs with $|iHS| > 2$ might lead one to conclude that this region does have patterns of patterns of variation associated with selection, however in the 600 SNP window centred on the focal SNP (rs1229984) suggests that there is considerable similarity between haplotypes on the derived and ancestral backgrounds. This could be a feature of the multiple recombination hotspots in the region, and further analysis would need to be undertaken to understand the nature of the evolution at this locus. Note that the method by Peter et al. [2012] calculates the posterior probability of neutral evolution in this region at 1×10^{-8} .

EDAR Another candidate region for populations of Asian ancestry is the *EDAR* locus. *EDAR* was shown by Fujimoto et al. [2008] to be associated with the thickness of hair, and has been identified by a number of previous genome wide selection scans [Akey et al., 2004, Voight et al., 2006]. Bryk et al. [2008] showed, using a series of in vitro experiments, that a SNP (rs3827760), associated with these phenotypes, appears to be under positive selection. As before we use this SNP as a focal point we centred a 1Mb window on the locus and looked at the variation in ABACUS and iHS values in the region; Figure 5.9

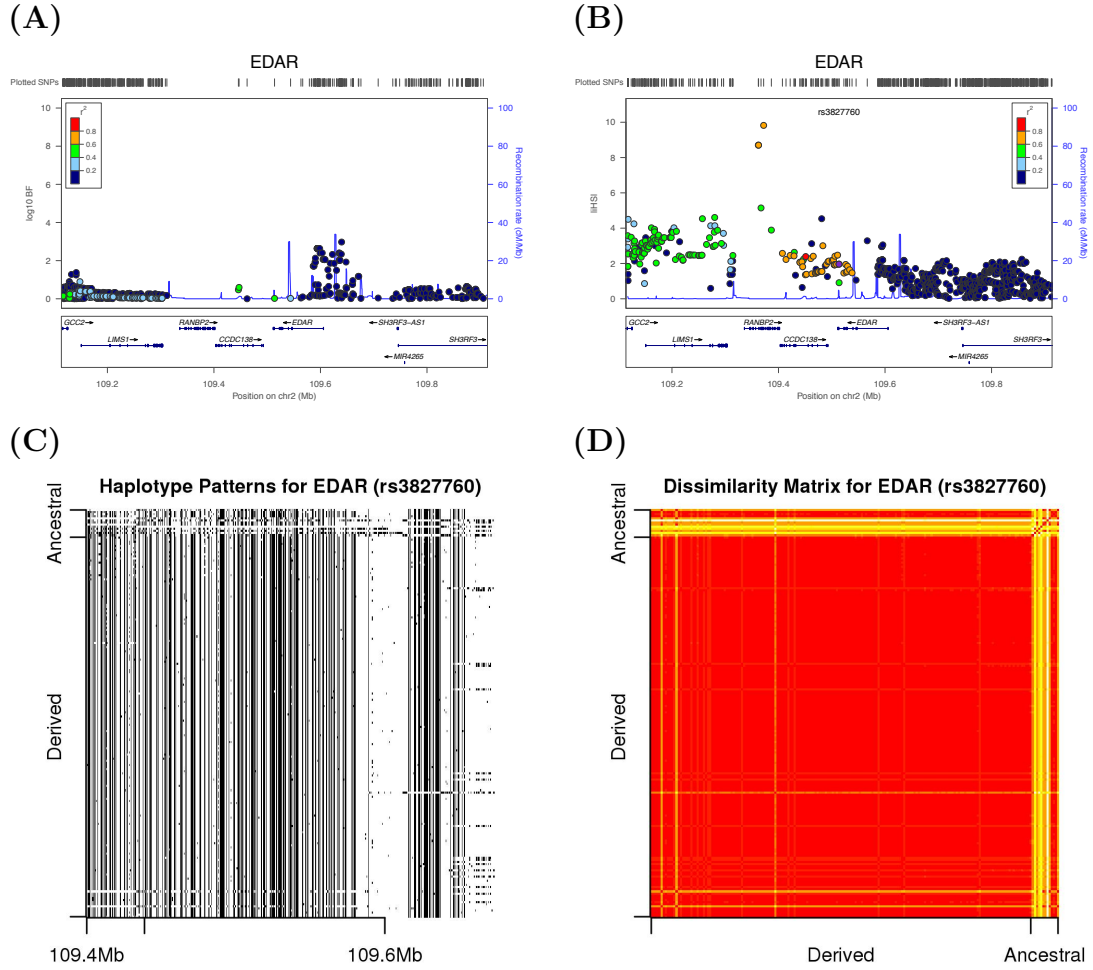


FIGURE 5.9: (A) and (B) show the regions surrounding the EDAR locus, centered on rs3827760, for ABACUS and iHS respectively. SNPs are coloured according to their R^2 value with the highlighted variant. Overlaid with the GRCh37 reference genome, and recombination map. Constructed using the LocusZoom web application [Pruim et al., 2010]. (C) Haplotype variation plot, with individuals as rows, and positions as columns. Coloured black or white depending on the allele that they carry. (D) Pairwise difference matrix between the haplotypes over this region. Dark red indicates low number of differences.

It would appear from Figures 5.9(A) & 5.9(B) that neither method identifies the proposed causal variant (rs3827760) as having the largest value, however ABACUS appears to identify a region of increased evidence upstream of the focal SNP. The patterns of variation in Figures 5.9 (C) & 5.9 (D), suggest that there is evidence of very strong classical selective sweep in this region. There appears to be one single haplotype on the derived background at a high frequency ($f = 93\%$). The painted lengths in Figure E.4 do appear to suggest that the time to common ancestor

among the derive alleles is much shorter, however because there only a handful of observations in the comparison group ABACUS has little power to detect the causal variant.

iHS does seem to have better power in this regions but considering the relative difference between the decays on the ancestral (red), and derived (blue), alleles in Figure E.4, it is surprising that we do not see a much stronger signal of selection in this region. One explanation may be that this region appears to be a recombination cold spot, therefore the decays are likely to be down weighted to account for the reduction in recombination. The empirical ranks for ABACUS and iHS are 0.348 and 0.049, again suggesting that at the 1% false positive rate we would reject the alternative hypothesis. Note, however, that iHS does have a considerably higher rank than the corresponding rank for the ABACUS value, suggesting that there is more evidence under iHS relative to ABACUS.

5.3.1.3 YRI Candidate

PSCA *PSCA* is involved with bladder and pancreatic cancer, and a particular SNP (rs2294008) has been associated with positive natural selection [Bhatia et al., 2011]. Using this SNP as a focal point we centred a 1Mb window on the locus and looked at the variation in ABACUS and iHS values in the region. Figures 5.10(A) & 5.10(B) show the results for ABACUS and iHS, respectively.

The results from Figures 5.10(A) & 5.10(B) indicated that both methods identify a region surrounding the focal SNP as having patterns of variation associated with selection, however there is some uncertainty as to which SNP is ‘causal’. We observe that the SNP with the largest value under ABACUS has an $R^2 = 1$ with the focal SNP (rs2294008). Nevertheless, the empirical ranks for the proposed causal SNP are 6×10^{-3} and 0.031 for ABACUS and iHS, respectively. As a result,

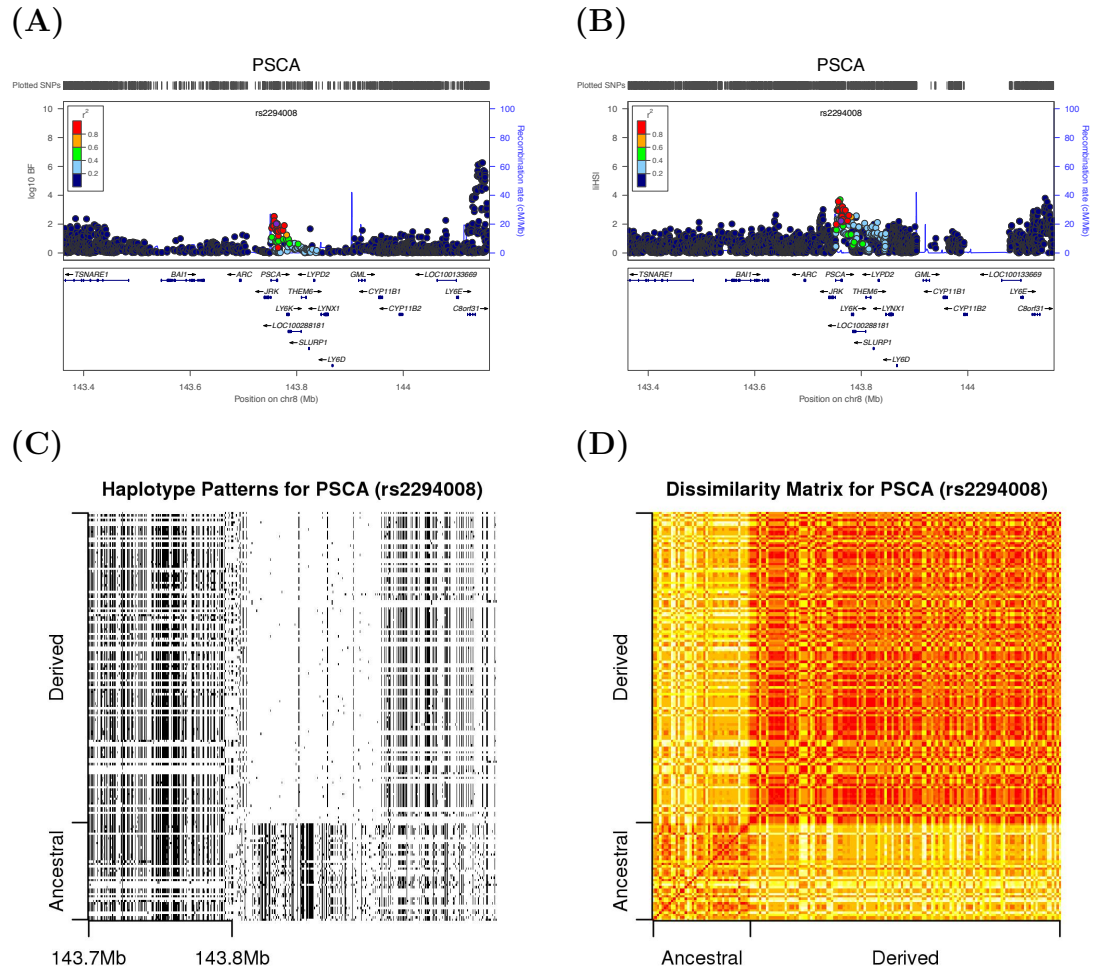


FIGURE 5.10: Comparison of YRI Candidates. **(A)** and **(B)** show the regions surrounding the PSCA locus, centered on rs2294008, for ABACUS and iHS respectively. SNPs are coloured according to their R^2 value with the highlighted variant. Overlaid with the GRCh37 reference genome, and recombination map. Constructed using the LocusZoom web application [Pruim et al., 2010]. **(C)** Haplotype variation plot, with individuals as rows, and positions as columns. Coloured black or white depending on the allele that they carry. **(D)** Pairwise difference matrix between the haplotypes over this region. Dark red indicates low number of differences.

under ABACUS, at a 1% false positive rate, we would reject the null hypothesis of selectively neutral evolution.

It would appear, from Figures 5.10(C) & 5.10(D) and the diagnostic plot in Figure E.5, that there are multiple haplotypes carrying the derived allele, as evidenced by the structure of the dissimilarity matrix in Figure 5.10(D), however they are relatively short. One reason for this is that the focal SNP lies between two regions of

increased recombination, therefore the haplotype will get broken up much quicker than in other regions. The results from this locus suggest that this region exhibits patterns of variation associated with selection on standing variation, which is in agreement with the results from Peter et al. [2012].

5.3.2 Summary

The results from the analysis above have been summarised in Table 5.5. Overall, we observe that the results fit well with our observations from the previous chapter. In classical selective sweeps we expect iHS to have better power when the frequency of the derived allele is high ($> 90\%$), e.g. *TRPV6* and *EDAR*, and at intermediate frequencies we expect both methods to have similar power, e.g. *LCT* and *ADH1B*. In soft sweeps we expect to ABACUS to out perform iHS across all frequencies, e.g. *PSCA*.

Gene	Population	SNP	Derived Fre- quency	$\log_{10}$ BF (rank)	iHS (rank)
LCT	CEU	rs4988235	0.71	12.04 (1.51×10^{-3})	4.21 (2.25×10^{-3})
ASPM	CEU	rs41310927	0.39	0.15 (0.24)	0.38 (0.62)
TRPV6	CEU	rs4987682	0.92	-0.24 (0.23)	0.57 (0.53)
ADH1B	CHB	rs1229984	0.71	-0.62 (0.76)	1.16 (0.21)
EDAR	CHB	rs3827760	0.93	-0.37 (0.35)	1.96 (4.90×10^{-2})
PSCA	YRI	rs2294008	0.76	2.02 (6.29×10^{-3})	2.22 (3.12×10^{-2})

TABLE 5.5: Summary of results from ABACUS and iHS applied to six autosomal candidate regions identified by Peter et al. [2012] using data from 1000 Genomes Project Consortium et al. [2012].

We also observed that a number of examples where ABACUS more clearly identifies the causal SNP, or a SNP that is in complete linkage with it, e.g. *LCT* and *PSCA*.

It would not be appropriate, or possible to, extend the conclusions of this fine scale analysis to the rest of the genome. As a result, in order to assess the quality of the results and the degree of overlap between the methods, and with previous

studies, across the genome it is necessary to construct an automated approach to identifying regions.

5.3.3 Overlap with Previous Studies

Akey [2009] constructed an integrated map of positive selection in the human genome, by analysing eight previous genome-wide selection studies, Table 5.6, and established degree of overlap between regions. Those regions that were identified by two or more studies are included. Using this dataset we can compare the results from both models across the three populations and determine which regions overlapped.

Reference	Feature	Sample
Carlson et al. [2005]	Frequency Spectrum	Perlegen
Wang et al. [2006]	LD	HapMap, Perlegen
Voight et al. [2006]	LD	HapMap
Kelley [2006]	Frequency Spectrum	Perlegen
Tang et al. [2007]	LD	HapMap, Perlegen
Kimura et al. [2007]	LD	HapMap
Williamson et al. [2007]	Frequency Spectrum	Perlegen
Sabeti et al. [2007]	Population Differentiation / LD	HapMap

TABLE 5.6: A list of methods used by Akey [2009] to produce an integrated map of natural selection in the human genome. HapMap data was obtained from International HapMap Consortium [2005], and Perlegen from Hinds et al. [2005].

To calculate the boundaries of a region that either method identifies proved to be challenging. Visual identification of these regions is relatively straightforward, however this is not feasible when the number of observations is on the order of 1×10^6 . To establish whether a signal of selection is present in either ABACUS or iHS we first identify all SNPs in the genome that have a value above a given threshold. Then using a 50Kb window centred on each SNP we determine the width of a region by the number of SNPs that have overlapping windows.

Within each population, and for each method we set the cut off value to be the upper 1% percentile for that metric, observed across the genome. In addition, to a universal 1% cut off for the values of $\log_{10} \text{BF}$ under ABACUS we also performed the same calculations when we considered different 1% cut off values for the distribution of $\log_{10} \text{BF}$ within 5% frequency bins. In the CEU example this results in 10.2% of the genome being highlighted as being under positive selection by ABACUS. Compared with 18.3% for iHS applied to the same sample, and 8.50% for the integrated map of positive selection [Akey, 2009].

We begin by looking at the distribution of $\log_{10} \text{BF}$ as a function of $|\text{iHS}|$. Figure 5.11 shows the relationship between these two values at every SNP in the CEU sample where it has a value of both metrics. We split the SNPs according to MAF to highlight the differences.

We observe that there appears to little SNP by SNP correlation between the two methods. The two simple linear models, $(\log_{10} \text{BF} = a + b|\text{iHS}|)$ predict that there is an increasing relationship between the two metrics ($b = 0.085$, and $b = 0.249$), however there is significant heteroscedasticity. We might expect considerable deviation at SNPs that show evidence of soft selective sweeps, which might, to some extent, explain the dense cloud of points with large $\log_{10} \text{BF}$ and low $|\text{iHS}|$. The large cloud of red observations with low $\log_{10} \text{BF}$ and larger $|\text{iHS}|$, is likely evidence of iHS having better power at lower frequencies.

