Title
Population Genetics of Ancient and Modern DNA

Author
Malaspinas, Anna-Sapfo

Publication Date
2011

Peer reviewed|Thesis/dissertation

Population Genetics of Ancient and Modern DNA

by

Anna-Sapfo Malaspinas

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Integrative Biology

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division
of the
University of California, Berkeley

Committee in charge:
Professor Montgomery Slatkin, Chair
Professor Steven Neil Evans
Professor Rasmus Nielsen

Fall 2011

Population Genetics of Ancient and Modern DNA

Abstract

Population Genetics of Ancient and Modern DNA

by

Anna-Sapfo Malaspinas

Doctor of Philosophy in Integrative Biology

University of California, Berkeley

Professor Montgomery Slatkin, Chair

In this work, I develop computational tools focused around the utilization of DNA sequence data to address questions relative to forensic science, medical genetics, human evolution and ancient DNA.

First, I compute the theoretical probability that two individual profiles match by chance at two loci in a subdivided population. This question is of particular interest in forensic science, where DNA evidence has become a widespread tool of investigation and criminal conviction. I find that the effect of ignoring population subdivision can be unfavorable to the defendant, but that the two loci can essentially be treated as unlinked.

Second, I develop a method to identify genes that are interacting, or in epistasis, to produce a particular phenotype. Determining interacting genes is indeed of particular relevance in medical genetics to help map disease genes. I validate the method with simulations and demonstrate an improved performance over existing approaches. I also apply the method to recently available genomic data from domesticated dogs, identifying genes in epistastis for the hair length phenotype - thus representing candidate genes for functional validation.

Third, I use a summary statistic of DNA sequences, the site frequency spectrum, to estimate parameters of recent human history, and to characterize the potential event of admixture between Neanderthals and humans. I find evidence for recent gene flow between Neanderthals and Europeans, and to a lesser extent between Neanderthals and Africans.

Finally, I develop a likelihood method to jointly estimate the age and selection coefficient of an identified mutation, along with the population size, by using time serial samples. Such datasets are widespread in the fields of ancient DNA as well as experimental and viral evolution. I validate the method through simulations. I re-analyze a recent dataset for a locus coding for the distribution of black pigmentation in horses - and estimate that the allele far predates domestication, arising between 20,000 and 13,000 years ago.

à Barbara

Un seul être vous manque, et tout est dépeuplé.

Alphonse de Lamartine, 1820

Contents

Acknowledgments

This dissertation is perhaps also an opportunity for me to thank in writing the people who have been instrumental for making it happen.

First, I would like to thank my advisor, Monty Slatkin. Over the years he has created for us a very stimulating work environment, giving me the opportunity to work on several projects at the forefront of the field - but also giving me the necessary freedom and encouragement to explore related areas. I thank him for his patience, kindness and constant availability. Finally, his support has been essential for making this experience enjoyable.

I am also grateful to the other members of my dissertation committee, Rasmus Nielsen and Steve Evans. Rasmus Nielsen has always given me relevant and expeditious advice. I also benefited tremendously from my interaction with Steve Evans, who has always taken the time to answer my myriad of questions with extreme patience and clarity.

For each chapter of my dissertation, aside from the members of my committee, I feel indebted to several people. For chapter 2, I would like to thank Gerasimos Mataragkas, Weiwei Zhai, Philip Johnson, and Anand Patil for helpful discussions, and Anand Bhaskar for independently checking the results. For chapter 3, I would like to thank Anders Albrechtsen, Lior Pachter, Yun Song, and Bernd Sturmfels for helpful discussion, as well as Yu Zhang for his help with BEAM and Heidi Parker and Elaine Ostrander for providing the dog dataset. For chapter 4, I would like to thank Ryan Gutenkunst, Ed Green, Jeff Wall, Eric Durand, Emilia Huerta-Sanchez, and in particular Philip Johnson who read an earlier version of the chapter. For chapter 5, I would like to thank Fernando Perez and Andreas Malaspinas for mathematical advice, Philip Johnson and Emilia Huerta-Sanchez for helpful discussions, and Arne Ludwig for providing the horse dataset.

I am grateful to present and past members of my lab who have created a great familial atmosphere for work. Emilia Huerta-Sanchez, Philip Johnson, and Weiwei Zhai's unconditional support on all possible matters (scientific, personal, but also related to cultural acclimatization) has meant a great deal to me. I would also like to thank Celine Becquet, Bastien Bousseau, Paola Bronson, Eric Durand, Anna Ferrer-Admetlla, Vera Kaiser, Kirk Lohmueller, Chris Nasrallah, Joshua Schraiber, Owen Solberg, Nicolas Vinckenbosch, Beatriz Vicoso, and Melinda Yang, who have always been present for scientific discussions, moral support, and entertainment!

Several professors at Berkeley and Geneva have been influential. First the members of my qualifying exam committee, Yun Song for his mentorship through chapter 2 of my dissertation, John Huelsenbeck for his encouragement during my post-qualifying exam melancholy, Nipam Patel for biological insight and Tim White for reminding me that intellectual merit is not everything. I would also like to thank several professors for their outstanding classes: Joshua Bloom, Michael Jordan, James Sethian, and Bernd Sturmfels. I thank Andreas Malaspinas for reading a previous version of my dissertation. And finally, I am very grateful to two professors in particular for their guidance through academic matters: Evris Stephanou and Jean-Pierre Eckmann.

Mei Griebenow for IB and Brian McClendon for the DE have both been crucial with all administrative issues. I thank them for their competence and kindness through the years. I also wish to thank all of the staff at the international office. They helped me feel a bit less like a convicted criminal through the various visa procedures.

I am grateful to many friends and housemates from Berkeley, Geneva and Copenhagen, who were supportive and fun during the whole process. In particular, Doris Bachtrog, David Buckley, Myriam Boumediane, Claire Brawand, James Bullard, Sabine Cimasoni, Crystal Chaw, Roberta Damasceno, Christopher DiVittorio, Maria-Jose Fernandez, Greg Goldsmith, Peter Huggins, Laetitia Krummenacher, Bonnie Kirkpatrick, Benoit Kornmann, Elsa Lapouille, Kitsos Louis, Jay McEntee, Ida Moltke, Camille Morend, Agnès Michel, Clair Null, Joshua Paul, Ricardo Pereira, Joey Pakes, Juan Parra, Cathleen Poon, Daniel Richter, Leonore Rougemont, Sean Schoville, Meromit Singer, Yaron Singer, Owen Solberg, Josuke Tanaka, Caroline Uhler, Rachel Walsh, Nancy Wang, Cyrille Zbinden, and Lorric Ziegler.

Ευχαριστω και πάλι τους γονείς μου, τη Μαρία και τον Αντρέα και τα αδέρφια μου, τον Ορέστη, την Ιλιόνα και τον Ορφέα (και πιο πρόσφατα το Τζακ!). Ελπίζω να τους έχω όλους κοντα μου για πολλα πολλα χρόνια ακόμα.

Finally, I would like to thank Τζεφούλι, who has helped me through many discussions, but also for reading and correcting earlier versions of this dissertation. In general, I would probably still be finishing my first chapter if it was not for him.

Last but not least, I am indebted to several sources of funding who have made this dissertation possible. The Janggen-Poehn foundation has generously funded two years of my PhD. The Ernst and Lucie Schmidheiny foundation, with the help and advice of Johanne Patenaude, has funded another year of my graduate school. NIH, through a grant to my advisor, has funded the remaining time. Other institutes have funded conference travels, notably SMBE, IB with the Hansen fund and the Beim, Gray, Wiley, Umbson, and Resetko fellowships and the graduate division at Berkeley - I thank them as well.

Chapter 1

Introduction

Historically, population genetics was born amid a fierce battle between the biometricians, those who believed in continuous evolution, and the supporters of Mendel. As is often the case, the resolution came with the realization by the pioneers of population genetics - Fisher, Haldane, and Wright - that the opponents were both right (Provine, 2001). Some 80 years later, we still rely on the fundamental work of these scientists as a starting point for most evolutionary studies. Each chapter of the current work represents natural extensions to the models they originally developed. In fact, the questions addressed here already preoccupied the first population geneticists. Perhaps the main difference today is that the molecular biology underlying the processes is much better characterized. Thus, for exampe, we know the mechanisms behind dominance, we have identified Mendel's factors etc. Another fundamental difference is the abundance of available data to test evolutionary hypotheses. Indeed, at the beginning of the 20th century most relevant data came from breeding experiments - which were both costly and time consuming. Today, a whole genome can be obtained in a day and the cost is accessible to many labs worldwide. Thus, the last decades have clearly been a reality check on the theory that was developed at the beginning of the 20th century, but also for a refinement of this theory using modern biological insight. The current work is well cast in this framework. I investigate here aspects of the main evolutionary forces that were put forward in the early developments of the field. Broadly, in chapter 2 I consider the effects of drift and population subdivision for a finite population. The object of chapter 3 is gene interaction, one of the main discoveries leading to the synthesis of Darwinism and Mendelism. In chapter 4, I try to characterize gene flow between two populations. Finally in chapter 5, I explore the standard selection-drift models when applied to a specific type of data. I will now give a brief overview of each chapter. Note that chapters 2-5 describe results from collaborative projects done with different groups of collaborators.

In chapter 2, I generalize a recently introduced graphical framework to compute the probability that haplotypes or genotypes of two individuals drawn from a finite, subdivided population match. As in previous work I assume an infinite-alleles model. I focus on the case of a population divided into two subpopulations, but the underlying framework can be applied to a general model of population subdivision. I examine the effect of population subdivision on the match probabilities and the accuracy of the product rule which approximates multi-locus match probabilities as a product of one-locus match probabilities. I quantify the deviation from predictions of the product rule by R, the ratio of the multi-locus match

probability to the product of the one-locus match probabilities. I carry out the computation for two loci and find that ignoring subdivision can lead to underestimation of the match probabilities if the population under consideration actually has subdivision structure and the individuals originate from the same subpopulation. On the other hand, under a given model of population subdivision, I find that the ratio R for two loci is only slightly greater than 1 for a large range of symmetric and asymmetric migration rates. Keeping in mind that the infinite-alleles model is not the appropriate mutation model for STR loci, I conclude that, for two loci and biologically reasonable parameter values, population subdivision may lead to results that disfavor innocent suspects because of an increase in identity-by-descent in finite populations. On the other hand, for the same range of parameters, population subdivision does not lead to a substantial increase in linkage disequilibrium between loci. Those results are consistent with established practice.

In chapter 3, I describe a two-stage method for detecting epistasis by combining the traditionally used single-locus search with a search for multiway interactions. Indeed, rapid research progress in genotyping techniques have allowed large genome-wide association studies. Existing methods often focus on determining associations between single loci and a specific phenotype. However, a particular phenotype is usually the result of complex relationships between multiple loci and the environment. Our method is based on an extended version of Fisher's exact test. To perform this test, a Markov chain is constructed on the space of multidimensional contingency tables using the elements of a Markov basis as moves. I test our method on simulated data and compare it to a two-stage logistic regression method and to a fully Bayesian method, showing that I am able to detect the interacting loci when other methods fail to do so. Finally, I apply our method to a genome-wide data set consisting of 685 dogs and identify epistasis associated with canine hair length for four pairs of SNPs.

In chapter 4, I characterize the Neanderthal admixture with modern humans assuming a particular demographic scenario. Neanderthals are believed to be the closest evolutionary relatives of modern humans. But how exactly the Neanderthals disappeared, and the nature of their relationship to modern human, remains a widely disputed topic. In particular, despite the wealth of archaeological and anthropological knowledge, the admixture question is still an open question. Recently, a whole Neanderthal genome was sequenced and evidence for admixture was found. Nevertheless, specific demographic scenarios for admixture were not tested. I use a summary statistics of DNA data, the site frequency spectrum, to characterize the Neanderthal admixture with modern humans. In particular, I infer parameters for several simplistic demographic models between humans and Neanderthals. The models I test have in common a single admixture event between a human population (CEU or YRI) and the Neanderthals. I use an alignement of the high coverage trio data of the 1,000 genome project and the Neanderthal to infer the time of admixture and the amount of admixture. I find evidence of admixture between Europeans and Neanderthals, consistent with previous results. I also find evidence for a small amount of admixture (less than 1%) between Africans and Neanderthals. I estimate the time of admixture to be around 45,000 years, consistent with the archeaological record. I conclude that the joint site frequency spectrum is informative to test hypotheses for early human evolution.

In chapter 5, I develop a likelihood method to jointly estimate the selection coefficient and the age of an allele from time serial data. Recent advances in sequencing technologies have made available an ever-increasing amount of ancient genomic data. In particular, it is

now possible to target specific single nucleotide polymorphisms in several samples at different time points. Such time series data is also available in the context of experimental or viral evolution. Time-series data should allow for a more precise inference of population genetic parameters, and to test hypotheses about the recent action of natural selection. I assume a single panmictic population evolving through time and a constant selection coefficient. The transition probabilities are calculated by approximating the standard diffusion equation of the Wright-Fisher model with a one step process. I show that our method produces almost unbiased estimates. The power of the method is tested via simulations. Finally, the usefullness of the method is illustrated with an application to a locus, the *ASIP* locus, encoding coat color in horses, a pattern that has previously been linked with domestication.