If we consider the overlap of ABACUS with previous studies, we observe that 88.6% (86.0% adjusting for derived frequency) of regions identified by Akey [2009] overlap with regions identified by ABACUS in at least one of the three populations that we explored from 1KG. Using an equivalent definition of a region for iHS we calculated that the degree of overlap is 90.4%, however iHS is in the list of methods used by Akey [2009]. Within each population, of those regions identified by ABACUS, 58.9% (57.4% adjusting for derived frequency) are also identified

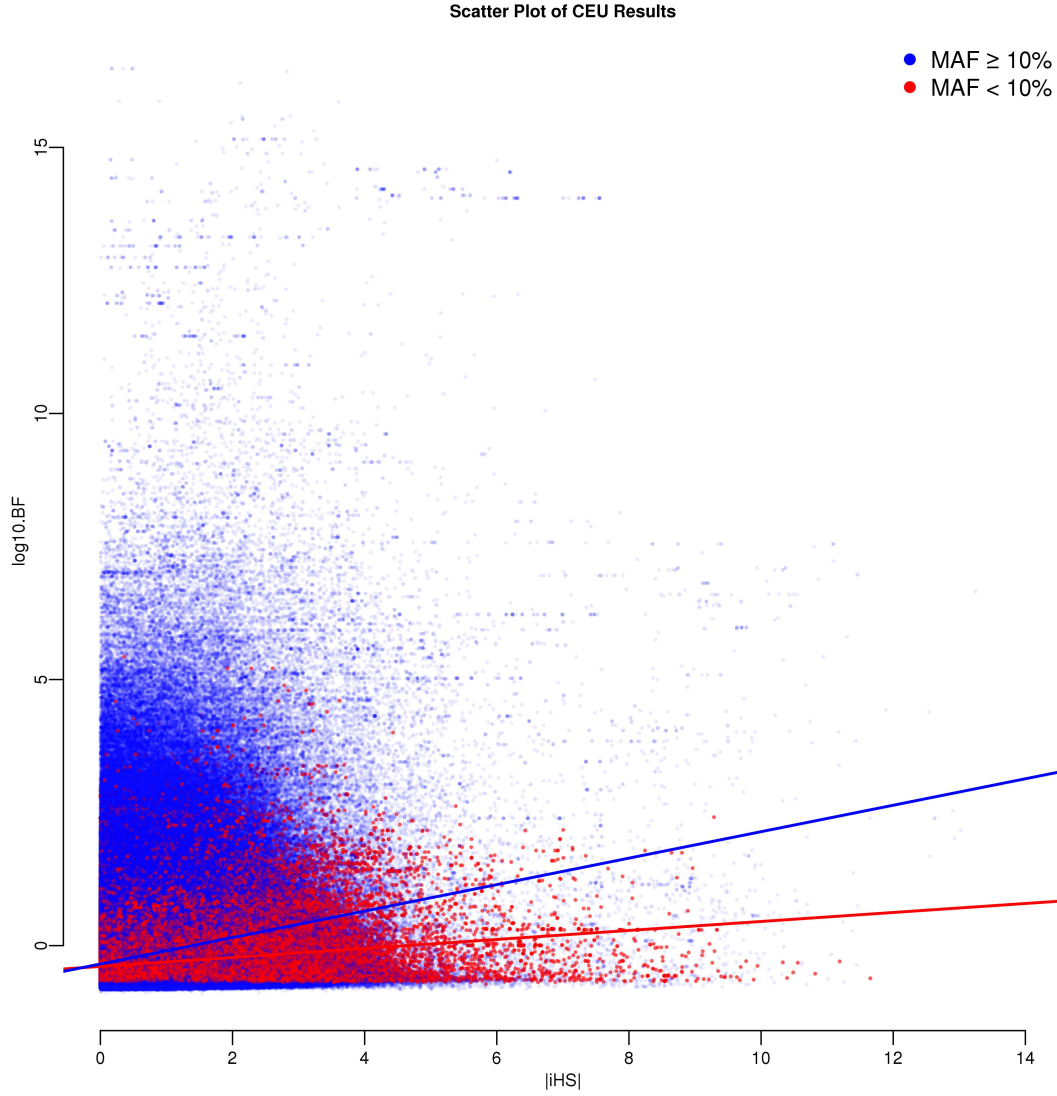


FIGURE 5.11: Scatter plot of $\log_{10} BF$ against $|iHS|$ for SNPs in the CEU sample, split by MAF. The red observations correspond to SNPs with MAF $< 10\%$, and blue observations correspond to SNPs with MAF $\geq 10\%$. The straight lines correspond to a simple linear model, $\log_{10} BF = a + b|iHS|$, within frequency class.

by iHS . Moreover, in regions that are identified by both ABACUS and iHS , they have on average higher maximum values (25.9%, and 10.3% higher respectively) compared to regions that are identified by just one of the methods. Note using the definition of a regions above the regions identified by iHS cover 34.5% of the genome and 32.5% for ABACUS. Therefore the probability of chance overlap between the two methods across the genome is 11.2% (calculated by take the product of the genome wide coverage for each method). Note that the observed

overlap is 8 fold that of the chance overlap.

The quality of this scheme depends entirely on the width of the windows; too wide and every SNP would be in the same region, too short and only a handful of regions would overlap. We tried a number of different window widths and found that 50Kb resulted in a good trade off between visual and automated identification. These results are quite encouraging, and suggest that there is good agreement previously constructed maps of positive selection in the human genome. Note, Akey [2009] observed that the degree of overlap between any two independent selection scans was on average 14.1%. Therefore the fact that we observe an overlap of $> 50\%$ with iHS suggests that ABACUS results in better agreement with existing techniques.

5.4 Novel Candidates for Positive Selection

Finally, we decided to look in more detail at examples of some regions that appear to have a considerably lower empirical p-values under ABACUS compared with iHS. The regions presented in this section were chosen from this group due to the fact that they occur in interesting regions of the genome, and because they appear to show evidence of soft selective sweeps.

DPY19L2 The first region focuses on SNP, rs4762993, in the *DPY19L2* gene in the CEU population. *DPY19L2* encodes a protein that is highly expressed in the testis and is required for sperm head elongation and acrosome formation during spermatogenesis, and it was discovered by ElInati et al. [2012] that deletion of *DPY19L2* is highly associated with *Globozoospermia*; a form of male infertility.

We observe in Figure 5.12(A) a large number of SNPs that have $\log_{10} \text{BF} > 2$ which align with the *DPY19L2* gene. The fact that we observe values $\log_{10} \text{BF} > 10$ in

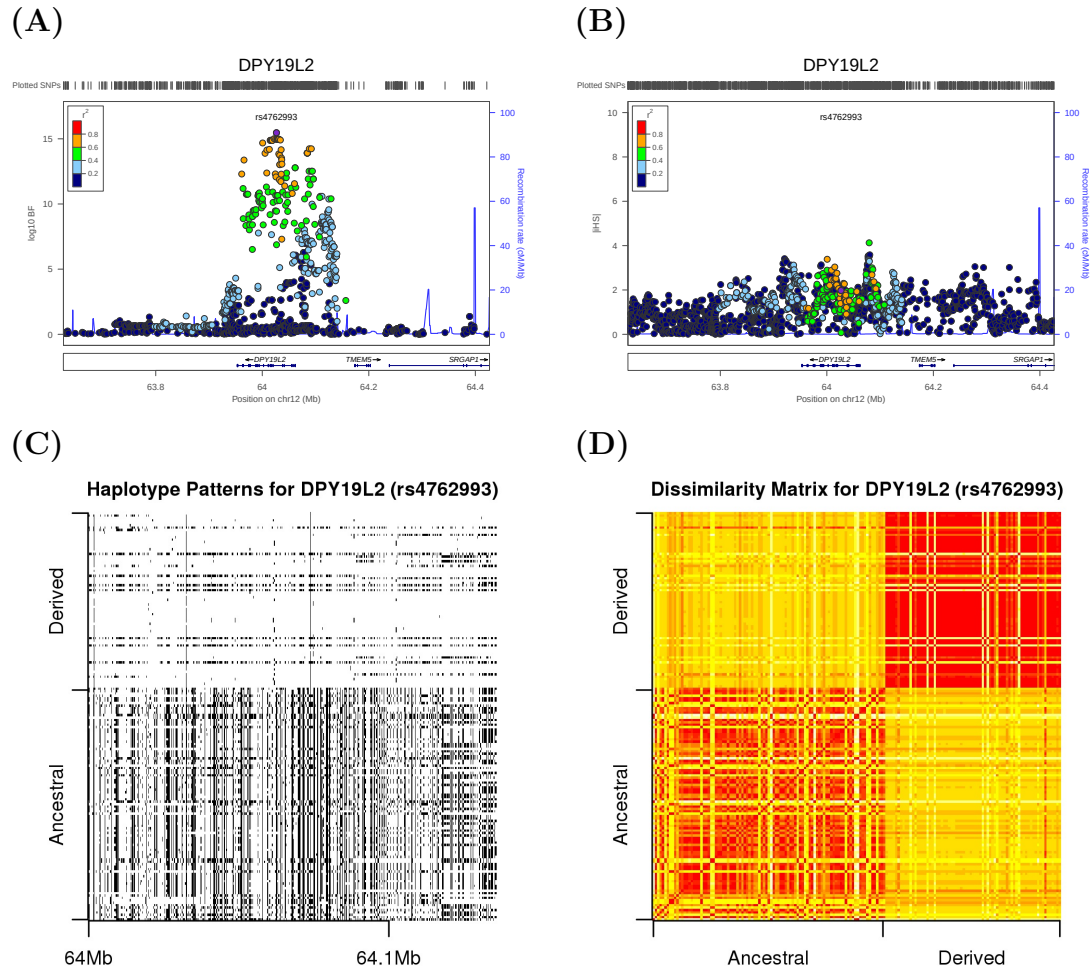


FIGURE 5.12: Comparison of CEU Candidates. (A) and (B) show the regions surrounding the DPY19L2 locus, centered on rs4762993, for ABACUS and iHS respectively. SNPs are coloured according to their R^2 value with the highlighted variant. Overlaid with the GRCh37 reference genome, and recombination map. Constructed using the LocusZoom web application [Pruim et al., 2010]. (C) Haplotype variation plot, with individuals as rows, and positions as columns. Coloured black or white depending on the allele that they carry. (D) Pairwise difference matrix between the haplotypes over this region. Dark red indicates low number of differences.

this regions suggests that a number of SNPs have alleles that are strongly associated with reduced time to common ancestor. The variation analysis in Figures 5.12(C) & 5.12(D), suggests that those haplotypes carrying the derived allele, have considerably reduced diversity compared to the ancestral haplotype.

Furthermore, the structure of the variation in Figures 5.12(C) & 5.12(D) suggests that there appears to be at least two haplotypes on the derived background. As

evidenced by the block-like pattern in the dissimilarity matrix between those haplotypes on the derived background. The diagnostic plot in Figure F.1, also shows that the painted segments extend up to 400Kb in either direction, and are significantly longer than the painted lengths of the ancestral allele, which is driving this large signal of selection.

The patterns of variation suggest that we cannot rule out the hypothesis that this region is currently experiencing, or has recently experienced, a soft selective sweep. iHS values in this region are also high ($|iHS| > 2$) however the signal is not as strong as those observed under ABACUS, and the highlighted SNP does not have the highest score in this region. As before, we calculated the empirical rank of the highlighted SNP as being 4.75×10^{-6} under ABACUS and 0.048 under iHS . Therefore, at the 1% false positive rate we would reject the null hypothesis under ABACUS and reject the alternative under iHS .

FADS The next region that we look at focuses on a 2Mb region surrounding a set of genes called *FADS1*, *FADS2*, *FADS3* in the CHB subpopulation. These genes are members of the fatty acid desaturase family which cluster together in this region, and variants have been associated with reduced desaturation of polyunsaturated fatty acids, e.g. *omega-3* [Steer et al., 2012]. It is thought that a reduction in *omega-3* concentrations can be associated with cardiovascular disease, type 2 diabetes, and also some developmental traits e.g. reduced IQ and visual development [Steer et al., 2012].

Figure 5.13 indicates that this region has a large number of SNPs with values of $\log_{10} BF > 2$. The variation plots in Figures 5.13(C) & 5.13(D), suggest that there is a significant reduction in variation on the derived allele, and possibly a single haplotype on the derived background. Figure F.2 suggest that the derived allele is significantly more associated with longer lengths under one allele. The

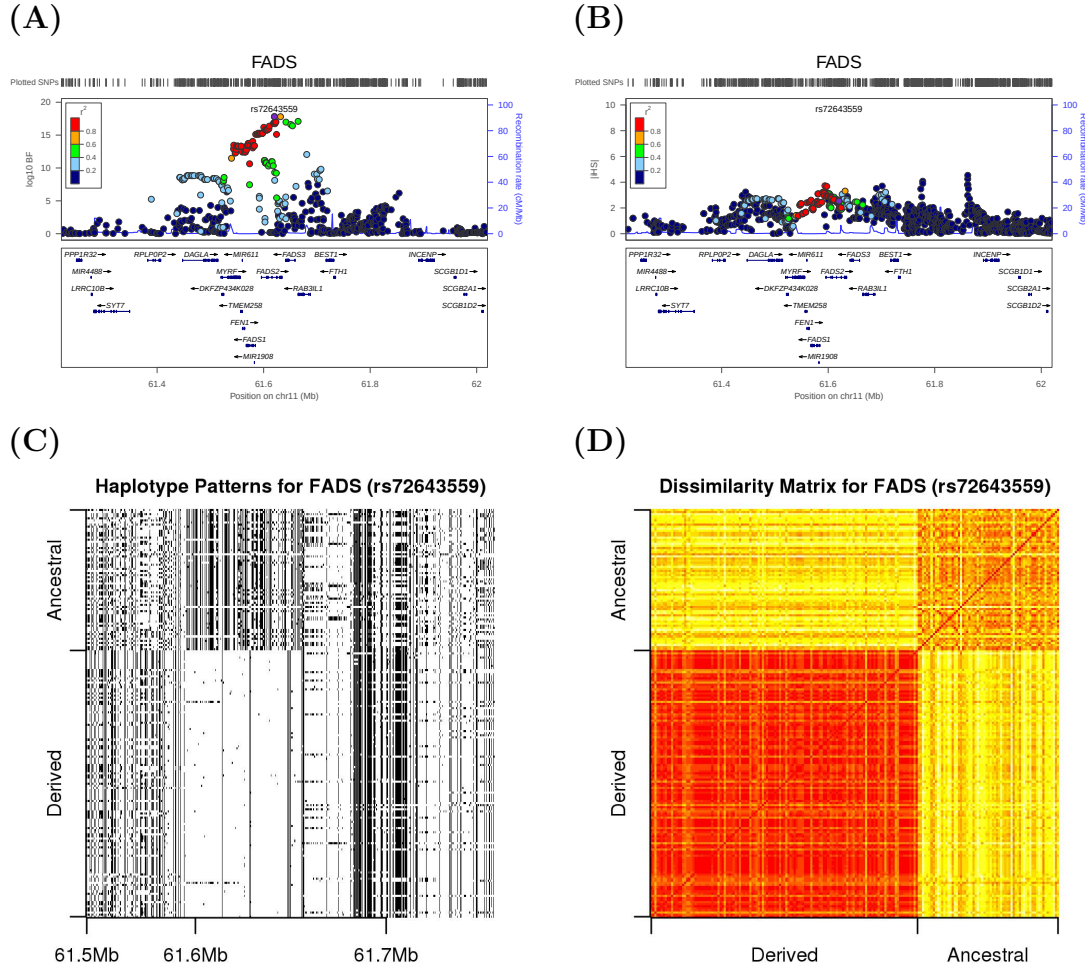


FIGURE 5.13: (A) and (B) show the regions surrounding the FADS locus, centered on rs72643559, for ABACUS and iHS respectively. SNPs are coloured according to their R^2 value with the highlighted variant. Overlaid with the GRCh37 reference genome, and recombination map. Constructed using the LocusZoom web application [Pruim et al., 2010]. (C) Haplotype variation plot, with individuals as rows, and positions as columns. Coloured black or white depending on the allele that they carry. (D) Pairwise difference matrix between the haplotypes over this region. Dark red indicates low number of differences.

painted segments extend up to 500Kb in each direction, even though this region has a number of recombination hotspots.

We also observe that there are a considerable number of SNPs at this locus with $|iHS| > 2$ adding to the evidence that this region has patterns of variation consistent with a selective sweep. Furthermore, this SNP has the highest empirical rank under ABACUS, and a rank of 0.018 under iHS. Considering the large single haplotype on the derived background it is surprising that we do not see a more extreme

value of iHS. One explanation may be the increased variation as a result of the multiple recombination hotspots, however as with ABACUS, iHS is designed to be robust to such features [Voight et al., 2006]. The fact that we observe such a large value of $\log_{10} \text{BF}$ at this locus, and considerably reduced variation on the derived background suggests that this region has experienced or is currently experiencing a classical selective sweep.

IGH The final region focuses on the immunoglobulin heavy locus (IGH) at the end of chromosome 14 in the CEU subpopulation. This locus has one of the largest values of $\log_{10} \text{BF}$ that is replicated across all the populations that we have looked at. The IGH locus encodes a large number of antibody molecules which are used by the immune system to protect an organism from infections [Tanaka and Nei, 1989]. By comparing synonymous to non-synonymous substitutions, which we discussed in Chapter 1, Tanaka and Nei [1989] showed that there was evidence of positive Darwinian selection.

Figures 5.14(A) & 5.14(B) show that there is a large number of SNPs that have values of $\log_{10} \text{BF} > 2$ and $|\text{iHS}| > 2$ in this locus. The variation plots in Figures 5.14(C) & 5.14(D) show that the region is experiencing strong recombination, as shown by the recombination maps in the top two panels. Despite the large recombination we observe that there is considerably reduced diversity on the derived allele. The empirical ranks for ABACUS and iHS at this SNP are 2.42×10^{-5} , and 0.0179 respectively. Again, this suggests that at the 1% false positive rate we would reject the null hypothesis under ABACUS.

From the patterns of variation it is difficult to identify whether this region has variation consistent with classical or soft sweep, however one might argue in Figure 5.14(C) that there are multiple haplotypes. Together with the block-like structure in the dissimilarity matrix, one might conclude that the region has experienced or is currently experiencing a soft selective sweep.

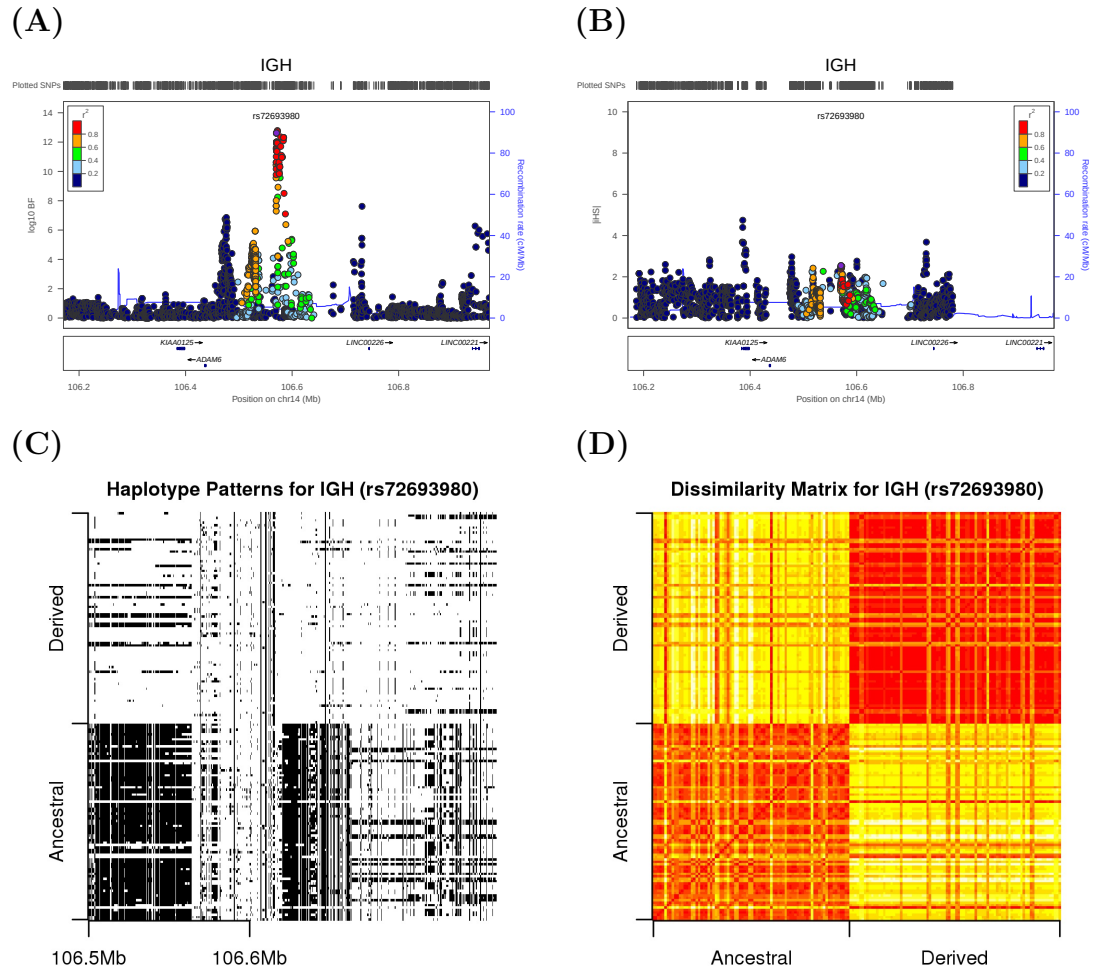


FIGURE 5.14: (A) and (B) show the regions surrounding the IGH locus, centered on rs72693980, for ABACUS and iHS respectively. SNPs are coloured according to their R^2 value with the highlighted variant. Overlaid with the GRCh37 reference genome, and recombination map. Constructed using the LocusZoom web application Pruim et al. [2010]. (C) Haplotype variation plot, with individuals as rows, and positions as columns. Coloured black or white depending on the allele that they carry. (D) Pairwise difference matrix between the haplotypes over this region. Dark red indicates low number of differences.

Due to the increased recombination, and polymorphisms in this region aligning the human sequence with other species is very challenging. As such it is often difficult to identify which of the alleles is derived and which is ancestral. One effect of this is that it is not possible to identify iHS values at these positions. This is one explanation as to why we see a significant number of missing SNPs in Figure F.3. This phenomenon is not unique to the *IGH* locus, indeed the *MHC* also suffers from a similar problem (data not shown).

5.5 Discussion

The goal for this final results chapter was to understand the challenges of applying ABACUS to real data and to attempt to quantify the differences between the results and previously studies, e.g. Voight et al. [2006] & Akey [2009]. ABACUS identified many previously identified regions of natural selection, e.g. *LCT*, and genes that are involved with immunity e.g. *MHC* [Borghans et al., 2004] and *IGH* [Tanaka and Nei, 1989].

In line with the results from Chapter 4, we observed that iHS had significantly better power to detect patterns of variation associated with selection, when the derived allele was at high frequency ($f > 80\%$). We observed that some of the strongest signals of selection under ABACUS included a number of regions where the patterns of variation might suggest evidence of selection by a soft sweep, e.g. *DPY19L2*. Moreover, in regions, e.g. *IGH*, where it is difficult to use ancestral genomes to identify which allele is derived, ABACUS will be able to calculate the evidence of selection, where iHS will not.

Finally, we observe across the genome that ABACUS resulted in good ($> 80\%$) overlap when compared with existing maps of positive selection in the human genome, e.g. Akey [2009], and we observe that there was a good degree (65%) of overlap between the results from iHS and ABACUS.

For this chapter we decided to focus on samples from just three of the fourteen populations from 1KG. We did this so that we could focus on explaining the differences between the methods in a number of regions, however we identified that the selection pressure may not be the same between populations with the same ancestry. As a result the conclusions that we made regarding the evolutionary development of some of the regions that were identified by Peter et al. [2012] may not be directly comparable.

Visually identifying whether a region highlighted by either method is evolving under a classical sweep, or a soft sweep of selection is very challenging. As we illustrated in Figure 1.11 different modes of selection can leave very distinct patterns of variation in the region, however without a statistical model like that described in Peter et al. [2012] it is not possible to assess the support of each model. Furthermore, the variability of the haplotypes between backgrounds is dependent on the recombination rate in the region, which will also affect the extent of the selective pressure.

Overall we are encouraged by the fact that we observe good agreement with previous studies and between iHS and ABACUS. Moreover results presented in this chapter together with the previous chapter suggest that ABACUS is a viable alternative to iHS when one wishes to quickly identify regions of a genome that exhibit patterns of variation associated with positive selection, either from a classical sweep or by a soft sweep.

Chapter 6

Conclusion and Future Direction

“If I have seen further it is by standing on the shoulders of giants.”

Sir Issac Newton

6.1 Conclusions

In this thesis, we set out to design a method that is capable of probabilistically identifying evidence of positive selection in genome wide variation data. Our aim was to construct the method so that it is sensitive to patterns of variation associated with classical selective sweeps and soft selective sweeps, at the same time as scaling to genome wide variation data.

6.1.1 Exact model, approximate inference

In Chapter 1 we highlighted a major challenge with using the likelihood of the coalescent to perform conditional inference. Since the space of trees, that could have generated a particular pattern of variation, is infinite it would be intractable

to apply this method directly to genome wide variation data. We decided to investigate the use of alternative inference frameworks that have been designed to work in intractable likelihood settings, e.g. Approximate Bayesian Computation.

We began by using machine learning tools together with coalescent simulation software to establish whether it was possible to infer the strength of selection, captured by the parameter σ , using a set of frequency spectrum statistics. These statistics were chosen based on their sensitivity to patterns of variation associated with a complete selective sweep. In Figure 2.6 we showed, even in the simple case of assuming everything except σ is known, that these machine learning methods resulted in highly variable estimates.

We then applied an Approximate Bayesian Computation method to establish whether the quality of the results improved. Using the same metrics as before together with the original rejection algorithm by Tavaré et al. [1997], we examined the results when we applied the method to four scenarios of differing strengths of selection, Figure 2.7. In general, we showed that despite the increased cost of computation the results were still highly variable, however in the case of strong selection $\sigma > 200$, and a derived allele frequency of 75%, we noted that the variability in the estimates did decrease considerably.

The results from our ABC analysis are consistent with those of other studies, e.g. Peter et al. [2012]. In their paper, the authors show that when selection is strong ($\sigma > 100$) and the beneficial allele is at a high frequency in the population ($f = 95\%$), their method is able to detect the strength of selection. Furthermore, when selection is very strong ($\sigma \geq 1000$) then their method is able to distinguish between patterns of variation associated with a classical selective sweep, and selection on standing variation.

One of the challenges that we identified with this approach is that it requires simulation from a model of evolution that is sufficiently complex so as to reduce

the effect of confounding variables, e.g. population structure, and varying population size. In real applications we will never know the true model and incorrect assumptions may introduce biases into the results. In general, we noted that these simulation based methods will not scale to genome wide variation for two reasons. First, it requires a choice of a window size and as we discussed in Chapter 2, changing the size can alter the conclusions. Second, in each window it will be necessary, as in Peter et al. [2012], to perform independent simulations as local recombination patterns will alter the genealogy of the sample. This will prove challenging if this method was applied to entire genome variation data.

6.1.2 Approximate model, exact inference

An alternative approach to using a coalescent model to perform inference is to use an approximate model that captures the main features of it. A method to do this, which has become popular for the analysis of genome wide variation data, was proposed by Li and Stephens [2003]. This model uses patterns of variation to relate haplotypes based on their most recent common ancestor, and has been used in a number of different applications in modern population genomics.

Using a version of this method, known as chromosome painting [Lawson et al., 2012], we demonstrated that it was possible to estimate the distance between recombination events in haplotypes for pairs of closely related haplotypes. Using a theoretical argument by Wiuf and Hein [1999] and simulated data, we showed that estimated distance is closely related, in expectation, to the time to common ancestor. In Figure 3.5 we showed that the longer the distance between estimates of the recombination distance, the shorter the time to common ancestor.

From the results from Chapter 2, we were aware of different sources of bias that might also impact the time to common ancestor, e.g. non-constant population size, and population structure. In expectation these sources will affect the whole

genome equally. Therefore we decided to use the fact that we have genome wide variation data to reduce their impact. In practice, we perform a relative comparison between haplotypes based on the estimated average distance between recombination events across the whole genome. Using simulations, and theoretical arguments we show that this step stabilises the false positive rate of the method across different demographic scenarios and population structures. Importantly though this step does not detrimentally impact the method's sensitivity to patterns of variation associated with positive selection.

One of the challenges of using this model to infer ancestral relationships is that it bases its inference on the tips of the tree [Li and Stephens, 2003]. As a result, inference on selection will be based on its effect on the recent parts of the tree. Whilst this is sufficient for recent selective pressures, older pressures will be much more difficult to detect.

This argument is linked to the observation, in Figure 4.14, that the power of the method to detect selection does not increase monotonically with sample size. We argued that this result can be explained by looking at the effect of sample size on the underlying ancestry. As sample sizes increase the times in the underlying ancestry will become shorter, increasing the distance between any two recombination events. As we showed at the end of Chapter 4, the effect of this will be to reduce the effect of the selective pressure to differentiate the recombination distances. As result, using current estimates of the distance between recombination events to infer evidence of selection in very large samples ($n > 1000$) will become more challenging.

6.1.3 Power to Detect Selection

In order to assess the evidence that a region of the genome exhibits patterns of variation consistent with partial selective sweep, we proposed a novel Bayesian

test of association. At each SNP we calculate the posterior probability that the estimated distance between recombination distances is longer than expected across the genome. We use this probability to label estimates of recombination distance as *long* relative to the genome wide average. Then using a logistic model we marginalise out the uncertainty in the labels to establish whether one allele is more associated with longer lengths compared with the alternative allele. Note, unlike existing methods e.g. iHS [Voight et al., 2006] we make no distinction between ancestral and derived alleles. This makes the method more applicable to regions of the genome where it is not possible to align the DNA with its ancestral ortholog, i.e. highly recombining or highly polymorphic regions.

We compared the power of our method, ABACUS, to detect both classical selective sweeps and soft selective sweeps relative to two alternative methods; iHS [Voight et al., 2006] and H_{12} [Garud et al., 2013]. Applying ABACUS across a range of simulated scenarios, we showed, under classical selective sweeps, that it has similar power as iHS and H_{12} to detect strong positive selection ($\sigma > 250$), Figure 4.8. Under soft sweeps by recurrent mutation, and soft sweeps by selection on standing variation, we showed that it had considerably better power than either of the two alternative methods, Figures 4.10 & 4.12.

One explanation, as to why ABACUS has similar power to detect classical and soft selective sweeps, is that it identifies regions of the genome where one allele is associated with longer distances between recombination events compared with the alternative allele. Controlling for the local recombination rates we expect that longer distances between recombination events are an effect of reduced time to common ancestor. Therefore if all haplotypes carrying a particular allele have considerably reduced time to common ancestor relative to the alternative allele, we argue that this is evidence of positive selection.

Since the derived allele is necessarily the younger allele, those haplotypes that

carry it will have relatively lower time to common ancestor than those haplotypes that do not. The effect of this will be to increase the distance between recombination events on the derived background. This will necessarily increase the association test at lower derived frequencies. The effect of this phenomenon, however, was actually quite small. As, in practice, we observe that at very low, and very high frequencies the test of association has reduced power due the low number of observations in one of the comparison groups. This is a feature of inference methods that are analogous to contingency tables. The fact that there are fewer observations in one of the comparison groups reduces the power to detect association. If one wished to achieve stability in distributions across the frequency spectrum one might wish to perform a standardisation not unlike that which is performed by Voight et al. [2006], however this will diminish the interpretability of the Bayes factor. A more useful alternative is to calculate empirical p-values using the distribution of $\log_{10} \text{BF}$.

Genome wide scans for selection, like ABACUS, rely on the assumption that selection is the only evolutionary pressure that might reduce the time to common ancestor among a subsample of chromosomes. Whilst true in expectation, other features, such as demography, act stochastically along the genome to alter the ancestry. We believe that ABACUS will be robust to these sources because we make a relative comparison between haplotypes at a single site and account for the average ancestry across the genome.

6.1.4 New candidates

Following on from the development of ABACUS and the demonstration of its power to detect both classical selective sweeps and soft selective sweeps, we focused in the final chapter on the applicability of the model to genome wide sequence data. Using samples from three populations (CEU, YRI, CHB) from 1000 Genomes

Project Consortium et al. [2012] we applied ABACUS and a scalable version of iHS. Using previously identified candidates of either classical selective sweeps or soft selective sweeps from Peter et al. [2012] we looked at the differences in the results from the two methods.

In line with results from simulations in Chapter 4, under classical sweeps, we showed that iHS has significantly better power when the beneficial allele has reached a very high frequency in the population, Figure 4.8. In a number of regions that exhibit patterns of variation associated with classical or soft sweeps of selection, at intermediate frequencies, we highlighted the fact that ABACUS identifies the causal variant as having the largest $\log_{10}$ BF in the region.

In Figure 5.11 we observed that there was little correlation between the SNP by SNP values of $\log_{10}$ BF and $|iHS|$, however when we looked at regions identified by either method we observed that there was considerable overlap (58.9%). By comparing the results of an integrated map of positive selection in the human genome [Akey, 2009] to the results from ABACUS we observed that 88.6% of the regions overlap. Some of the differences in the SNP level relationship can be explained by the fact that ABACUS is more sensitive to patterns of variation associated with soft selective sweeps, and iHS is more powerful to patterns of variation consistent with positive selection when the minor allele frequency is $< 10\%$.

Building on this we propose three novel candidates of positive selection in the human genome; *DPY19L2*, *FADS*, and *IGH*. All three regions have some of the largest $\log_{10}$ BF under ABACUS, and patterns of variation consistent with either a classical selective sweep, or in the case of *DPY19L2* and *IGH* patterns of variation associated with a soft selective sweep. We observed in Figures 5.12 to 5.14 that iHS also has a considerable number of SNPs in this regions with $|iHS| > 2$, however they had a much lower rank. Furthermore, in the *IGH* locus, we observed that

many SNPs had no iHS value because the method was not able to establish which allele was ancestral.

We mentioned at the end of Chapter 5 that visually identifying the mode of selection from the variation data is fraught with challenges, and it would be unwise to relying on these conclusions without quantifying the evidence; e.g Peter et al. [2012]. Moreover, it is difficult to make conclusions about whether iHS is able to identify a causal variant because stochastic variation caused by the standardisation of the metric within frequency windows.

The results from our analysis of the 1000 Genomes data together with the power estimates suggest that ABACUS is capable of detecting patterns of variation associated with both classical and soft positive selective sweeps. In the case of soft sweeps the results suggest that it is more powerful than either of the two alternative methods that we considered. Furthermore, as we have shown, the inference framework and implementation of ABACUS means that it is readily applicable to genome wide sequence variation data.

6.2 Future work

An aim for this thesis is to establish a new method to scan genome wide variation data for evidence of positive selection. The model that we designed is based on a method that relates patterns of observed variation to recent ancestry. This is just one possible approach, and over the course of this research we explored a number of alternative methods.

6.2.1 Alternative inference methods

The Viterbi algorithm is a very quick and efficient way of making inference in HMMs, however it only produces a point estimate summary, i.e. a single path through the state space, with no additional information about the variation. At the end of Chapter 3, we proposed that an alternative approach would be to use posterior decoding, to calculate, at each position, the posterior probability of any chromosome being most closely related to all other chromosomes. Using this approach one could sample paths through the HMM and calculate the Bayesian test of association for each sample, and take an average of the results at each SNP. This approach has merits, as one is able to establish the variation in sampled paths, however we abandoned it because we could not establish whether it was appropriate, or more powerful, to combine Bayes factors across realisations of the HMM.

Both Viterbi decoding, and path sampling, produce a single path through the state space of the HMM. We explored, an alternative approach that removed the need to summarise the HMM using paths. At each position we use the posterior decoding probabilities to calculate for each pair, in expectation, the distance between consecutive recombination events. At each position this produces a matrix of the expected distance (Equation 3.11) between recombination events for each pair of haplotypes. In simulated data this, visually, produced a very accurate painting of the selected haplotypes. We abandoned this method because, unlike previous models, this method no longer has linear complexity in the number of positions. Thus applying this method to genome wide variation data would not be possible.

An aim throughout this thesis is to make inference using the ancestral history relating a sample of haplotypes. Since inferring the history itself is challenging, we used alternative methods that capture its underlying features. We chose the

method by Li and Stephens [2003], because it uses information regarding the relationship between closely related haplotypes to make inference on recent ancestry, however more ancient events are beyond its capability. To improve the accuracy of our method we explored a number of alternative models that should allow one to make inference on more ancient ancestries.

In particular, we looked at an HMM based method by Zhang [2013] which was originally designed to simultaneously genotype and phase haplotypes using the raw output from sequencing machines. It infers a set of ancestral haplotypes from which it is possible to infer patterns of relatedness in a similar manner to the LSHMM. Our investigations suggest that this could be a useful approach, however whilst this method is very scalable, it was shown by Zhang [2013] to get stuck in local modes. The authors propose using a MCMC jump to resolve this problem, however we found in practice that it still produced unstable conclusions.