Chapter 2

Match probabilities in a finite, subdivided population

Le dix-huitième siècle, c'est là une partie de sa gloire, a aboli la torture; le dix-neuvième siècle abolira la peine de mort. Victor Hugo, 1848.

2.1 Introduction

In forensic science, the analysis of DNA has become increasingly important. The multi-locus genotype, *the DNA profile*, of a biological sample from a crime scene is compared with the DNA profile of one or more suspects and often with the set of profiles in a large database. In the United States, the CODIS (Combined DNA Index System) is comprised of genotypes for 13 tetranucleotide microsatellite loci (`http://www.fbi.gov/hq/lab/html/codis1.htm`).

In the absence of laboratory error, a match between an individual's profile and that from the crime scene can be explained in one of two ways: either the individual is the source of the crime-scene sample or the individual is not the source but has the same profile by chance alone. The probability of such a match has been defined in different ways that also depend on the underlying population genetic model. In this paper we define the *match probability* as the probability of a match between two individuals drawn at random. We consider the cases where the two individuals come from the same subpopulation or from different subpopulations. Given that a genetic match is often sufficient to uphold a conviction (Song et al., 2009), the match probability plays an important role in the US and other judicial systems.

In US courts, the match probability is usually computed by following the recommendations of the second National Research Council Report (Committee on DNA Forensic Science: An Update, 1996), called NRCII. Recommendation 4.1 of the NRCII advocates for the use of the profile frequencies to compute the match probability. But if the subpopulation origin of the sample is known but not the allele frequencies for the specific subpopulation, NRCII recommends the use of the equation developed by Balding and Nichols (1994, 1995) (recommendation 4.2). The NRC report has been criticized in the scientific community (see for example Evett and Weir (1998); Balding (2005)). In particular it has been argued that the equation of Balding and Nichols (1994, 1995) should be used even in cases where the

origin of the sample is unknown. The latter equation describes the conditional probability that given an observed profile we find another individual with the same profile. This conditional match probability depends on a parameter, θ, which accounts for small deviations from Hardy-Weinberg frequencies caused by population subdivision or other deviations from random mating, and which is equivalent to F_{ST} in the population genetics literature. NRCII recommend a value of θ between 0.01 and 0.03 to account for the observed genotype frequencies in all known human populations, but others have argued that in some minority groups a value of 0.05 may be more appropriate (e.g. Balding (2005)). NRCII recommends that the multi-locus genotypic match probability be computed by multiplying the one-locus match probabilities, which, following convention in the forensics literature, we will call the *product rule*. For the 13 CODIS loci, the match probability of two unrelated individuals computed using the product rule is on the order of $10^{-14} \sim 10^{-15}$ (Song et al., 2009).

The use of the product rule remains controversial (Laurie and Weir, 2003; Bhaskar and Song, 2009). Because of the difficulty of analyzing multi-locus models in finite populations, there has been relatively little theoretical work on the conditions under which the product rule provides an accurate approximation to the match probabilities of multi-locus profiles. For a small number of loci, Laurie and Weir (2003) and Song and Slatkin (2007) computed the match probability in a finite randomly mating population and found that the deviations from the predictions of the product rule are small unless mutation rates are unrealistically high. Bhaskar and Song (2009) later generalized the graphical method of Song and Slatkin (2007) to compute the match probability of haploid genotypes for as many as 10 loci in a Wright-Fisher model and 13 loci in a Moran model. They showed that deviations from the predictions of the product rule are sensitive to the assumed mutation rate but, for mutation rates consistent with the observed levels of heterozygosity at CODIS loci, deviations from the predictions of the product rule were relatively small for up to 13 loci.

All of these theoretical studies of match probabilities assume an infinite-alleles model of mutation at each locus (i.e., every mutation leads to a new allele never seen before in the population). They predict the deviation from the product rule resulting from multi-locus identity-by-descent created by genetic drift in a finite population. These models cannot predict the overall deviations from the product rule in practice, however, because the loci used for forensic purposes are short tandem repeat (STR) loci for which the infinite-alleles model is not realistic. The same allele can be created more than once by mutation. A stepwise mutation model is more appropriate for STR loci but does not lead to a tractable theory of match probabilities even in a single population. The match probabilities calculated from the infinite-alleles model indicate the magnitude of deviations from the product rule expected because of the accumulated identity-by-descent, which is likely to be the primary effect of population subdivision.

Here, we extend the graphical framework derived by Song and Slatkin (2007) to explicitly account for population subdivision. We model two linked loci in two subpopulations that exchange migrants. We derive a system of coupled linear recurrence equations for the match probabilities. The equilibrium match probabilities are then found by assuming stationarity and solving the recurrence equations. As in the previous studies, we assume an infinite-alleles model. Our goals are to determine the effect of population subdivision on the match probabilities, and to study what assumptions about population subdivision lead to two-locus match probabilities that differ substantially from the predictions made using the product

rule. The model allows us to address two scenarios of practical interest.

1. *Same-subpopulation scenario*: Suppose there are two subpopulations α and β, and the source of the crime-scene sample and the defendant are both from the same subpopulation, say subpopulation α. In this case, the correct thing to do would be to compute the match probability between two individuals taken at random within the subpopulation α. However, the subpopulation origin of the individuals is not known and one considers the match of two individuals taken at random from the entire population assuming no subdivision, i.e., assuming random mating between all individuals of the whole population. This case corresponds to the situation described by Balding and Nichols (1994, 1995).

2. *Cryptic subdivision scenario*: One is interested in the match probability between two individuals randomly drawn from the entire population. He or she computes this probability assuming no population subdivision, but what is thought to be a single randomly mating population actually consists of two subpopulations with some gene flow between them.

This paper is organized as follows. In Section 2.2, we first describe the Wright-Fisher random mating model for a subdivided population and give a more precise definition of match probabilities. We then describe in Section 2.3 our extension of the aforementioned graphical framework to incorporate population structure. In Section 2.4, we present the results for the haplotypic and genotypic match probabilities for one and two loci for various migration rates. We show that the aforementioned same-subpopulation scenario can lead to substantial underestimation of the match probabilities, while both scenarios considered result in relatively small deviations from the product rule for reasonable mutation rates at the two loci.

2.2 Random mating model for a subdivided population and match probabilities

In this paper, we consider a population consisting of two subpopulations with migration. Provided below is a brief description of the assumed random mating model. See Table 2.1 for a summary of notation.

1. The total population is finite and of constant size N ($2N$ gametes).

2. Generations are non-overlapping.

3. The population has two subpopulations, α and β, with constant sizes N_α and N_β. Note that $N = N_\alpha + N_\beta$.

4. We assume a standard Wright-Fisher random union of gametes within each subpopulation as in Laurie and Weir (2003), extended to incorporate migration.

Table 2.1: Summary of Notation.

Notation	Signification
$2N$	The total number of gametes in each generation.
α, β	Subpopulation labels.
$2N_\alpha$, $2N_\beta$	Number of gametes in subpopulations α and β, respectively.
μ_i	Mutation rate at locus i per generation, per gamete (independent of the subpopulation label).
$m_{\alpha\beta}$, $m_{\beta\alpha}$	Backward migration rate from α to β, and from β to α, respectively, per generation, per gamete.
x_i	Allele at locus i of a gamete from an unspecified subpopulation.
$\boldsymbol{x}$	A haplotypic sequence $\boldsymbol{x} = x_1x_2$.
a_i	Allele at locus i of a gamete from subpopulation α.
$\boldsymbol{a}$	A haplotypic sequence $\boldsymbol{a} = a_1a_2$.
b_i	Allele at locus i of a gamete from subpopulation β.
$\boldsymbol{b}$	A haplotypic sequence $\boldsymbol{b} = b_1b_2$.
$x_i \equiv x_i'$	For locus i, alleles x_i and x_i' match.
$\boldsymbol{x} \equiv \boldsymbol{x}'$	For all loci i, allele x_i of $\boldsymbol{x}$ and allele x_i' of $\boldsymbol{x}'$ match.

5. Mutations occur at locus i with probability μ_i per generation and are independent of the subpopulation, and of other loci. Each mutation event produces a new allele never seen before in the population, i.e., we assume an infinite-alleles model.

6. Recombination may occur each generation between loci with probability r per generation for gametes within a subpopulation, at a rate independent of the subpopulation.

7. Forward in time, the probability a particular individual from subpopulation α migrates to subpopulation β is $m_{\beta\alpha}N_\beta/N_\alpha$ per generation. The analogous probability for migration from β to α is $m_{\alpha\beta}N_\alpha/N_\beta$. The parameters $m_{\alpha\beta}$ and $m_{\beta\alpha}$ are *backward* migration fractions.

8. All loci are neutral.

9. Migration is conservative (Nagylaki, 1980), i.e., $m_{\alpha\beta}N_\alpha = m_{\beta\alpha}N_\beta$, or equivalently $N_\alpha = \frac{m_{\beta\alpha}}{m_{\alpha\beta}+m_{\beta\alpha}}N$ and $N_\beta = \frac{m_{\alpha\beta}}{m_{\alpha\beta}+m_{\beta\alpha}}N$.

A new generation is obtained by drawing individuals from the previous generation. For two loci, a gamete from an unspecified subpopulation is denoted by $\boldsymbol{x} = x_1x_2$, where x_i denotes the allele at locus i. Further, we denote by $\boldsymbol{a} = a_1a_2$ and $\boldsymbol{b} = b_1b_2$ gametes from

subpopulation α and β, respectively. Considering only two loci, our assumptions imply the following scheme for the generation of offsprings.

1. Two gametes, $\boldsymbol{x}^1 = x_1^1 x_2^1$ and $\boldsymbol{x}^2 = x_1^2 x_2^2$, are drawn with replacement from within a subpopulation.

2. The two gametes drawn recombine with probability r to create the offspring gamete. With probability $\frac{1}{2}(1-r)$, $\frac{1}{2}(1-r)$, $\frac{1}{2}r$, $\frac{1}{2}r$, the offspring gametes are, respectively, $x_1^1 x_2^1$, $x_1^2 x_2^2$, $x_1^1 x_2^2$, $x_1^2 x_2^1$.

3. As described above, each offspring gamete produced in a subpopulation may migrate to the other subpopulation with a certain probability.

4. At each locus i, each offspring gamete may have undergone mutation with probability μ_i, independent of the subpopulation. Every mutation creates a new allele never seen before in the population.

The above procedure is repeated N_α times in subpopulation α and N_β times in subpopulation β.

We are interested in the one-locus and the two-locus random match probability for the haplotypic and genotypic case. We define it as the probability that two individuals drawn at random match at 1 or 2 loci, respectively (Laurie and Weir, 2003; Song and Slatkin, 2007). We will consider three different cases; the two haploid or diploid individuals come from the same subpopulation (either α or β) or each of them comes from a different subpopulation. We assume that the alleles of each locus for each individual are known. Our definition differs from the match probability defined by some authors (e.g. Weir (2004)) in that it is not a conditional match probability.

2.3 Description of the graphical framework

Song and Slatkin (2007) derive recursion equations by representing match relations as graphs and by performing operations on them. The four parts in the above-mentioned mating scheme will be captured by four different operations on graphs, namely, vertex-merge, vertex-split, vertex-recolor, and vertex-count, respectively. Our graphical method is the same as theirs', except that vertices are now colored to distinguish gametes from different subpopulations.

2.3.1 Correspondence between match probabilities and graphs

The graphs we consider are undirected graphs with edge labels and vertex colors. Each vertex represents a gamete and the color of the vertex indicates the subpopulation to which the gamete belongs. Two vertices are joined by an edge (or an arc) labeled i if the associated gametes match at locus i. Three examples of match graphs are shown in Figure 2.1. We sometimes omit edge labels for convenience of drawing, in which case we adopt the convention of drawing locus 1 (respectively, locus 2) edges above (respectively, below) the vertices.