6.2.2 New applications

The main advantage of using an HMM based model to perform inference is that it will scale to very large datasets. This is especially important as genome wide sequencing technologies become more common in population genetics. The benefit of this new data is that the causal or beneficial SNP will always be typed. Therefore methods that rely on bi-partitions of alleles to perform inference, e.g. ABACUS should have a better chance of detecting the causal variant.

Due to data availability it was only possible to present the results from applying ABACUS to genome wide sequence data from 1000 Genomes Project Consortium et al. [2012]. Over the past few months we have applied the method to sequence data from *Mus musculus castaneus*, and *Mus musculus domesticus*.

In the next application we will be focusing on genome wide variation data from 10,000 individuals from eight African populations. Ultimately, however, our goal is to apply ABACUS to genome wide variation data from *Plasmodium falciparum*, however the patterns of recombination may mean that the method will need to be adapted slightly to account for these differences. Despite these challenges, ABACUS has a distinct overall existing methods to detect patterns of variation associated with selection, in that it has good power to detect both selection on standing variation and classical selective sweeps. This is particularly useful in drug resistance applications where environment pressures may mean that beneficial mutations rise to high frequency in the population from variation that is already present in the population, i.e. selection on standing variation. Furthermore the Bayesian design of the method means that if samples were collected, and sequenced, regularly from populations we might be able to use ABACUS to leverage this temporal information to perform drug resistance surveillance at a population level.

6.2.3 A final word

ABACUS is designed to be a fast, efficient and robust method to detect patterns of variation associated with positive selection in genome wide genetic variation datasets. To our knowledge it is the first scalable model that has to similar power to detect patterns of variation associated with both incomplete classical selective sweeps and incomplete soft selective sweeps. The fact that ABACUS performs this inference in a model based way makes it more robust to confounding effects such as demographic history and population structure.

Recall, one of the motivations for doing this research was to construct a method that would be capable of detecting and monitoring drug resistant traits in pathogens. ABACUS represents a first step in constructing such a method, however there is

still a considerable amount of work still left to do. Ultimately one of the key features of this model should be that it is able to combine information from multiple sources, e.g. clinical records and genetic variation data. By constructing a unified model we believe that it will help to reduce the impact of drug resistant pathogens.

Appendix A

Laplace Approximation Equations

To calculate the Laplace Approximation to the posterior is necessary to determine the score vector and information matrix under each hypothesis. To do that in this application we need to establish the derivatives of $p(\theta_i)$ in Equation 3.25.

$$\frac{\delta}{\delta\mu}p(\theta_i) = \frac{\exp(\mu + \beta\theta_i)}{(1 + \exp(\mu + \beta\theta_i))^2} \quad (\text{A.1})$$

$$\frac{\delta}{\delta\beta}p(\theta_i) = \begin{cases} \frac{\exp(\mu+\beta)}{(1+\exp(\mu+\beta))^2} & \theta_i = 1 \\ 0 & \theta_i = 0 \end{cases} \quad (\text{A.2})$$

$$\frac{\delta}{\delta\mu\delta\beta}p(\theta_i) = \begin{cases} \frac{\exp(\mu+\beta)(1-\exp(\mu+\beta))}{(1+\exp(\mu+\beta))^3} & \theta_i = 1 \\ 0 & \theta_i = 0 \end{cases} \quad (\text{A.3})$$

$$\frac{\delta^2}{\delta\mu^2}p(\theta_i) = \frac{\exp(\mu + \beta\theta_i)(1 - \exp(\mu + \beta\theta_i))}{(1 + \exp(\mu + \beta\theta_i))^3} \quad (\text{A.4})$$

$$\frac{\delta^2}{\delta\beta^2}p(\theta_i) = \begin{cases} \frac{\exp(\mu+\beta)(1-\exp(\mu+\beta))}{(1+\exp(\mu+\beta))^3} & \theta_i = 1 \\ 0 & \theta_i = 0 \end{cases} \quad (\text{A.5})$$

We following the argument in Marchini et al. [2007] and set the priors for μ and β to be Gaussian with mean 0 and variance σ_μ^2 and σ_β^2 , respectively. Defining

$l(\mu, \beta) = \log(L(\mu, \beta|\Pi, G))$ we can express the posterior score function under H_1 as,

$$U_{H_1}(\mu, \beta) = \left(\frac{\delta}{\delta\mu} l(\mu, \beta) + \frac{\delta}{\delta\mu} \log \mathbb{P}(\mu|H_1), \right. \\ \left. \frac{\delta}{\delta\beta} l(\mu, \beta) + \frac{\delta}{\delta\beta} \log \mathbb{P}(\beta|H_1) \right) \quad (\text{A.6})$$

$$= \left(\sum_{i=1}^n \frac{\delta}{\delta\mu} \log(f_i) + \frac{\delta}{\delta\mu} \log \mathbb{P}(\mu|H_1), \right. \\ \left. \sum_{i=1}^n \frac{\delta}{\delta\beta} \log(f_i) + \frac{\delta}{\delta\beta} \log \mathbb{P}(\beta|H_1) \right) \quad (\text{A.7})$$

$$= \left(\sum_{i=1}^n \frac{\frac{\delta f_i}{\delta\mu}}{f_i} - \frac{\mu}{\sigma_\mu^2}, \sum_{i=1}^n \frac{\frac{\delta f_i}{\delta\beta}}{f_i} - \frac{\beta}{\sigma_\beta^2} \right). \quad (\text{A.8})$$

where

$$\frac{\delta f_i}{\delta\mu} = \begin{cases} \frac{\delta}{\delta\mu} p(1) \pi_i + \frac{\delta}{\delta\mu} p(0) (1 - \pi_i) & g_i = 1 \\ -\frac{\delta}{\delta\mu} p(1) \pi_i - \frac{\delta}{\delta\mu} p(0) (1 - \pi_i) & g_i = 0 \end{cases} \quad (\text{A.9})$$

$$\frac{\delta f_i}{\delta\beta} = \begin{cases} \frac{\delta}{\delta\beta} p(1) \pi_i & g_i = 1 \\ -\frac{\delta}{\delta\beta} p(1) \pi_i & g_i = 0 \end{cases} \quad (\text{A.10})$$

Using a similar argument we can express the posterior information matrix as follows,

$$I_{H_1}(\mu, \beta) = \begin{pmatrix} \frac{\delta^2}{\delta\mu^2} l(\mu, \beta) & \frac{\delta^2}{\delta\mu\beta} l(\mu, \beta) \\ \frac{\delta^2}{\delta\mu\beta} l(\mu, \beta) & \frac{\delta^2}{\delta\beta^2} l(\mu, \beta) \end{pmatrix} + \begin{pmatrix} \frac{\delta^2}{\delta\mu^2} \log \mathbb{P}(\mu|H_1) & 0 \\ 0 & \frac{\delta^2}{\delta\beta^2} \log \mathbb{P}(\beta|H_1) \end{pmatrix} \quad (\text{A.11})$$

$$= \sum_{i=1}^n \begin{pmatrix} \frac{\delta^2}{\delta\mu^2} \log(f_i) & \frac{\delta^2}{\delta\mu\beta} \log(f_i) \\ \frac{\delta^2}{\delta\mu\beta} \log(f_i) & \frac{\delta^2}{\delta\beta^2} \log(f_i) \end{pmatrix} + \begin{pmatrix} -\frac{1}{\sigma_\mu^2} & 0 \\ 0 & -\frac{1}{\sigma_\beta^2} \end{pmatrix} \quad (\text{A.12})$$

$$= \sum_{i=1}^n \begin{pmatrix} \frac{\frac{\delta^2 f_i}{\delta\mu^2} - \left(\frac{\delta f_i}{\delta\mu}\right)^2}{f_i^2} & \frac{\frac{\delta^2 f_i}{\delta\mu\beta} - \frac{\delta f_i}{\delta\mu} \frac{\delta f_i}{\delta\beta}}{f_i^2} \\ \frac{\frac{\delta^2 f_i}{\delta\mu\beta} - \frac{\delta f_i}{\delta\mu} \frac{\delta f_i}{\delta\beta}}{f_i^2} & \frac{\frac{\delta^2 f_i}{\delta\beta^2} - \left(\frac{\delta f_i}{\delta\beta}\right)^2}{f_i^2} \end{pmatrix} + \begin{pmatrix} -\frac{1}{\sigma_\mu^2} & 0 \\ 0 & -\frac{1}{\sigma_\beta^2} \end{pmatrix} \quad (\text{A.13})$$

$$(\text{A.14})$$

where

$$\frac{\delta^2 f_i}{\delta\mu^2} = \begin{cases} \frac{\delta^2}{\delta\mu^2} p(1) \pi_i + \frac{\delta^2}{\delta\mu^2} p(0) (1 - \pi_i) & g_i = 1 \\ -\frac{\delta^2}{\delta\mu^2} p(1) \pi_i - \frac{\delta^2}{\delta\mu^2} p(0) (1 - \pi_i) & g_i = 0 \end{cases} \quad (\text{A.15})$$

$$\frac{\delta^2 f_i}{\delta\beta^2} = \begin{cases} \frac{\delta^2}{\delta\beta^2} p(1) \pi_i & g_i = 1 \\ -\frac{\delta^2}{\delta\beta^2} p(1) \pi_i & g_i = 0 \end{cases} \quad (\text{A.16})$$

$$\frac{\delta^2 f_i}{\delta\mu\delta\beta} = \begin{cases} \frac{\delta^2}{\delta\mu\delta\beta} p(1) \pi_i & g_i = 1 \\ -\frac{\delta^2}{\delta\mu\delta\beta} p(1) \pi_i & g_i = 0 \end{cases} \quad (\text{A.17})$$

Appendix B

Simulations

B.1 Data Generation

In all these simulations, population frequency f , and selection coefficient, σ , unless otherwise stated can take the following values.

$$f = \{5\%, 10\%, 20\%, 40\%, 50\%, 60\%, 80\%, 90\%, 95\%\}$$

$$\sigma = \{10, 50, 100, 250, 500, 1000\}$$

B.1.1 Classical Sweep

B.1.1.1 Null

We simulated 1,000 regions in each of the nine frequencies. *msms* simulates a trajectory conditional on the frequency of the allele in the current day population.

We set $\sigma = 0.1$, $\theta_A = 0$, $\mu = 1.2 \times 10^{-8}$ per generation per base pair, $r = 1 \times 10^{-8}$ per generation per base pair, $l = 2 \times 10^6$ base pair, and $N_e = 10,000$.

```
msms -ms 200 1000 -t 960.0 -r 800.0 2000000 -SAA 0.1 -SaA 0.05
```

```
-SF 0.0 frequency -N 10000 -Sp 0.5 -Smark
```

B.1.1.2 Alternative

We simulated 250 regions at each intersection of the nine frequencies and six σ . *msms* simulates a trajectory conditional on the frequency of the allele in the current day population. We set $\theta_A = 0$, $\mu = 1.2 \times 10^{-8}$ per generation per base pair, $r = 1 \times 10^{-8}$ per generation per base pair, $l = 2 \times 10^6$ base pair, and $N_e = 10,000$.

```
msms -ms 200 250 -t 960.0 -r 800.0 2000000 -SAA sigma -SaA sigma/2
-SF 0.0 frequency -N 10000 -Sp 0.5 -Smark
```

B.1.2 Soft Sweep

B.1.2.1 Null

We simulated 1,000 regions in each of the nine frequencies. *msms* simulates a trajectory conditional on the frequency of the allele in the current day population. We set $\sigma = 0.1$, $\theta_A = 10$, $\mu = 1.2 \times 10^{-8}$ per generation per base pair, $r = 1 \times 10^{-8}$ per generation per base pair, $l = 2 \times 10^6$ base pair, and $N_e = 10,000$.

```
msms -ms 200 1000 -t 960.0 -r 800.0 2000000 -SAA 0.1 -SaA 0.05
-Smu 10 -SF 0.0 frequency -N 10000 -Sp 0.5 -Smark
```

B.1.2.2 Alternative

We simulated 250 regions at each intersection of the nine frequencies and six σ . *msms* simulates a trajectory conditional on the frequency of the allele in the current day population. We set $\theta_A = 10$, $\mu = 1.2 \times 10^{-8}$ per generation per

base pair, $r = 1 \times 10^{-8}$ per generation per base pair, $l = 2 \times 10^6$ base pair, and $N_e = 10,000$.

```
msms -ms 200 250 -t 960.0 -r 800.0 2000000 -SAA sigma -SaA sigma/2
-Smu 10 -SF 0.0 frequency -N 10000 -Sp 0.5 -Smark
```

B.2 Calculating ABACUS

We perform the painting as described in Chapter 4 setting input effective population size parameter $N_e = 10,000$. We then average the total copying length across all pairs and repetitions under the null, to estimate the average pairwise copying length.

We then calculate ABACUS as described in Chapter 3, but setting the estimate of the copying lengths across the genome for all pairs equal to that which we calculated under the null. We then take the $\log_{10}$ BF value at the selected site.

B.3 Calculating uIHS

At the target locus we calculate the integrated decay of on the ancestral and derived backgrounds; iHH_A and iHH_D , respectively. We then calculate $uIHS = \log_e iHH_A - \log_e iHH_D$.

B.4 Calculating H_{12}

Following the derivation in Garud et al. [2013] we calculate the H_{12} statistic as follows. At the target locus we take a 401-SNP window it and calculate the frequency of each haplotype. We rank each haplotype, in descending order according

to its frequency; i.e. p_i is the i^{th} most frequent haplotype. Then

$$H_{12} = (p_1 + p_2)^2 + \sum_{i>2} p_i^2. \quad (\text{B.1})$$

Appendix C

ABACUS Implementation

To maximise the flexibility and scalability of the ABACUS model we implemented it using a combination of Linux *bash* and concurrent *Java*. In its current form the model reads the variation files into memory and then calculates the Viterbi paths, and Bayes factors in parallel. One of the challenges with our implementation however is that the Viterbi algorithm requires a significant amount of memory. This can be a problem when the number of samples is $> 1 \times 10^3$. One solution that we will discuss in the next chapter is using a Map-Reduce [Dean and Ghemawat, 2008] design, or exploiting the Viterbi algorithm to split the genome up and reconstruct the Viterbi paths. For our purposes however, the current implementation is quite sufficient.

A schematic of the current implementation of ABACUS has been produced in Figure C.1. As we can see the implementation takes a phased haplotype file in *.gen* format.

```
rs58108140 rs58108140 10583 G A 0 1 0 ...  
rs180734498 rs180734498 13302 C T 0 1 0 ...
```

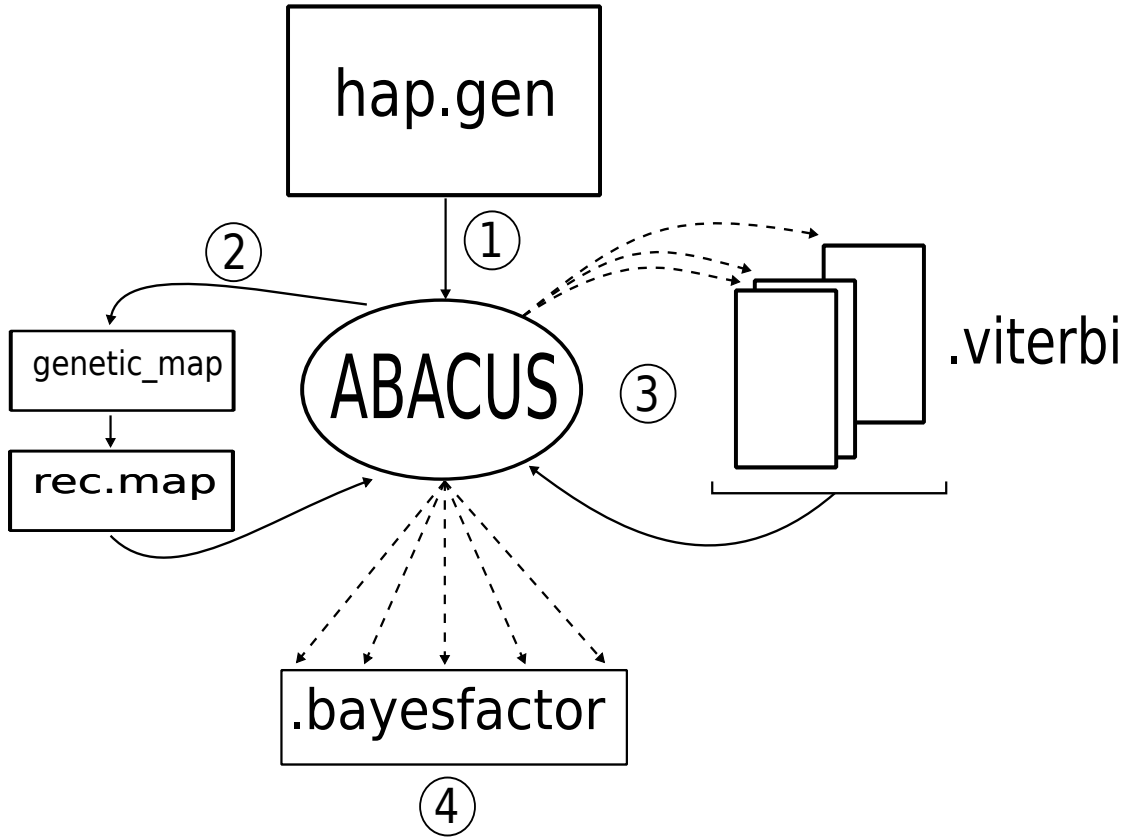


FIGURE C.1: Schematic detailing the ABACUS pipeline. (1) Read in haplotype file in .gen format. (2) Use position information and genetic map to produce linear interpolated map in centi-Morgans for every position in .gen file. (3) Parallel run the Viterbi to produce one .viterbi file for every haplotype. (4) Read Viterbi information into ABACUS and parallel create .bayesfactor file. Parallel process are denoted by dashed lines.

Each space separated line represents a SNP; the first two columns uniquely identify the SNP via the *RSID*, the third column denotes the position in the chromosome, the fourth and fifth columns denote the nucleotides for the 0 and 1 alleles respectively, and the last columns denote the alleles for each haplotype.