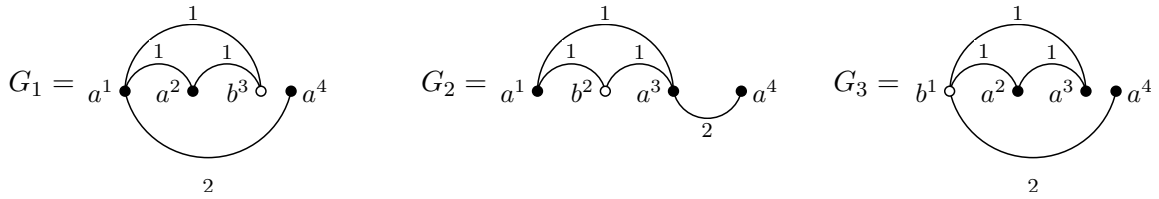

Figure 2.1: Example of graphical representation of match probabilities. Filled circles represent gametes from subpopulation α, while open circles represent gametes from subpopulation β. The graph G_1 corresponds to the probability $\mathbb{P}(a_1^1 \equiv a_1^2,\ a_1^1 \equiv b_1^3,\ a_1^2 \equiv b_1^3,\ a_2^1 \equiv a_2^4)$ that three gametes a^1, a^2 and b^3 match at locus 1, and that two gametes a^1 and a^4 match at locus 2. The graph G_2 corresponds to $\mathbb{P}(a_1^1 \equiv b_1^2,\ a_1^1 \equiv a_1^3,\ b_1^2 \equiv a_1^3,\ a_2^3 \equiv a_2^4)$ and the graph G_3 corresponds to $\mathbb{P}(b_1^1 \equiv a_1^2,\ b_1^1 \equiv a_1^3,\ a_1^2 \equiv a_1^3,\ b_2^1 \equiv a_2^4)$. Note that the superscripts 1, 2, 3, and 4 allow us to refer to specific vertices, but otherwise they have no meaning. In particular, two graphs obtained by permuting superscripts are equivalent. For example, graphs G_1 and G_2 are isomorphic, while G_1 and G_3 are not.

As pointed out in Song and Slatkin (2007), two match probabilities are equal under random mating if they are related by some permutation of gamete labels. More formally, two graphs are equivalent if they are isomorphic as vertex-colored edge-labeled graphs. For example in Figure 2.1, G_1 and G_2 are isomorphic, but neither G_1 nor G_2 is isomorphic to G_3. If we permute the superscripts on gametes as $1 \to 3$, $2 \to 1$, $3 \to 2$, $4 \to 4$, we can transform G_1 into G_2, but there exists no permutation that can transform G_1 (or G_2) into G_3.

In the graphical framework, the goal is to relate an offspring graph G^O (corresponding to a match relationship at time t) to a set of parental graphs $G_1^P, G_2^P, \ldots$ (corresponding to a set of match relationships at time $t-1$). Let $\mathcal{A}$ denote the set of all possible ancestries (i.e., a set of mutation, migration, recombination and coalescent events) one-generation back in time for the gametes in G^O. Then, the probability of G^O can be decomposed as

$$\mathbb{P}(G^O) = \sum_{A \in \mathcal{A}} \mathbb{P}(G^O \mid A)\mathbb{P}(A). \tag{2.3.1}$$

For many ancestries $A \in \mathcal{A}$, the conditional probability $\mathbb{P}(G^O \mid A)$ may be zero; our method considers only the set of ancestries $A \in \mathcal{A}$ with positive $\mathbb{P}(G^O \mid A)$ and sums over that set. We say that an ancestry A is *valid* if $\mathbb{P}(G^O \mid A) > 0$. Given a particular valid ancestry $A_k \in \mathcal{A}$, there corresponds a parental graph G_k^P such that $\mathbb{P}(G_k^P) = \mathbb{P}(G^O \mid A_k)$. At stationarity, isomorphic parental and offspring graphs have the same probability, and hence we can construct a closed system of linear equations by repeatedly using (2.3.1). We assume that mutation (U), migration (M), recombination (R) and coalescence (C) events are independent of each other, and hence we can decompose the probability of a particular one-generation ancestry A_k into four parts as $\mathbb{P}(A_k) = Q_k^U Q_k^M Q_k^R Q_k^C$, with the superscript denoting the type of event.

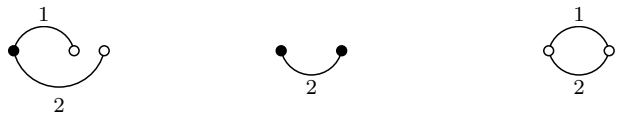

Figure 2.2: Graphical representation of two-locus match probabilities. From left to right, Q^U is equal to $(1-\mu_1)^2(1-\mu_2)^2$, $(1-\mu_2)^2$, and $(1-\mu_1)^2(1-\mu_2)^2$, respectively.

2.3.2 Operations on graphs

We now describe a graphical method to find, for a given offspring graph G^O, the set of all valid one-generation ancestries A. Further, we describe the computation of Q_k^U, Q_k^M, Q_k^R, and Q_k^C for each ancestral A_k.

Mutation or vertex count

Since we consider an infinite-alleles model, if two offspring gametes match at a particular locus, then their parental gametes in the previous generation must also match at that locus. Moreover, no mutation could have happened at that locus in producing the offspring gametes.

If n gametes match at locus i in the offspring graph G^O, an overall factor of $(1-\mu_i)^n$ will contribute to Q_k^U. More generally, for every valid ancestry A_k, we obtain $Q_k^U = \prod_{i=1}^{2}(1-\mu_i)^{\delta_i(G^O)}$, where $\delta_i(G^O)$ is the number of vertices in G^O that are involved in a match relation for locus i (i.e., the number of vertices incident with an edge corresponding to locus i). Examples are provided in Figure 2.2. This procedure is referred to as a "vertex-count" operation.

The intermediate graph resulting from a mutation event or a vertex-count operation is identical to the original graph. We will denote by G^M, the resulting graph to refer to the fact that the next operation corresponds to a migration event.

2.3.3 Migration or vertex recolor

The graphs considered in Song and Slatkin (2007) have uncolored vertices. To take into account the population subdivision, we color the vertices. A migration event corresponds to a "vertex recoloring" operation in the graphical framework (see Figure 2.3).

Consider a graph with n_α vertices from subpopulation α and n_β vertices from subpopulation β. Since the migration rates are typically small, we linearize the probability of migration per generation, and therefore $Q_k^M = 1-n_\alpha m_{\alpha\beta}-n_\beta m_{\beta\alpha}$ if no migration happens, $Q_k^M = m_{\alpha\beta}$ if a gamete from α migrates to β or $Q_k^M = m_{\beta\alpha}$ if a gamete from β migrates to α.

The graph resulting from a recoloring operation is an intermediate graph denoted by G^R, indicating that the next operation is for a recombination event.

2.3.4 Recombination or vertex split

Going backward in time, if a recombination event occurs in an offspring gamete, then two parental gametes must be tracked, each contributing a new vertex. Recombination is relevant only if we consider two or more loci.

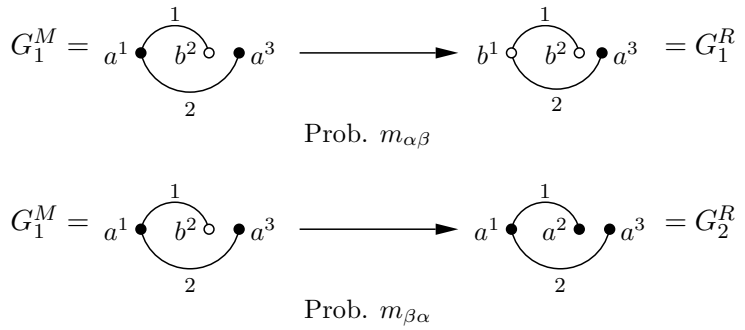

Figure 2.3: Illustration of vertex-recolor operations on match graphs for two loci. In the top figure, vertex a^1 is recolored and the label changes to b^1; this event occurs with probability $m_{\alpha\beta}$. In the bottom figure, vertex b^2 is recolored (and the label changes to a^2) with probability $m_{\beta\alpha}$.

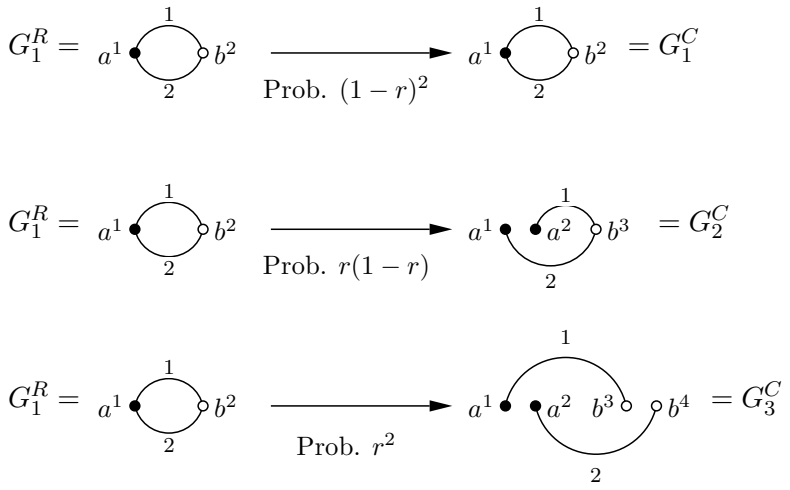

Figure 2.4: Illustration of vertex-split operations on match graphs for two loci. In G_1^R, vertices a^1 and b^2 each have δ-degree 2, so there can be up to two split operations. The top figure corresponds to there being no split. In the middle figure only vertex a^1 is split, while both a^1 and b^2 are split in the bottom figure.

The operation on graphs that corresponds to recombination is a "vertex-split" operation. Any vertex with at least two incident edges with different labels can undergo a split operation. We can associate a number $\delta(v)$ (the δ-degree) to each vertex v, such that for a given match graph the δ-degree is the number of loci at which the vertex is involved in a match relationship. Denote by n the number of vertices with δ-degree larger than one. Hence, if s out of the n vertices split, then $Q_k^R = r^s(1-r)^{n-s}$. See Figure 2.4 for an example.

The graph obtained from a vertex-split operation is an intermediate graph denoted by G^C. The next operation to consider corresponds to coalescent events.

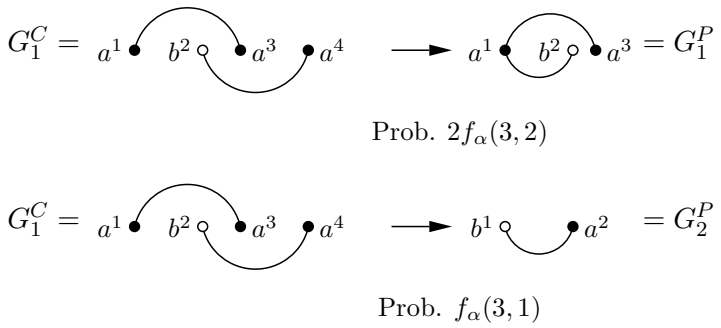

Figure 2.5: Illustration of vertex-merge operations on match graphs for two loci. In the top figure, there are two different vertex-merge operations that can transform G_1^C into G_1^P, explaining the factor of 2 in the associated probability. Specifically, a^1 and a^4 can be merged into a single vertex, or a^3 and a^4 can be merged. In the bottom figure, a^1, a^3 and a^4 are merged into a single vertex. Vertices from different subpopulations cannot be merged.

2.3.5 Coalescence or vertex merge

Finally, because parental gametes are drawn with replacement, a parental gamete may be drawn more than once. Going backward in time, this means that two (or more) vertices in G^C may undergo a "merge operation," provided that the vertices have the same color. Adopting Convention 1 of Song and Slatkin (2007), merging a set of vertices entails removing all edges between the vertices in that set. Any set of vertices in the same subpopulation can merge. The probability of such a merge is the probability of drawing the same gamete two or more times from the same subpopulation. Such a merged graph will be called a parental graph G^P.

The probability of a merge event given a graph G^C is almost identical to Equation 2 of Song and Slatkin (2007). Since two gametes from different subpopulations cannot be merged, each subpopulation can be considered independently. Having n_α (respectively, n_β) vertices of type α (respectively, β) before the merge operation and j_α (respectively, j_β) after merging, implies that j_α (respectively, j_β) distinct gametes were drawn. The probability of such an event is given by

$$Q_k^C = f_\alpha(n_\alpha, j_\alpha) f_\beta(n_\beta, j_\beta),$$

where

$$f_\alpha(n_\alpha, j_\alpha) = \frac{2N_\alpha(2N_\alpha - 1)\cdots(2N_\alpha - j_\alpha + 1)}{(2N_\alpha)^{n_\alpha}},$$

and $f_\beta(n_\beta, j_\beta)$ is similarly defined with N_α replaced with N_β.

Note that two different merge operations may produce isomorphic graphs. Isolated vertices are not involved in any match relationship, and hence can be ignored. A graph containing only isolated vertices has probability one. See Figure 2.5 for an example.