This haplotype file is read into ABACUS and using the recombination map for the species a linear interpolated map is produced for each position in the .gen file. This is then used to produce the .viterbi files (one for each haplotype) which are then used to produce the .bayesfactor file. As we mentioned above the last two steps are implemented in a parallel which means that the application can take advantage of multicore compute nodes.

Appendix D

ABACUS and iHS Results Tables

Pop.	Chr.	Position (Mb)	Genes	$\log_{10}\text{BF}$	iHS	Replicated
CEU	6	29.56 - 31.69	MHC	17.52	NO	CHB + YRI
CEU	1	248.3 - 248.6	OR14, OR2	16.48	NO	CHB + YRI
CEU	10	134.75 - 135.2	GPR123	16.42	YES	CHB
CEU	7	6.8 - 7.16	CCZ1B	16.21	NO	-
CEU	12	63.88 - 64.21	DPY19L2	15.47	NO	CHB
CEU	2	135.11 - 137.13	MCM6, LCT	14.76	YES	-
CEU	4	78.11 - 78.32	CCNG2	14.59	YES	CHB + YRI
CEU	3	97.44 - 98.01	OR5, EPHA6, GABRR3, CRYBG3	13.38	NO	CHB + YRI
CEU	9	89.14 - 89.29	ZCCHC6	13.19	NO	CHB
CEU	14	106.41 - 106.78	IGH	12.79	NO	CHB + YRI
YRI	2	89.2 - 89.6	IGK	12.43	NO	CEU + CHB
YRI	19	21.6 - 21.89	ZNF100, LINC00664	12.01	NO	-
YRI	6	29.65 - 32.83	MHC	10.78	NO	CEU + CHB
YRI	2	112.03 - 112.41	LOC541471, MIR4435	10.70	YES	-
YRI	14	106.14 - 106.33	IGH	10.14	NO	CEU + CHB
YRI	8	8.02 - 8.14	FAM86B3P	9.29	NO	-
YRI	7	149.67 - 150.17	ACTR3C	8.68	YES	CHB
YRI	16	29.1 - 29.37	RRN3P2, SNX29P2	8.33	YES	CEU + CHB
YRI	17	80.23 - 80.41	UTS2R, OGFOD3	8.33	NO	CHB
YRI	4	168.78 - 168.9	ANXA10	8.32	YES	-
CHB	11	61.24 - 61.91	FADS	17.81	YES	CEU + YRI
CHB	1	22.19 - 22.4	CELA3B	17.55	YES	CEU + YRI
CHB	2	56.57 - 57.64	CCDC85A	16.89	NO	CEU
CHB	10	134.73 - 135.08	GPR123	16.72	YES	CEU
CHB	8	127.1 - 127.54	LINC00861, FAM84B	15.94	YES	-
CHB	16	29.08 - 29.37	RRN3P2, SNX29P2	15.89	YES	CEU + YRI
CHB	1	1.84 - 1.97	KIAA1751	15.47	YES	CEU
CHB	11	11.12 - 11.42	CSNK2A3, GALNT18	14.74	YES	CEU + YRI
CHB	6	30.9 - 31.66	MHC	14.34	NO	CEU + YRI
CHB	4	96.48 - 96.71	UNC5C, PDHA2	14.16	YES	CEU

TABLE D.1: Shows the top 10 regions in each population. Regions were determined by the distance from the maximum that $\log_{10}\text{BF}$ fell below 3. Genes that are in boldface overlap the region and genes that aren't are the closest to the tagged region. The iHS column indicates whether a similar signal is observed in the iHS analysis. The Replicated column indicates whether we observe a similar BF signal in other populations.

Pop.	Chr.	Position (Mb)	Genes	iHS	BF	Replicated
CEU	10	67.62 - 68.8	CTNNA3	13.48	YES	CHB + YRI
CEU	2	12.93 - 14.34	LINC00276	12.80	YES	CHB + YRI
CEU	13	107.33 - 107.8	LINC00443	11.63	YES	CHB + YRI
CEU	8	33.8 - 34.78	UNC5D	11.50	YES	YRI
CEU	10	59.81 - 60.39	IPMK	11.35	NO	CHB + YRI
CEU	13	58.82 - 59.57	DIAPH3	11.28	YES	-
CEU	15	74.05 - 75.05	ARID3B	10.92	YES	CHB + YRI
CEU	4	59.95 - 60.4	LPHN3	10.60	YES	CHB + YRI
CEU	1	34.2 - 34.4	CSMD2	10.57	YES	CHB + YRI
CEU	6	77.86 - 78.68	HTR1B	10.44	NO	CHB + YRI
YRI	18	62.23 - 62.43	LOC284294, CDH7	11.73	NO	CEU + CHB
YRI	9	30.95 - 31.28	LOC401497	10.36	YES	-
YRI	3	98.7 - 98.81	MIR548G, DCBLD2	9.50	NO	-
YRI	8	34.6 - 34.88	UNC5D	9.49	YES	CEU
YRI	10	67.93 - 68.29	CTNNA3	9.35	YES	CEU + CHB
YRI	11	86.74 - 88.06	TMEM135	9.30	YES	CHB
YRI	19	29.77 - 29.89	LOC284395	9.27	YES	CEU + CHB
YRI	2	81.61 - 81.72	CTNNA2, LOC1720	9.18	YES	-
YRI	20	24.83 - 24.94	SYNDIG1, CST7	8.64	YES	CEU, CHB
YRI	18	12.03 - 12.32	C18orf61	8.54	NO	CEU, CHB
CHB	4	65.26 - 65.63	TECRL, LOC401134	12.59	YES	CEU + YRI
CHB	10	25.42 - 25.82	GPR158	12.51	YES	CEU + YRI
CHB	4	70.13 - 70.65	UGT2B4, UGT2A1	12.34	NO	CEU + YRI
CHB	4	120.15 - 120.95	FLJ14186	12.21	NO	CEU + YRI
CHB	12	58.29 - 58.56	XRCC6BP1, LRIG3	12.19	YES	CEU + YRI
CHB	11	86.85 - 87.1	TMEM135	11.53	YES	YRI
CHB	16	48.42 - 49.05	N4BP1, CBLN1	11.24	NO	YRI
CHB	2	75.95 - 76.68	MRPL19, LRRTM4	11.21	NO	CEU + YRI
CHB	4	59.93 - 60.4	IGFBP7, LPHN3	10.99	YES	CEU + YRI
CHB	15	52.67 - 53.77	ONECUT1, WDR72	10.84	YES	YRI

TABLE D.2: Shows the top 10 regions in each population. Regions were determined by the distance from the maximum that iHS fell below 2. Genes that are in boldface overlap the region and genes that aren't are the closest to the tagged region. The BF column indicates whether a similar signal is observed in the ABACUS analysis. The Replicated column indicates whether we observe a similar iHS signal in other populations.

Appendix E

Existing Candidate Region Analysis

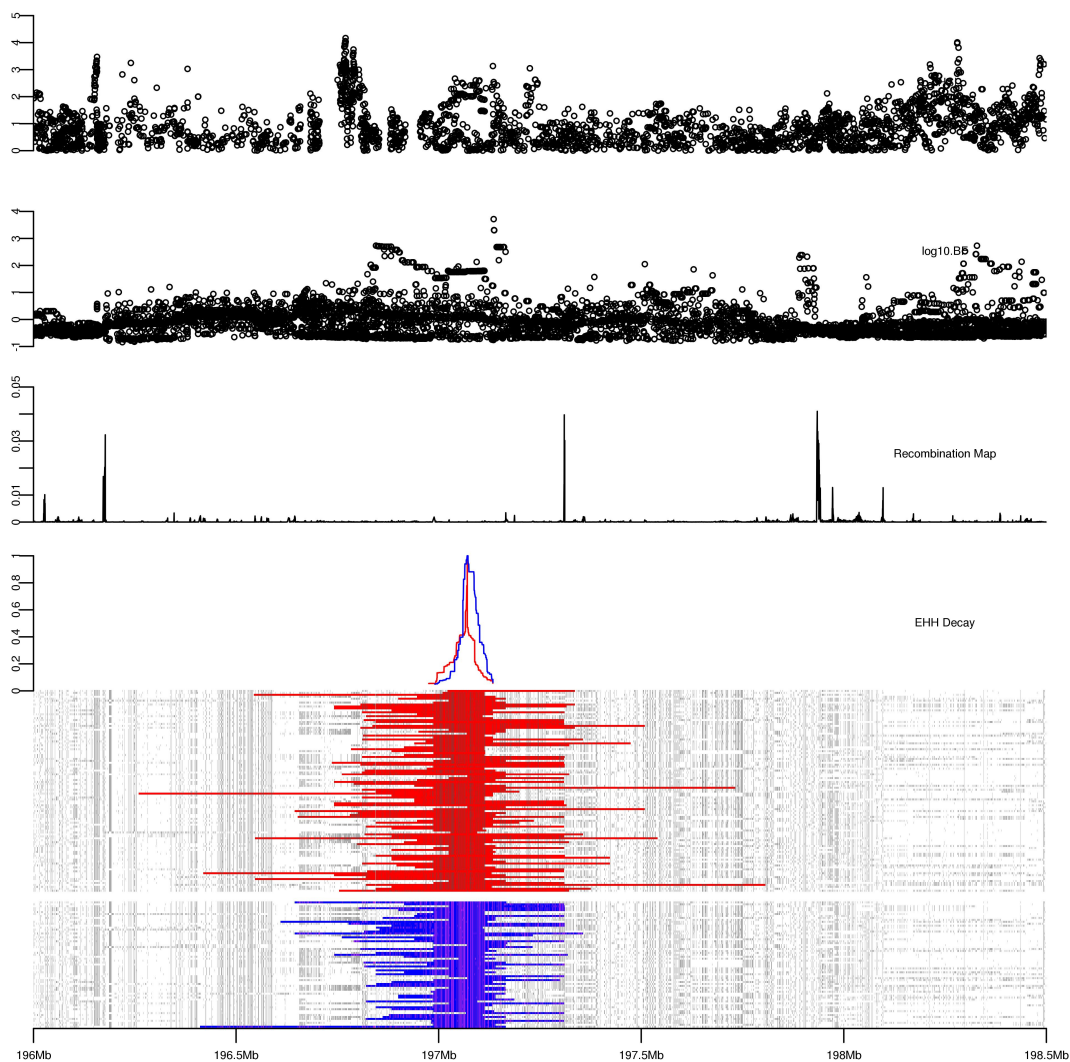


FIGURE E.1: Diagnostic plot focusing on a 2Mb region surrounding the *ASPM* gene on chromosome 1 for the CEU subpopulation. The top panel two panels show the $|iHS|$ and $\log_{10}$ Bayes Factor across the region. The middle plot shows the recombination map in cM / BP. The lower middle panel shows the EHH decay. The bottom plot shows the Viterbi lengths that overlap the SNP of interest, in this case *rs31410927*, overlaid on to the haplotype map.

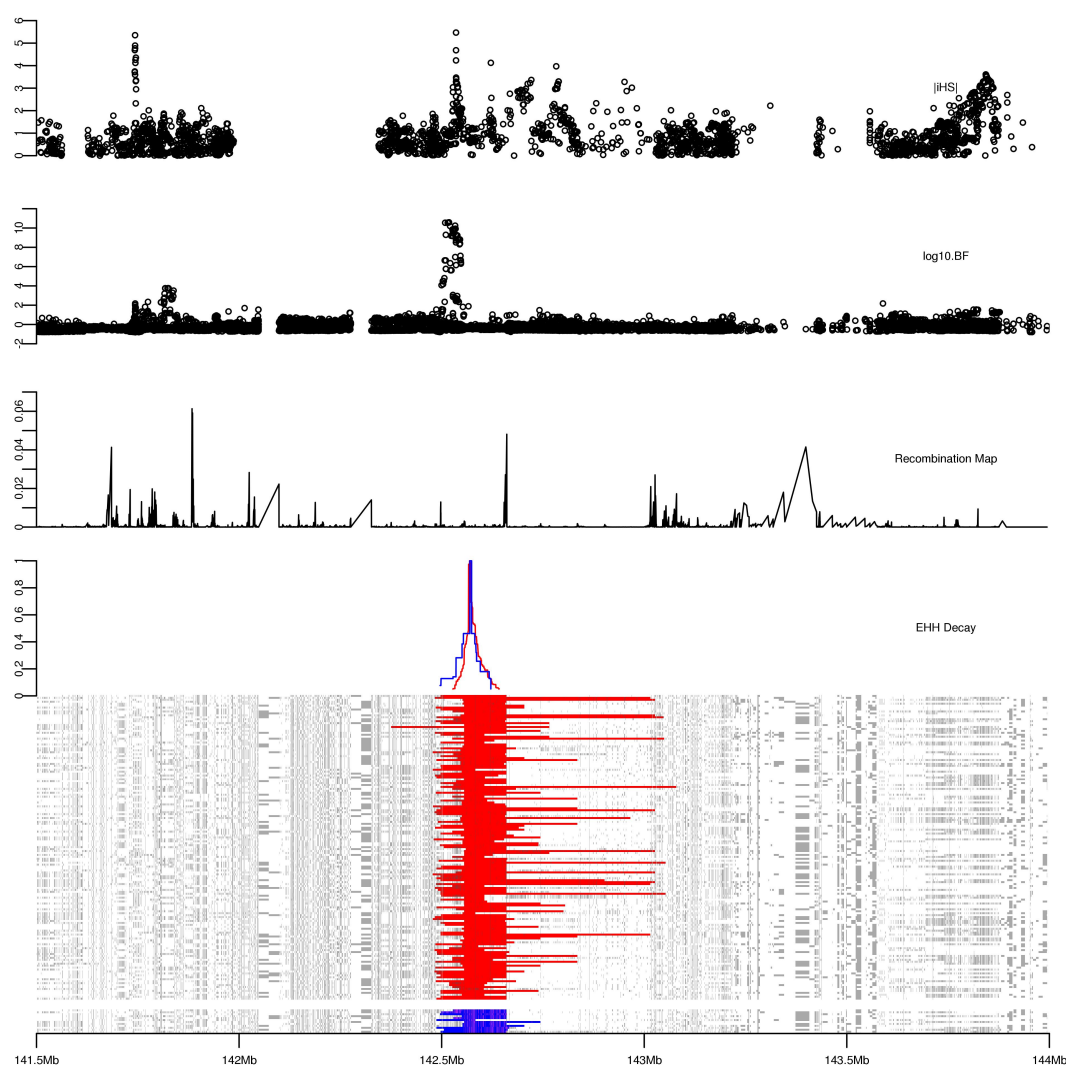


FIGURE E.2: Diagnostic plot focusing on a 2Mb region surrounding the *TRPV6* gene on chromosome 7 for the CEU subpopulation. The top panel two panels show the $|iHS|$ and $\log_{10}$ Bayes Factor across the region. The middle plot shows the recombination map in cM / BP. The lower middle panel shows the EHH decay. The bottom plot shows the Viterbi lengths that overlap the SNP of interest, in this case *rs4987682*, overlaid on to the haplotype map.

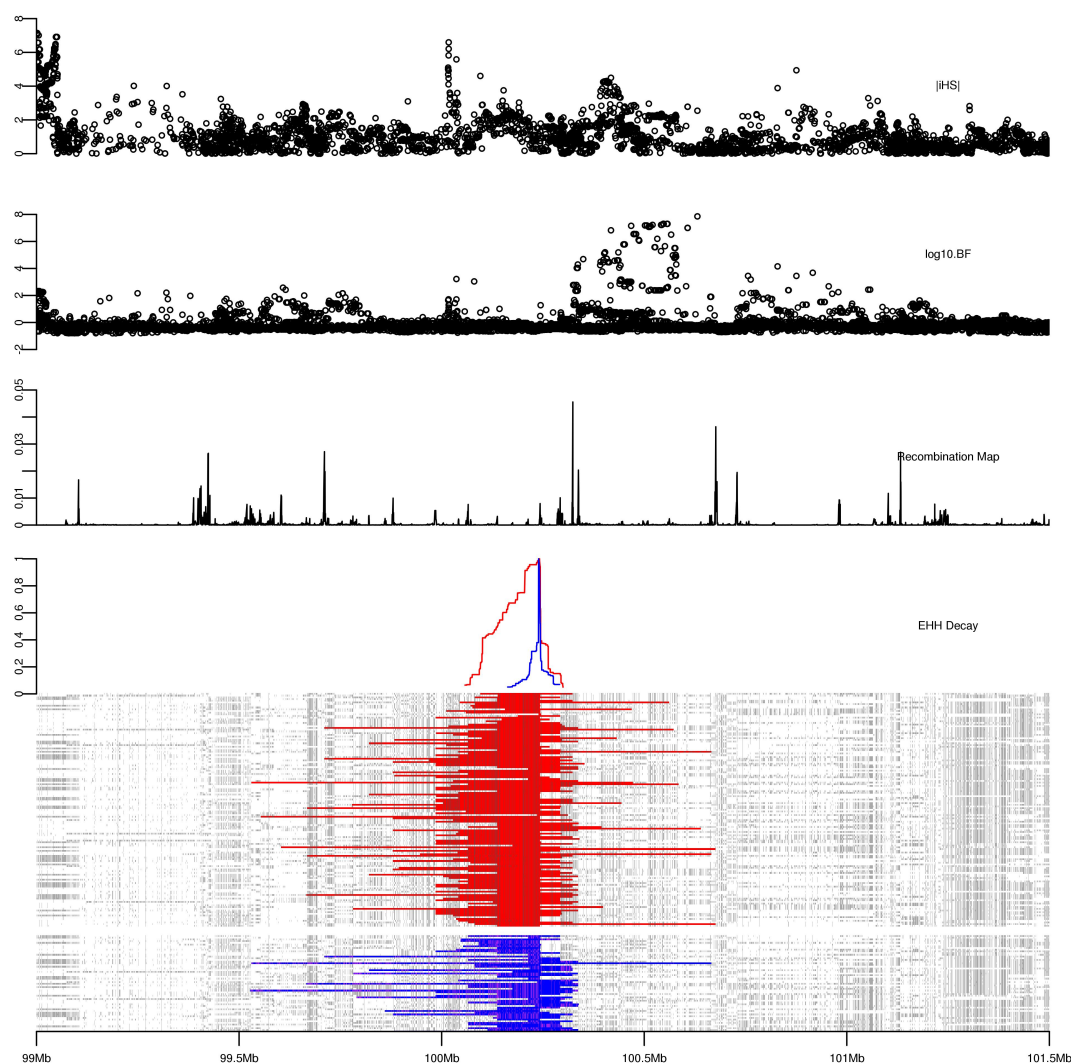


FIGURE E.3: Diagnostic plot focusing on a 2Mb region surrounding the *ADH1B* gene on chromosome 4 for the CHB subpopulation. The top panel two panels show the $|iHS|$ and $\log_{10}$ Bayes Factor across the region. The middle plot shows the recombination map in cM / BP. The lower middle panel shows the EHH decay. The bottom plot shows the Viterbi lengths that overlap the SNP of interest, in this case *rs1229984*, overlaid on to the haplotype map.