2.3.6 Summary: count, recolor, split, and merge

There are four graphical operations corresponding to different evolutionary events: vertex-count, vertex-recolor, vertex-split, and vertex-merge. The first three operations produce

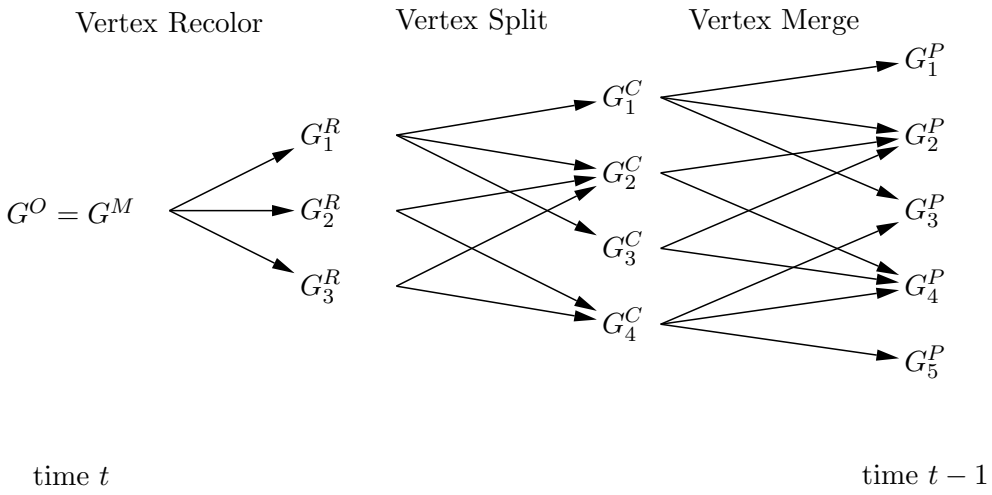

Figure 2.6: Schematic summary of the graphical approach of Song and Slatkin (2007) modified to account for population subdivision. Our notation reflects the operation to be performed: the vertices of G^M will be recolored to account for migration, the vertices of G_i^R will be split to account for recombination and finally the vertices of G_i^C will be merged to account for coalescence. The offspring graph G^O on the left can be written as a linear combination of the resulting parental graphs, G_i^P, on the right. The coefficients correspond to all possible single generation evolutionary histories, starting with the offspring graph.

intermediate graphs G^M, G^R, and G^C. They are used to relate an offspring graph G^O representing a match probability at time t to a set of parental match graphs G_k^P at time $t-1$. A schematic summary (corresponding to Figure 8 of Song and Slatkin 2007) is shown in Figure 2.6. At stationarity, the match probabilities are independent of t.

2.3.7 An example of a closed system of equations

Illustrated in Figure 2.7 is an example of a closed system of recurrence equations in our graphical framework. It is for the simplest case (i.e., the one-locus haplotypic match), involving three equations and three unknown variables. Because we consider only a single locus, no vertex-split operations are involved. The one-locus match probability $\mathbb{P}_h(x_i \equiv x_i')$ appears as one of the unknown variables in Figure 2.7, and solving the coupled equations allows us to compute that stationary probability.

Note that, when $\mu_i = 0$ for all loci i, every match probability is equal to one, and, therefore, the right-hand side of each equation in Figure 2.7 must sum to one. This can be easily verified. Such consistency conditions are useful for checking that the coefficients in the recurrence equations have been determined correctly.

2.4 Results on match probabilities

We are interested in computing two quantities: the two-locus match probability (denoted by $\mathbb{P}_h$ in the haplotypic case and by $\mathbb{P}_g$ in the genotypic case) computed without assuming

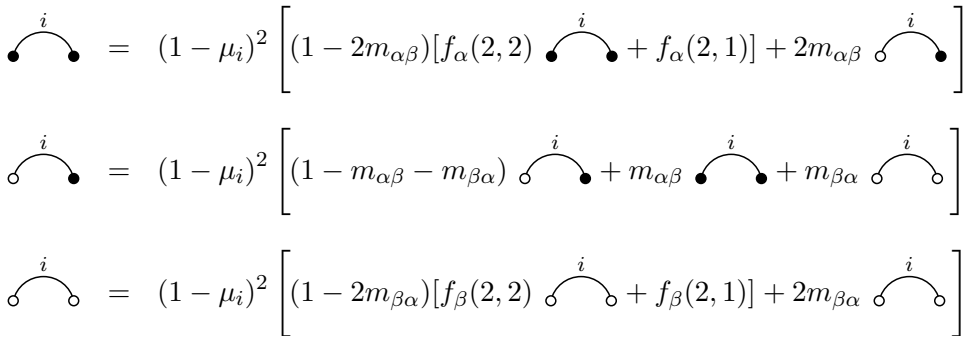

Figure 2.7: Graphical representation of the closed system of equations involving one-locus match probabilities for locus i. Recall that the probability of migration is linearized.

independence of the loci, and the ratio (denoted by R_h for the haplotypic case and by R_g for the genotypic case) of that probability to the approximate two-locus match probability obtained by multiplying the one-locus match probabilities. Under linkage equilibrium, the multi-locus match probability is given by the product of the one-locus match probabilities at the loci involved. Therefore, assuming linkage equilibrium the ratios R_h and R_g are equal to one. If the ratios are greater than 1, the actual match probabilities exceed those obtained by assuming the product rule because of linkage disequilibrium created between unlinked loci by the interaction of mutation and genetic drift.

To examine the effect of population subdivision on the above quantities, we present below some numerical results for both haplotypic and genotypic cases. We implemented our graphical method in *Python* to generate the system of equations, and then used *Mathematica* to solve the system. For simplicity, we present results for the case of the same mutation rate for all loci, i.e., $\mu_i = \mu$ for all i.

2.4.1 Haplotypic match probability

One-locus haplotypic match

The system of equations shown in Figure 2.7 can be used to obtain the haplotypic match probability $\mathbb{P}_h(x_i \equiv x_i')$ for locus i. Note that the ratio R_h is equal to 1 by definition.

Numerical values for one-locus haplotypic match probabilities are given in Table 2.2. The population size is set to $N = 10,000$, which is an approximate long-term effective population size of humans (Nei M. and Graur D., 1984; Harding et al., 1997; Harpending et al., 1998). The two mutation rates used, 10^{-4} and 10^{-3}, correspond to expected heterozygosities in a randomly mating population of $4N\mu/(1 + 4N\mu) = 0.8$ and 0.976. In the case of symmetric migration, the range of migration rates from 10^{-6} to 10^{-2} respectively correspond to 0.02 to 200 migrants per generation. Shown in this and later tables are 1) the F_{ST} value, 2) the match probability for two different individuals sampled randomly from the same subpopulation, and 3) the match probability of two different individuals chosen at random without regard to subpopulation. The last probability, denoted by $\mathbb{P}^{\text{random}}$, is a weighted average defined as

follows:

$$\mathbb{P}_h^{\text{random}}(x \equiv x') = \frac{N_\alpha^2}{N^2}\mathbb{P}_h(a \equiv a') + \frac{2N_\alpha N_\beta}{N^2}\mathbb{P}_h(a \equiv b') + \frac{N_\beta^2}{N^2}\mathbb{P}_h(b \equiv b'), \tag{2.4.2}$$

where a and a' (respectively, b and b') denote alleles from subpopulation α (respectively, β). The version of F_{ST} we adopt is

$$F_{ST} = \frac{\frac{N_\alpha}{N} \cdot \mathbb{P}_h(a_i \equiv a_i') + \frac{N_\beta}{N} \cdot \mathbb{P}_h(b_i \equiv b_i') - \mathbb{P}_h^{\text{random}}(x \equiv x')}{1 - \mathbb{P}_h^{\text{random}}(x \equiv x')}.$$

In the asymmetric case, the within-subpopulation match probability decreases as the subpopulation size increases. In the symmetric case, the more isolated the subpopulations, the higher the within-subpopulation and the smaller the between-subpopulations match probabilities. The effect of subdivision on the match probabilities is further discussed below in Section 2.4.2.