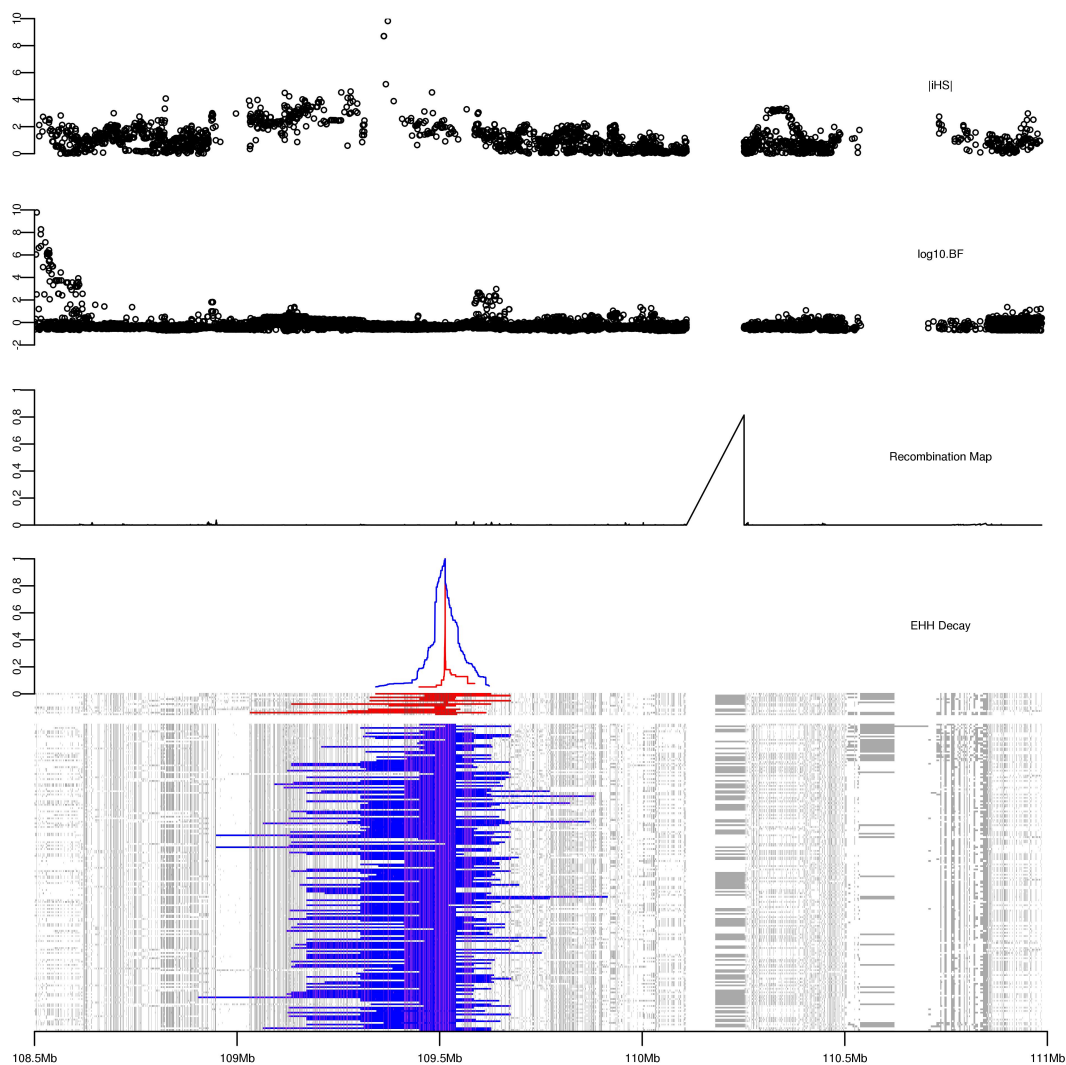


FIGURE E.4: Diagnostic plot focusing on a 2Mb region surrounding the *EDAR* gene on chromosome 2 for the CHB subpopulation. The top panel two panels show the $|iHS|$ and $\log_{10}$ Bayes Factor across the region. The middle plot shows the recombination map in cM / BP. The lower middle panel shows the EHH decay. The bottom plot shows the Viterbi lengths that overlap the SNP of interest, in this case *rs3827760*, overlaid on to the haplotype map.

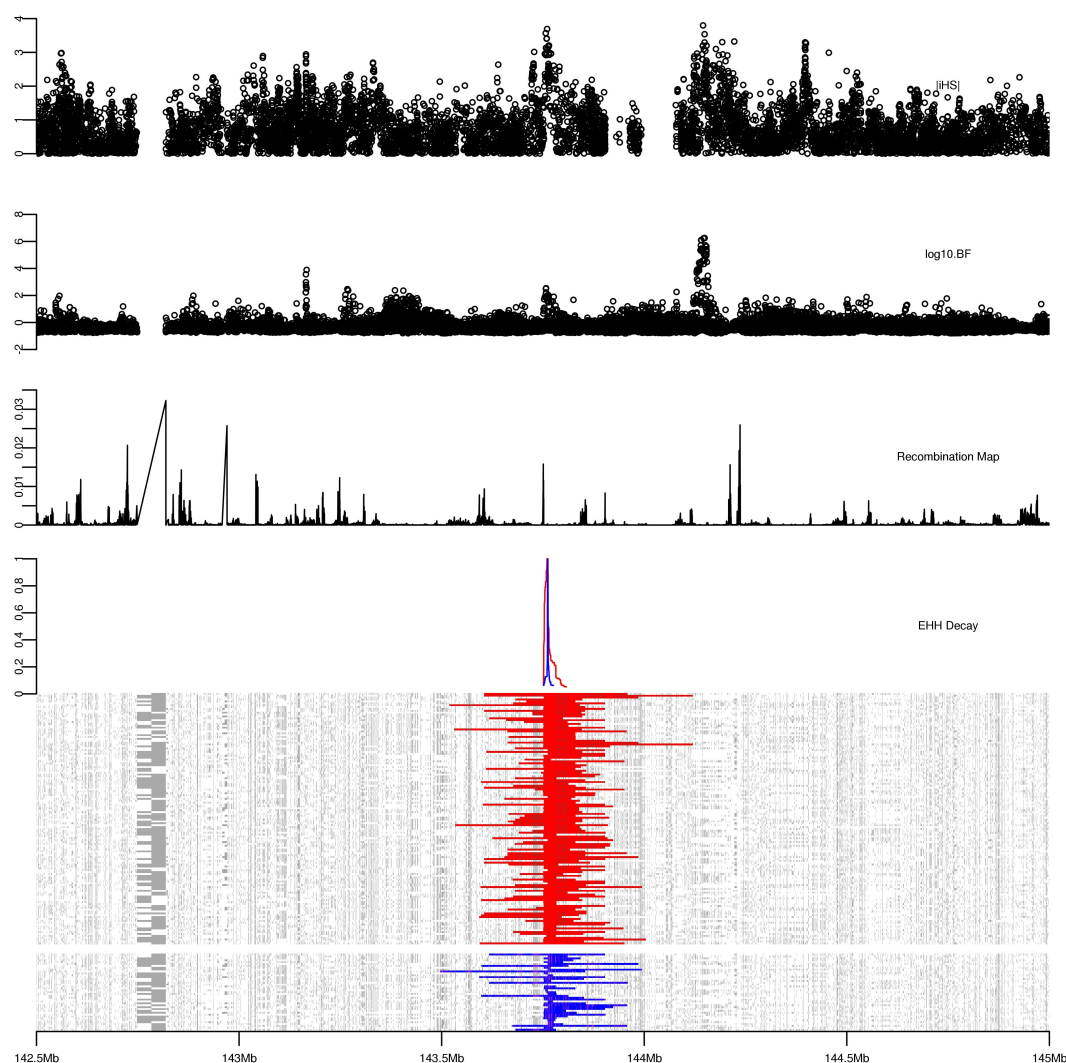


FIGURE E.5: Diagnostic plot focusing on a 2Mb region surrounding the *PSCA* gene on chromosome 8 for the YRI subpopulation. The top panel two panels show the $|iHS|$ and $\log_{10}$ Bayes Factor across the region. The middle plot shows the recombination map in cM / BP. The lower middle panel shows the EHH decay. The bottom plot shows the Viterbi lengths that overlap the SNP of interest, in this case *rs2294008*, overlaid on to the haplotype map.

Appendix F

Novel Candidate Region Analysis

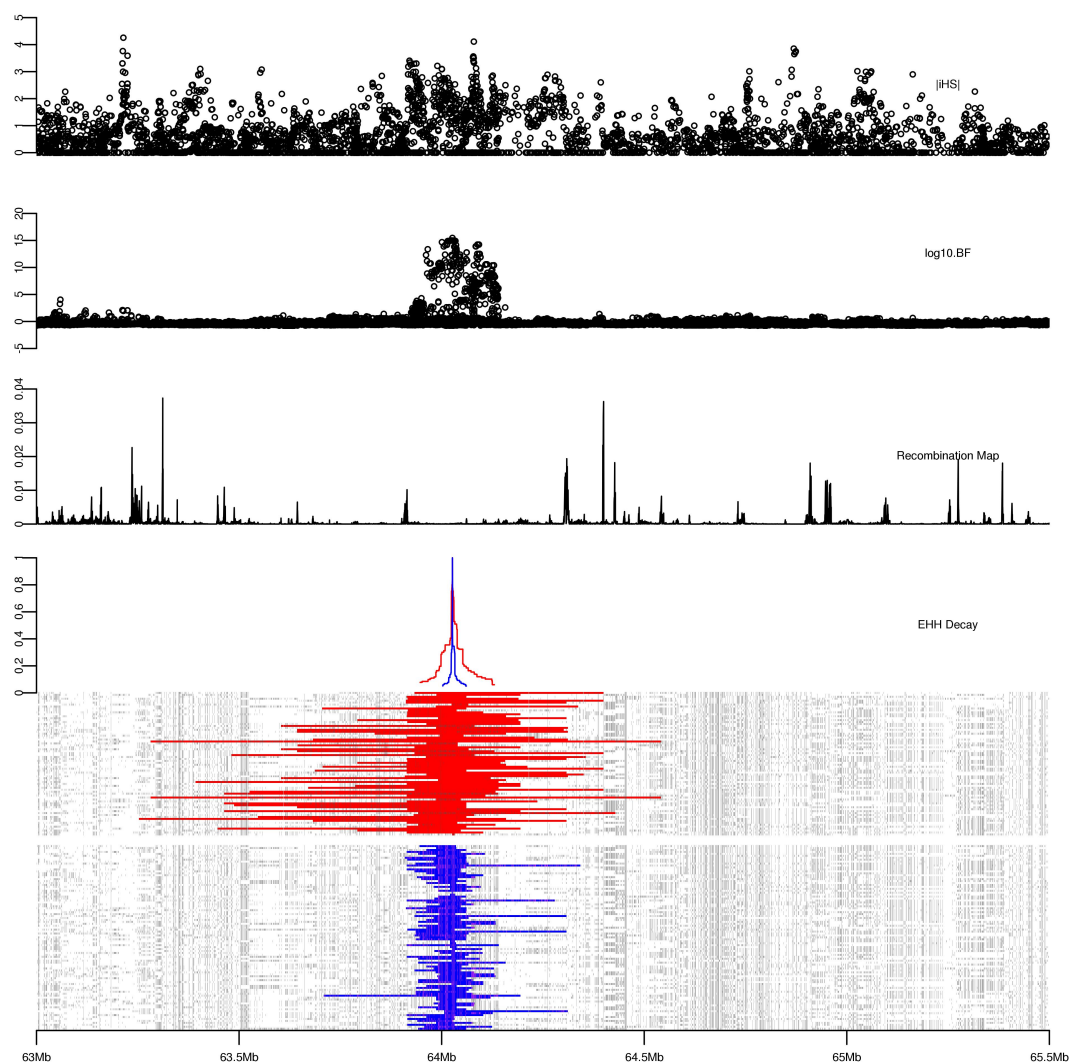


FIGURE F.1: Diagnostic plot focusing on a 2Mb region surrounding the *DPY19L2* gene on chromosome 12 for the CEU subpopulation. The top panel two panels show the $|iHS|$ and $\log_{10}$ Bayes Factor across the region. The middle plot shows the recombination map in cM / BP. The lower middle panel shows the EHH decay. The bottom plot shows the Viterbi lengths that overlap the SNP of interest, in this case rs4762993, overlaid on to the haplotype map.

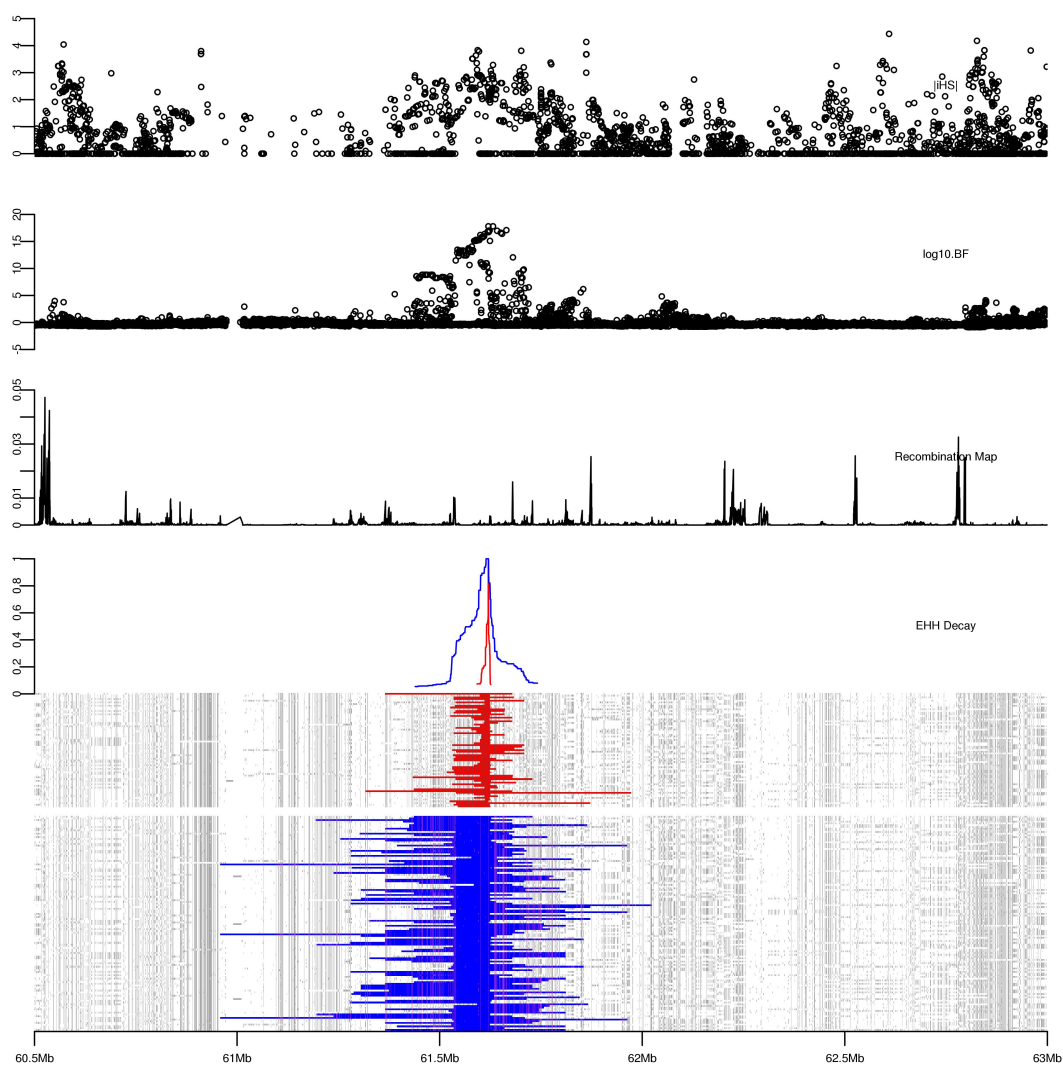


FIGURE F.2: Diagnostic plot focusing on a 2Mb region surrounding the *FADS* region on chromosome 11 for the CHB subpopulation. The top panel two panels show the $|iHS|$ and $\log_{10}$ Bayes Factor across the region. The middle plot shows the recombination map in cM / BP. The lower middle panel shows the EHH decay. The bottom plot shows the Viterbi lengths that overlap the SNP of interest, in this case rs72643559, overlaid on to the haplotype map.

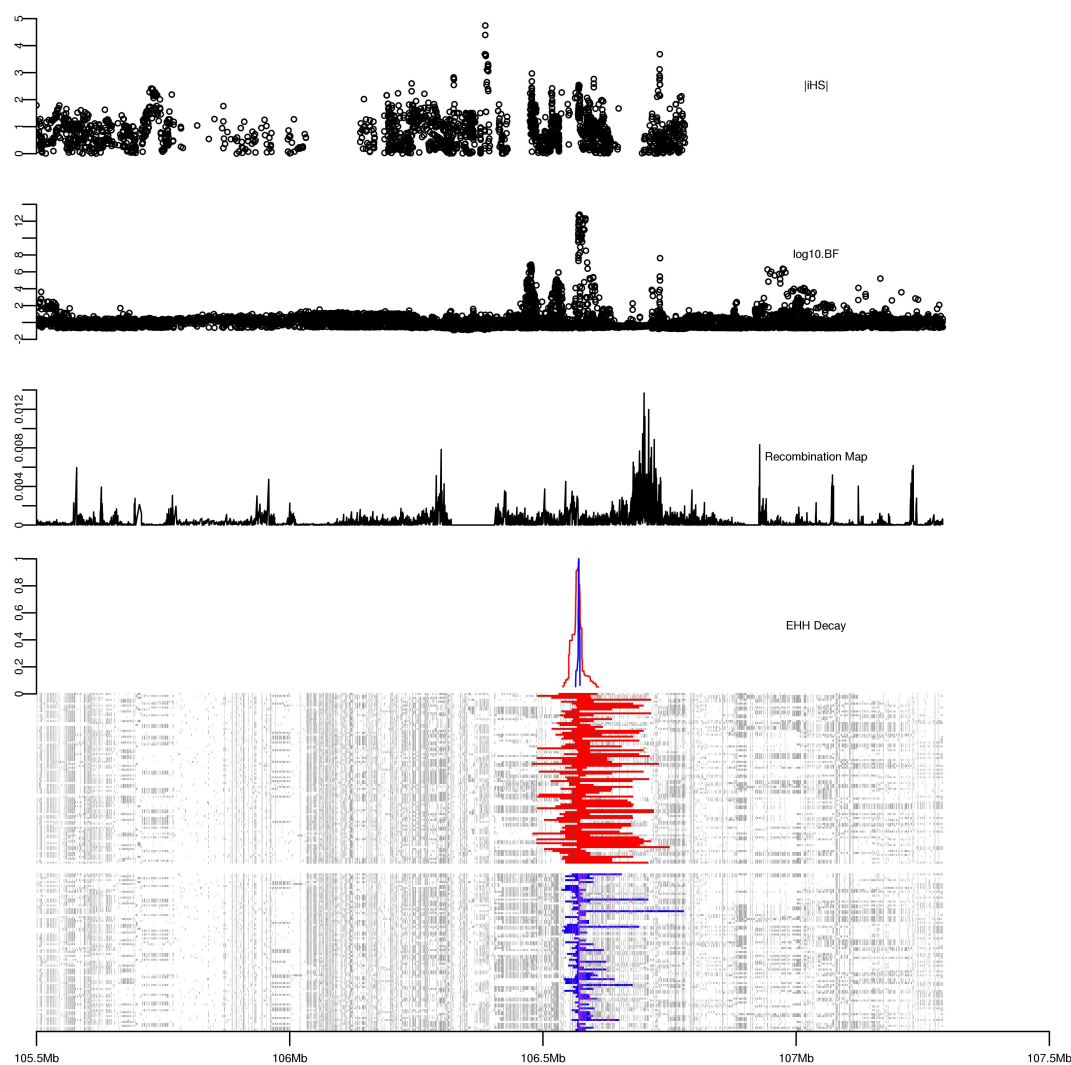


FIGURE F.3: Diagnostic plot focusing on the *IGH* locus on chromosome 14 for the CEU subpopulation. The top panel two panels show the $|iHS|$ and $\log_{10}$ Bayes Factor across the region. The middle plot shows the recombination map in cM / BP. The lower middle panel shows the EHH decay. The bottom plot shows the Viterbi lengths that overlap the SNP of interest, in this case rs72693980, overlaid on to the haplotype map.

Bibliography

- 1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., and McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.
- Achaz, G. (2009). Frequency spectrum neutrality tests: one for all and all for one. *Genetics*, 183(1):249–258.
- Ackerman, H., Usen, S., Jallow, M., Sisay Joof, F., Pinder, M., and Kwiatkowski, D. P. (2005). A comparison of case-control and family-based association methods: the example of sickle-cell and malaria. *Annals of human genetics*, 69(Pt 5):559–565.
- Akey, J. M. (2006). TRPV6 exhibits unusual patterns of polymorphism and divergence in worldwide populations. *Human molecular genetics*, 15(13):2106–2113.
- Akey, J. M. (2009). Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Research*, 19(5):711–722.
- Akey, J. M., Eberle, M. A., Rieder, M. J., Carlson, C. S., Shriver, M. D., Nickerson, D. A., and Kruglyak, L. (2004). Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biology*, 2(10):e286.

- Albrechtsen, A., Moltke, I., and Nielsen, R. (2010). Natural selection and the distribution of identity-by-descent in the human genome. *Genetics*, 186(1):295–308.
- Ashley-Koch, A., Yang, Q., and Olney, R. S. (2000). Sick cell hemoglobin (Hb S) allele and sick cell disease: a HuGE review. *American Journal of Epidemiology*, 151(9):839–845.
- Bamshad, M. and Wooding, S. P. (2003). Signatures of natural selection in the human genome. *Nature Reviews Genetics*, 4(2):99–111.
- Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge University Press.
- Barrett, R. D. H. and Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, 23(1):38–44.
- Baum, L. E. (1972). An equality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3:1–8.
- Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.
- Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E., and Hirschhorn, J. N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human Genetics*, 74(6):1111–1120.
- Bhatia, G., Patterson, N., Pasaniuc, B., Zaitlen, N., Genovese, G., Pollack, S., Mallick, S., Myers, S., Tandon, A., Spencer, C., Palmer, C. D., Adeyemo, A. A., Akylbekova, E. L., Cupples, L. A., Divers, J., Fornage, M., Kao, W. H. L., Lange, L., Li, M., Musani, S., Mychaleckyj, J. C., Ogunniyi, A., Papanicolaou,

- G., Rotimi, C. N., Rotter, J. I., Ruczinski, I., Salako, B., Siscovick, D. S., Tayo, B. O., Yang, Q., McCarroll, S., Sabeti, P., Lettre, G., De Jager, P., Hirschhorn, J., Zhu, X., Cooper, R., Reich, D., Wilson, J. G., and Price, A. L. (2011). Genome-wide comparison of African-ancestry populations from CARE and other cohorts reveals signals of natural selection. *American journal of human genetics*, 89(3):368–381.
- Black, W. C., Baer, C. F., Antolin, M. F., and DuTeau, N. M. (2001). Population genomics: genome-wide sampling of insect populations. *Annual review of entomology*, 46:441–469.
- Bond, J., Roberts, E., Mochida, G. H., Hampshire, D. J., Scott, S., Askham, J. M., Springell, K., Mahadevan, M., Crow, Y. J., Markham, A. F., Walsh, C. A., and Woods, C. G. (2002). ASPM is a major determinant of cerebral cortical size. *Nature Genetics*, 32(2):316–320.
- Borghans, J. A. M., Beltman, J. B., and De Boer, R. J. (2004). MHC polymorphism under host-pathogen coevolution. *Immunogenetics*, 55(11):732–739.
- Bright, R. A., Shay, D. K., Shu, B., Cox, N. J., and Klimov, A. I. (2006). Adamantane resistance among influenza A viruses isolated early during the 2005-2006 influenza season in the United States. *JAMA : the journal of the American Medical Association*, 295(8):891–894.
- Browning, S. R. and Browning, B. L. (2007). Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *The American Journal of Human Genetics*, 81(5):1084–1097.
- Bryk, J., Hardouin, E., Pugach, I., Hughes, D., Strotmann, R., Stoneking, M., and Myles, S. (2008). Positive selection in East Asians for an EDAR allele that enhances NF-kappaB activation. *PloS one*, 3(5):e2209.

- Carlson, C. S., Thomas, D. J., Eberle, M. A., Swanson, J. E., Livingston, R. J., Rieder, M. J., and Nickerson, D. A. (2005). Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Research*, 15(11):1553–1565.
- Casella, G. and Berger, R. L. (1990). *Statistical inference*, volume 70. Duxbury Press Belmont, CA.
- Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, 2(4):e64.
- Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4):327–335.
- Cirulli, E. T. and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, 11(6):415–425.
- Clavel, F. and Hance, A. J. (2004). HIV drug resistance. *The New England journal of medicine*, 350(10):1023–1035.
- Coop, G. and Griffiths, R. C. (2004). Ancestral inference on gene trees under selection. *Theoretical Population Biology*, 66(3):219–232.
- Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., and François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, 25(7):410–418.
- Darwin, C. (1859). On the origins of species by means of natural selection. *London: Murray*.
- Darwin, C. and Wallace, A. (1858). 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*, 3(9):45–62.
- Dean, J. and Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113.
- Del Moral, P., Doucet, A., and Jasra, A. (2011). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, pages 1–12.
- Delaneau, O., Coulonges, C., and Zagury, J.-F. (2008). Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics*, 9:540.
- Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods*, 10(1):5–6.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.
- Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *The Annals of statistics*, 7(2):269–281.
- Doucet, A., De Freitas, N., and Gordon, N. (2001). An introduction to sequential Monte Carlo methods. pages 3–14.
- Durbin, R. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.
- Durbin, R. M., Altshuler, D. L., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S. B., Gibbs, R. A., Knoppers, B. M., Lander, E. S., Lehrach, H., Mardis, E. R., McVean, G. A., Nickerson, D. A., Peltonen, L., Schafer, A. J., Sherry, S. T., Wang, J., Wilson, R. K., Gibbs, R. A., Deiros, D., Metzker, M.,

Muzny, D., Reid, J., Wheeler, D., Wang, J., Li, J., Jian, M., Li, G., Li, R., Liang, H., Tian, G., Wang, B., Wang, J., Wang, W., Yang, H., Zhang, X., Zheng, H., Ambrogio, L., Bloom, T., Cibulskis, K., Fennell, T. J., Jaffe, D. B., Shefler, E., Sougnez, C. L., Gormley, N., Humphray, S., Kingsbury, Z., Koko-Gonzales, P., Stone, J., McKernan, K. J., Costa, G. L., Ichikawa, J. K., Lee, C. C., Sudbrak, R., Borodina, T. A., Dahl, A., Davydov, A. N., Marquardt, P., Mertes, F., Nietfeld, W., Rosenstiel, P., Schreiber, S., Soldatov, A. V., Timmermann, B., Tolzmann, M., Affourtit, J., Ashworth, D., Attiya, S., Bachorski, M., Buglione, E., Burke, A., Caprio, A., Celone, C., Clark, S., Connors, D., Desany, B., Gu, L., Guccione, L., Kao, K., Kebbel, A., Knowlton, J., Labrecque, M., McDade, L., Mealmaker, C., Minderman, M., Nawrocki, A., Niazi, F., Pareja, K., Ramenani, R., Riches, D., Song, W., Turcotte, C., Wang, S., Dooling, D., Fulton, L., Fulton, R., Weinstock, G., Burton, J., Carter, D. M., Churcher, C., Coffey, A., Cox, A., Palotie, A., Quail, M., Skelly, T., Stalker, J., Swerdlow, H. P., Turner, D., De Witte, A., Giles, S., Bainbridge, M., Challis, D., Sabo, A., Yu, F., Yu, J., Fang, X., Guo, X., Li, Y., Luo, R., Tai, S., Wu, H., Zheng, H., Zheng, X., Zhou, Y., Marth, G. T., Garrison, E. P., Huang, W., Indap, A., Kural, D., Lee, W.-P., Fung Leong, W., Quinlan, A. R., Stewart, C., Stromberg, M. P., Ward, A. N., Wu, J., Lee, C., Mills, R. E., Shi, X., Daly, M. J., DePristo, M. A., Ball, A. D., Banks, E., Browning, B. L., Garimella, K. V., Grossman, S. R., Handsaker, R. E., Hanna, M., Hartl, C., Kernytsky, A. M., Korn, J. M., Li, H., Maguire, J. R., McCarroll, S. A., McKenna, A., Nemesh, J. C., Philippakis, A. A., Poplin, R. E., Price, A., Rivas, M. A., Sabeti, P. C., Schaffner, S. F., Shlyakhter, I. A., Cooper, D. N., Ball, E. V., Mort, M., Phillips, A. D., Stenson, P. D., Sebat, J., Makarov, V., Ye, K., Yoon, S. C., Bustamante, C. D., Boyko, A., Degenhardt, J., Gravel, S., Gutenkunst, R. N., Kaganovich, M., Keinan, A., Lacroute, P., Ma, X., Reynolds, A., Clarke, L., Cunningham, F., Herrero, J., Keenen, S., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Smith, R. E., Zalunin, V., Zheng-Bradley, X., Korbel, J. O., Stütz, A. M.,

- Bauer, M., Keira Cheetham, R., Cox, T., Eberle, M., James, T., Kahn, S., Murray, L., Ye, K., Fu, Y., Hyland, F. C. L., Manning, J. M., McLaughlin, S. F., Peckham, H. E., Sakarya, O., Sun, Y. A., Tsung, E. F., Batzer, M. A., Konkel, M. K., Walker, J. A., Albrecht, M. W., Amstislavskiy, V. S., Herwig, R., Parkhomchuk, D. V., Agarwala, R., Khouri, H. M., Morgulis, A. O., Paschall, J. E., Phan, L. D., Rotmistrovsky, K. E., Sanders, R. D., Shumway, M. F., Xiao, C., Auton, A., Iqbal, Z., Lunter, G., Marchini, J. L., Moutsianas, L., Myers, S., Tumian, A., and Knigh... (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.
- Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of statistics*, 7(1):1–26.
- ElInati, E., Kuentz, P., Redin, C., Jaber, S., Vanden Meerschaut, F., Makarian, J., Kosciński, I., Nasr-Esfahani, M. H., Demiroğlu, A., Gurgan, T., Louanjli, N., Iqbal, N., Bisharah, M., Pigeon, F. C., Gourabi, H., De Briel, D., Brugnon, F., Gitlin, S. A., Grillo, J.-M., Ghaedi, K., Deemeh, M. R., Tanhaei, S., Modarres, P., Heindryckx, B., Benkhalifa, M., Nikiforaki, D., Oehninger, S. C., De Sutter, P., Muller, J., and Viville, S. (2012). Globozoospermia is mainly due to DPY19L2 deletion via non-allelic homologous recombination involving two recombination hotspots. *Human molecular genetics*, 21(16):3695–3702.
- Enattah, N. S., Sahi, T., Savilahti, E., Terwilliger, J. D., Peltonen, L., and Järvelä, I. (2002). Identification of a variant associated with adult-type hypolactasia. *Nature Genetics*, 30(2):233–237.
- Ewing, G. and Hermisson, J. (2010). MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16):2064–2065.
- Fay, J. C. and Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, 155(3):1405–1413.

- Fearnhead, P. and Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics*, 159(3):1299–1318.
- Fearnhead, P. and Prangle, D. (2010). Constructing summary statistics for approximate Bayesian computation: Semi-automatic ABC. *Arxiv preprint arxiv:1004.1112*.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376.
- Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., García-Girón, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kähäri, A. K., Keenan, S., Komorowska, M., Kulesha, E., Longden, I., Maurel, T., McLaren, W. M., Muffato, M., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H. S., Ritchie, G. R. S., Ruffier, M., Schuster, M., Sheppard, D., Sobral, D., Taylor, K., Thormann, A., Trevanion, S., White, S., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Harrow, J., Herrero, J., Hubbard, T. J. P., Johnson, N., Kinsella, R., Parker, A., Spudich, G., Yates, A., Zadissa, A., and Searle, S. M. J. (2013). Ensembl 2013. *Nucleic acids research*, 41(Database issue):D48–55.
- Fu, Y. X. and Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics*, 133(3):693–709.
- Fujimoto, A., Ohashi, J., Nishida, N., Miyagawa, T., Morishita, Y., Tsunoda, T., Kimura, R., and Tokunaga, K. (2008). A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in Asia. *Human genetics*, 124(2):179–185.
- Galton, F. (1869). Hereditary genius - Sir Francis Galton - Google Books.

- Garud, N. R., Messer, P. W., Buzbas, E. O., and Petrov, D. A. (2013). Soft selective sweeps are the primary mode of recent adaptation in *Drosophila melanogaster*. Technical report.
- Gelman, A. and Meng, X. L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185.
- Gonzalez-Angulo, A. M., Hennessy, B. T. J., and Mills, G. B. (2010). Future of personalized medicine in oncology: a systems biology approach. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 28(16):2777–2783.
- Goymer, P. (2008). Natural selection: The evolution of cancer. *Nature*, 454(7208):1046–1048.
- Griffiths, R. C. (1981). Neutral two-locus multiple allele models with recombination. *Theoretical Population Biology*, 19(2):169–186.
- Griffiths, R. C. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, 3(4):479–502.
- Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10):e1000695.
- Han, L. and Abney, M. (2013). Using identity by descent estimation with dense genotype data to detect positive selection. *European journal of human genetics : EJHG*, 21(2):205–211.
- Hartl, D. L. and Clark, A. G. (2007). Principles of population genetics. 116.
- Hastie, T. J., Tibshirani, R. J., and Friedman, J. J. H. (2009). *The elements of statistical learning*. Springer.

- Hein, J., Schierup, M. H., and Wiuf, C. (2005). *Gene genealogies, variation and evolution*. Oxford University Press, Oxford.
- Hellenthal, G. and Stephens, M. (2007). msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics*, 23(4):520–521.
- Hill, A. V., Allsopp, C. E., Kwiatkowski, D., Anstey, N. M., Twumasi, P., Rowe, P. A., Bennett, S., Brewster, D., McMichael, A. J., and Greenwood, B. M. (1991). Common west African HLA antigens are associated with protection from severe malaria. *Nature*, 352(6336):595–600.
- Hill, W. G. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38(6):226–231.
- Hinch, A. G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C. D., Chen, G. K., Wang, K., Buxbaum, S. G., Akylbekova, E. L., Aldrich, M. C., Ambrosone, C. B., Amos, C., Bandera, E. V., Berndt, S. I., Bernstein, L., Blot, W. J., Bock, C. H., Boerwinkle, E., Cai, Q., Caporaso, N., Casey, G., Cupples, L. A., Deming, S. L., Diver, W. R., Divers, J., Fornage, M., Gillanders, E. M., Glessner, J., Harris, C. C., Hu, J. J., Ingles, S. A., Isaacs, W., John, E. M., Kao, W. H. L., Keating, B., Kittles, R. A., Kolonel, L. N., Larkin, E., Le Marchand, L., McNeill, L. H., Millikan, R. C., Murphy, A., Musani, S., Neslund-Dudas, C., Nyante, S., Papanicolaou, G. J., Press, M. F., Psaty, B. M., Reiner, A. P., Rich, S. S., Rodriguez-Gil, J. L., Rotter, J. I., Rybicki, B. A., Schwartz, A. G., Signorello, L. B., Spitz, M., Strom, S. S., Thun, M. J., Tucker, M. A., Wang, Z., Wiencke, J. K., Witte, J. S., Wrensch, M., Wu, X., Yamamura, Y., Zanetti, K. A., Zheng, W., Ziegler, R. G., Zhu, X., Redline, S., Hirschhorn, J. N., Henderson, B. E., Taylor, H. A., Price, A. L., Hakonarson, H., Chanock, S. J., Haiman, C. A., Wilson, J. G., Reich, D., and Myers, S. R. (2011). The landscape of recombination in African Americans. *Nature*, 476(7359):170–175.

- Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A., and Cox, D. R. (2005). Whole-genome patterns of common DNA variation in three human populations. *Science*, 307(5712):1072–1079.
- Hollox, E. J., Poulter, M., Zvarik, M., Ferak, V., Krause, A., Jenkins, T., Saha, N., Kozlov, A. I., and Swallow, D. M. (2001). Lactase haplotype diversity in the Old World. *The American Journal of Human Genetics*, 68(1):160–172.
- Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6):e1000529.
- Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2):183–201.
- Hudson, R. R. (2001). Two-locus sampling distributions and their application. *Genetics*, 159(4):1805–1817.
- Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338.
- Hudson, R. R., Bailey, K., Skarecky, D., Kwiatowski, J., and Ayala, F. J. (1994). Evidence for positive selection in the superoxide dismutase (Sod) region of *Drosophila melanogaster*. *Genetics*, 136(4):1329–1340.
- International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320.
- Itan, Y., Jones, B. L., Ingram, C. J., Swallow, D. M., and Thomas, M. G. (2010). A worldwide correlation of lactase persistence phenotype and genotypes. *BMC evolutionary biology*, 10(1):36.
- Joyce, P. and Marjoram, P. (2008). Approximately sufficient statistics and bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 7(1):Article26.

- Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. In *Mammalian protein metabolism*, pages 21–132. Academic Press, N. Y.
- Kandpal, R., Saviola, B., and Felton, J. (2009). The era of 'omics unlimited. *BioTechniques*, 46(5):351–2– 354–5.
- Kelley, J. L. (2006). Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Research*, 16(8):980–989.
- Kiefer, J. (1953). Sequential minimax search for a maximum. *Proceedings of the American Mathematical Society*, 4(3):502–506.
- Kim, H. L., Igawa, T., Kawashima, A., Satta, Y., and Takahata, N. (2010). Divergence, demography and gene loss along the human lineage. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1552):2451–2457.
- Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893.
- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16(2):111–120.
- Kimura, M. and Crow, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics*, 49(4):725.
- Kimura, R., Fujimoto, A., Tokunaga, K., and Ohashi, J. (2007). A practical genome scan for population-specific strong selective sweeps that have reached fixation. *PloS one*, 2(3):e286.
- Kingman, J. F. C. (1982a). On the genealogy of large populations. *Journal of Applied Probability*, (Special Vol. 19A):27–43.
- Kingman, J. F. C. (1982b). The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248.

- Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. (2012). The Big Data Bootstrap. *arXiv preprint arXiv:1206.6415*.
- Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S. T., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R., and Stefansson, K. (2002). A high-resolution recombination map of the human genome. *Nature Genetics*, 31(3):241–247.
- Kouprina, N., Pavlicek, A., Mochida, G. H., Solomon, G., Gersch, W., Yoon, Y.-H., Collura, R., Ruvolo, M., Barrett, J. C., Woods, C. G., Walsh, C. A., Jurka, J., and Larionov, V. (2004). Accelerated evolution of the ASPM gene controlling brain size begins prior to human brain expansion. *PLoS Biology*, 2(5):E126.
- Krone, S. and Neuhauser, C. (1997). Ancestral Processes with Selection. *Theoretical Population Biology*, 51(3):210–237.
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish,

W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglu, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F.,

- Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.
- Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics*, 8(1):e1002453.
- Leffler, E. M., Gao, Z., Pfeifer, S., Séguirel, L., Auton, A., Venn, O., Bowden, R., Bontrop, R., Wall, J. D., Sella, G., Donnelly, P., McVean, G., and Przeworski, M. (2013). Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science*, 339(6127):1578–1582.
- Li, H., Mukherjee, N., Soundararajan, U., Tarnok, Z., Barta, C., Khaliq, S., Mohyuddin, A., Kajuna, S. L. B., Mehdi, S. Q., Kidd, J. R., and Kidd, K. K. (2007). Geographically separate increases in the frequency of the derived ADH1B*47His allele in eastern and western Asia. *The American Journal of Human Genetics*, 81(4):842–846.
- Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233.
- Li, Y., Sidore, C., Kang, H. M., Boehnke, M., and Abecasis, G. R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. *Genome Research*, 21(6):940–951.
- Lopes, J. S. and Beaumont, M. A. (2010). ABC: A useful Bayesian tool for the analysis of population data. *Infection, Genetics and Evolution*, 10(6):825–832.

- Luikart, G., Allendorf, F. W., Cornuet, J. M., and Sherwin, W. B. (1998). Distortion of allele frequency distributions provides a test for recent population bottlenecks. *The Journal of heredity*, 89(3):238–247.
- Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39(7):906–913.
- Mardis, E. R. (2011). A decade’s perspective on DNA sequencing technology. *Nature*, 470(7333):198–203.
- Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324.
- Marth, G., Schuler, G., Yeh, R., Davenport, R., Agarwala, R., Church, D., Whee-
lan, S., Baker, J., Ward, M., Kholodov, M., Phan, L., Czabarka, E., Murvai,
J., Cutler, D., Wooding, S., Rogers, A., Chakravarti, A., Harpending, H. C.,
Kwok, P.-Y., and Sherry, S. T. (2003). Sequence variations in the public hu-
man genome data reflect a bottlenecked population history. *Proceedings of the
National Academy of Sciences*, 100(1):376–381.
- McDonald, J. H. and Kreitman, M. (1991). Adaptive protein evolution at the Adh
locus in *Drosophila*. *Nature*, 351(6328):652–654.
- McVean, G. (2007). The structure of linkage disequilibrium around a selective
sweep. *Genetics*, 175(3):1395–1406.
- McVean, G. A. T. (2002). A genealogical interpretation of linkage disequilibrium.
Genetics, 162(2):987–991.
- McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and
Donnelly, P. (2004). The fine-scale structure of recombination rate variation in
the human genome. *Science*, 304(5670):581–584.

- Mekel-Bobrov, N., Gilbert, S. L., Evans, P. D., Vallender, E. J., Anderson, J. R., Hudson, R. R., Tishkoff, S. A., and Lahn, B. T. (2005). Ongoing adaptive evolution of ASPM, a brain size determinant in *Homo sapiens*. *Science*, 309(5741):1720–1722.
- Messer, P. W. and Petrov, D. A. (2013). Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution*.
- Miotto, O., Almagro-Garcia, J., Manske, M., MacInnis, B., Campino, S., Rockett, K. A., Amaratunga, C., Lim, P., Suon, S., Sreng, S., Anderson, J. M., Duong, S., Nguon, C., Chuor, C. M., Saunders, D., Se, Y., Lon, C., Fukuda, M. M., Amenga-Etego, L., Hodgson, A. V. O., Asoala, V., Imwong, M., Takala-Harrison, S., Nosten, F. c. o., Su, X.-z., Ringwald, P., Ariey, F. e. d. e. r., Dolecek, C., Hien, T. T., Boni, M. F., Thai, C. Q., Amambua-Ngwa, A., Conway, D. J., eacute, A. A. D., Doumbo, O. K., Zongo, I., Ouedraogo, J.-B., Alcock, D., Drury, E., Auburn, S., Koch, O., Sanders, M., Hubbard, C., Maslen, G., Ruano-Rubio, V., Jyothi, D., Miles, A., O'Brien, J., Gamble, C., Oyola, S. O., Rayner, J. C., Newbold, C. I., Berriman, M., Spencer, C. C. A., McVean, G., Day, N. P., White, N. J., Bethell, D., Dondorp, A. M., Plowe, C. V., Fairhurst, R. M., and Kwiatkowski, D. P. (2013). Multiple populations of artemisinin-resistant *Plasmodium falciparum* in Cambodia. *Nature Genetics*, pages 1–10.
- Mita, T., Venkatesan, M., Ohashi, J., Culleton, R., Takahashi, N., Tsukahara, T., Ndounga, M., Dysoley, L., Endo, H., Hombhanje, F., Ferreira, M. U., Plowe, C. V., and Tanabe, K. (2011). Limited Geographical Origin and Global Spread of Sulfadoxine-Resistant dhps Alleles in *Plasmodium falciparum* Populations. *Journal of Infectious Diseases*, 204(12):1980–1988.
- Moran, J. V. (1958). A general theory of the distribution of gene frequencies. I. Overlapping generations. *Proceedings of the Royal Society of London. Series*

- B*, *Containing papers of a Biological character*. Royal Society (Great Britain), 149(934):102–112.
- Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324.
- Neuhauser, C. (1999). The ancestral graph and gene genealogy under frequency-dependent selection. *Theoretical Population Biology*, 56(2):203–214.
- Nielsen, R. (2001). Statistical tests of selective neutrality in the age of genomics. *Heredity*, 86(Pt 6):641–647.
- Noedl, H., Se, Y., Schaecher, K., Smith, B. L., Socheat, D., Fukuda, M. M., and Consortium, A. S. (2008). Evidence of Artemisinin-Resistant Malaria in Western Cambodia. *The New England journal of medicine*, 359(24):2619–2620.
- Norris, J. R. (1998). *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- Novelli, G., Predazzi, I. M., Mango, R., Romeo, F., and Mehta, J. L. (2010). Role of genomics in cardiovascular medicine. *World journal of cardiology*, 2(12):428–436.
- Nunes, M. A. and Balding, D. J. (2010). On optimal selection of summary statistics for approximate Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 9(1):1–14.
- Paul, J. S. and Song, Y. S. (2012). Blockwise HMM computation for large-scale population genomic inference. *Bioinformatics*, 28(15):2008–2015.
- Payne, D. (1987). Spread of Chloroquine Resistance in Plasmodium-Falciparum. *Parasitology Today*, 3(8):241–246.

- Pennings, P. S. and Hermisson, J. (2006). Soft sweeps II—molecular population genetics of adaptation from recurrent mutation or migration. *Molecular Biology and Evolution*, 23(5):1076–1084.
- Peter, B. M., Huerta-Sanchez, E., and Nielsen, R. (2012). Distinguishing between Selective Sweeps from Standing Variation and from a De Novo Mutation. *PLoS Genetics*, 8(10):e1003011.
- Piel, F. B., Patil, A. P., Howes, R. E., Nyangiri, O. A., Gething, P. W., Williams, T. N., Weatherall, D. J., and Hay, S. I. (2010). Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nature communications*, 1:104.
- Prezeworski, M., Coop, G., and Wall, J. D. (2005). The signature of positive selection on standing genetic variation. *Evolution*, 59(11):2312–2323.
- Provine, W. B. (1987). *The origins of theoretical population genetics*. The Chicago History of Science and Medicine. University of Chicago Press, Chicago, IL.
- Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., Boehnke, M., Abecasis, G. R., and Willer, C. J. (2010). LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, 26(18):2336–2337.
- Rabiner, L. and Juang, B. H. (1986). An introduction to hidden Markov models. *ASSP Magazine, IEEE*, 3(1):4–16.
- Robert, C. P., Beaumont, M. A., Marin, J. M., and Cornuet, J. M. (2008). Adaptivity for ABC algorithms: the ABC-PMC scheme. *Arxiv preprint arXiv:0805.2256*.
- Roper, C., Pearce, R., Nair, S., Sharp, B., Nosten, F., and Anderson, T. (2004). Intercontinental spread of pyrimethamine-resistant malaria. *Science*, 305(5687):1124–1124.

- Ruwende, C. and Hill, A. (1998). Glucose-6-phosphate dehydrogenase deficiency and malaria. *Journal of Molecular Medicine*, 76(8):581–588.
- Sabeti, P., Usen, S., Farhadian, S., Jallow, M., Doherty, T., Newport, M., Pinder, M., Ward, R., and Kwiatkowski, D. (2002a). CD40L association with protection from severe malaria. *Genes and Immunity*, 3(5):286–291.
- Sabeti, P. C. (2008). Natural Selection: Uncovering Mechanisms of Evolutionary Adaptation to Infectious Disease. *Nature Education*, 1(1).
- Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., and McDonald, G. J. (2002b). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):830–832.
- Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S., Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Wayne, M. M. Y., Tsui, S. K. W., Xue, H., Tze-Fei Wong, J., Galver, L. M., Fan, J.-B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J.-F., Phillips, M. S., Roumy, S., Sallée, C., Verner, A., Hudson, T. J., Kwok, P.-Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L.,

Taillon-Miller, P., Xiao, M., Tsui, L.-C., Mak, W., Qiang Song, Y., Tam, P. K. H., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., Daly, M. J., de Bakker, P. I. W., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D. J., Sabeti, P., Saxena, R., Sham, P. C., Stein, L. D., Krishnan, L., Vernon Smith, A., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Tsunoda, T., Johnson, T. A., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D. M., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Yakub, I., Richter, D. J., Birren, B. W., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archevêque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Ota Wang, V., Peterson, J. L., and Sh... (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164):913–918.

- Scally, A. and Durbin, R. (2012). Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics*, 13(10):745–753.
- Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629–644.
- Schlebusch, C. M., Sjödin, P., Skoglund, P., and Jakobsson, M. (2013). Stronger signal of recent selection for lactase persistence in Maasai than in Europeans. *European journal of human genetics : EJHG*, 21(5):550–553.
- Simonsen, K. L., Churchill, G. A., and Aquadro, C. F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, 141(1):413–429.
- Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760.
- Slatkin, M. (2001). Simulating genealogies of selected alleles in a population of variable size. *Genetical research*, 78(1):49–57.
- Slatkin, M. and Hudson, R. R. (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, 129(2):555–562.
- Smith, J. M. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical research*, 23(1):23–35.
- Spencer, C. C. A. and Coop, G. (2004). SelSim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics*, 20(18):3673–3675.
- Steer, C. D., Hibbeln, J. R., Golding, J., and Davey Smith, G. (2012). Polyunsaturated fatty acid levels in blood during pregnancy, at birth and at 7 years:

- their associations with two common FADS2 polymorphisms. *Human molecular genetics*, 21(7):1504–1512.
- Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):605–635.
- Stephens, M. and Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *The American Journal of Human Genetics*, 73(5):1162–1169.
- Stephens, M., Smith, N. J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68(4):978–989.
- Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2):437–460.
- Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595.
- Tanaka, T. and Nei, M. (1989). Positive darwinian selection observed at the variable-region genes of immunoglobulins. *Molecular Biology and Evolution*, 6(5):447–459.
- Tang, K., Thornton, K. R., and Stoneking, M. (2007). A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biology*, 5(7):e171.
- Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505.
- Teshima, K. M. and Innan, H. (2009). mbs: modifying Hudson’s ms software to generate samples of DNA sequences with a biallelic site under selection. *BMC Bioinformatics*, 10(1):166.

- Vanhoucke, V., Senior, A., and Mao, M. Z. (2011). Improving the speed of neural networks on CPUs.
- Venables, W. N., Ripley, B. D., and Venables, W. N. (1994). *Modern applied statistics with S-PLUS*, volume 250. Springer-verlag New York.
- Verrelli, B. C., McDonald, J. H., Argyropoulos, G., Destro-Bisol, G., Froment, A., Drousiotou, A., Lefranc, G., Helal, A. N., Loiselet, J., and Tishkoff, S. A. (2002). Evidence for balancing selection from nucleotide sequence analyses of human G6PD. *The American Journal of Human Genetics*, 71(5):1112–1128.
- Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260–269.
- Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biology*, 4(3):e72.
- Wang, E. T., Kodama, G., Baldi, P., and Moyzis, R. K. (2006). Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proceedings of the National Academy of Sciences*, 103(1):135–140.
- Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2):256–276.
- Wellems, T. E. and Plowe, C. V. (2001). Chloroquine-resistant malaria. *The Journal of infectious diseases*, 184(6):770–776.
- Williamson, S. H., Hubisz, M. J., Clark, A. G., Payseur, B. A., Bustamante, C. D., and Nielsen, R. (2007). Localizing Recent Adaptive Evolution in the Human Genome. *PLoS Genetics*, 3(6):e90.
- Wiuf, C. and Hein, J. (1999). Recombination as a point process along sequences. *Theoretical Population Biology*, 55(3):248–259.

- Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, 16(2):97–159.
- Yang, Z. and Nielsen, R. (2000). Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models.
- Zhang, J. (2003). Evolution of the human ASPM gene, a major determinant of brain size. *Genetics*, 165(4):2063–2070.
- Zhang, J., Chiodini, R., Badr, A., and Zhang, G. (2011). The impact of next-generation sequencing on genomics. *Journal of genetics and genomics = Yi chuan xue bao*, 38(3):95–109.
- Zhang, Y. (2013). A Dynamic Bayesian Markov Model for Phasing and Characterizing Haplotypes in Next Generation Sequencing. *Bioinformatics*.
- Zhu, L. and Bustamante, C. D. (2005). A composite-likelihood approach for detecting directional selection from DNA sequence data. *Genetics*, 170(3):1411–1421.