Two-locus haplotypic match

The two-locus system has a total of 26 match relations. The corresponding graphs are illustrated in Figure 2.8. Given two gametes $\boldsymbol{x} = x_1x_2$ and $\boldsymbol{x'} = x_1'x_2'$, the haplotypic match probability is $\mathbb{P}_h(\boldsymbol{x} \equiv \boldsymbol{x'})$. If the loci are independent, then the product rule holds, yielding $\mathbb{P}_h(x_1 \equiv x_1')\mathbb{P}_h(x_2 \equiv x_2')$ for the match probability. To quantify the departure from the product rule, we consider the following ratio:

$$R_h(\boldsymbol{x} \equiv \boldsymbol{x'}) = \frac{\mathbb{P}_h(\boldsymbol{x} \equiv \boldsymbol{x'})}{\mathbb{P}_h(x_1 \equiv x_1')\mathbb{P}_h(x_2 \equiv x_2')}.$$

Both the match probability $\mathbb{P}_h$ and the ratio R_h depend on the subpopulations to which the gametes $\boldsymbol{x}$ and $\boldsymbol{x'}$ belong. There are three different cases. 1) They both belong to subpopulation α. 2) They both belong to subpopulation β. 3) One belongs to subpopulation α, while the other belongs to subpopulation β. Graphical representations of $\mathbb{P}_h$ for the three cases are illustrated in Figures 2.9.

Numerical values of match probabilities $\mathbb{P}_h$ and ratios R_h are provided in Table 2.3. The same parameters were used as in the one-locus case were used. In all cases, R_h is not much greater than 1, even for the smallest migration rate we considered. Therefore, the product rule provides an adequate approximation for computing the match probabilities.

2.4.2 Genotypic match probability

One-locus genotypic match probability

Consider two unordered pairs $\{x_i, \bar{x}_i\}$ and $\{x_i', \bar{x}_i'\}$ of alleles. The probability of genotypic match at locus i is denoted $\mathbb{P}_g(\{x_i, \bar{x}_i\} \equiv \{x_i', \bar{x}_i'\})$. There are two possible ways to have a one-locus genotypic match: 1) $x_i \equiv x_i'$ and $\bar{x}_i \equiv \bar{x}_i'$, or 2) $x_i \equiv \bar{x}_i'$ and $\bar{x}_i \equiv x_i'$. However, these possibilities are not mutually exclusive, and as explained in Song and Slatkin (2007), the probability of a genotypic match can be computed using the inclusion-exclusion principle.

Table 2.2: One-locus haplotypic match probabilities for $N = 10,000$.

							$\mathbb{P}_h(x_i = x_i')$		
	μ	$m_{\alpha\beta}$	$m_{\beta\alpha}$	N_α	N_β	F_{ST}	Within α	Within β	Random
Sym	1×10^{-4}	1×10^{-2}	1×10^{-2}	5000	5000	0.001	2.01×10^{-1}	2.01×10^{-1}	2.00×10^{-1}
	1×10^{-4}	1×10^{-4}	1×10^{-4}	5000	5000	0.077	2.50×10^{-1}	2.50×10^{-1}	1.87×10^{-1}
	1×10^{-4}	1×10^{-6}	1×10^{-6}	5000	5000	0.197	3.31×10^{-1}	3.31×10^{-1}	1.67×10^{-1}
Asym	1×10^{-4}	1×10^{-3}	1×10^{-4}	909	9091	0.018	3.30×10^{-1}	1.99×10^{-1}	1.97×10^{-1}
	1×10^{-4}	1×10^{-5}	1×10^{-4}	9091	909	0.064	2.06×10^{-1}	5.97×10^{-1}	1.90×10^{-1}
No sub	1×10^{-4}	–		10000		–	2.00×10^{-1}		
Sym	1×10^{-3}	1×10^{-2}	1×10^{-2}	5000	5000	0.001	2.54×10^{-2}	2.54×10^{-2}	2.43×10^{-2}
	1×10^{-3}	1×10^{-4}	1×10^{-4}	5000	5000	0.020	4.38×10^{-2}	4.38×10^{-2}	2.39×10^{-2}
	1×10^{-3}	1×10^{-6}	1×10^{-6}	5000	5000	0.024	4.75×10^{-2}	4.75×10^{-2}	2.38×10^{-2}
Asym	1×10^{-3}	1×10^{-3}	1×10^{-4}	909	9091	0.011	1.26×10^{-1}	2.54×10^{-2}	2.41×10^{-2}
	1×10^{-3}	1×10^{-5}	1×10^{-4}	9091	909	0.019	2.65×10^{-2}	2.00×10^{-1}	2.39×10^{-2}
No sub	1×10^{-3}	–		10000		–	2.44×10^{-2}		

The probability of mutation per generation per gamete is assumed to be μ for all loci. Subpopulations α and β have $2N_\alpha$ and $2N_\beta$ gametes, respectively, and $m_{\alpha\beta}$ denotes the backward migration fraction from α to β; $m_{\beta\alpha}$ is similarly defined. "Sym" indicates symmetric migration; "Asym" indicates asymmetric migration; and "No sub" indicates a single randomly mating population. "Within α" and "Within β" indicate the match probabilities for two different individuals drawn randomly from the same subpopulation. "Random" indicates the match probability for two different individuals drawn randomly from the entire combined population, computed as described in (2.4.2).

Table 2.3: Two-locus haplotypic match probabilities and their ratios R_h to the approximate two-locus haplotypic match probabilities obtained using the product rule.

							$\mathbb{P}_h(\boldsymbol{x} \equiv \boldsymbol{x}')$			$R_h(\boldsymbol{x} \equiv \boldsymbol{x}')$		
	μ	$m_{\alpha\beta}$	$m_{\beta\alpha}$	N_α	N_β	F_{ST}	Within α	Within β	Random	Within α	Within β	Random
Sym	1×10^{-4}	1×10^{-2}	1×10^{-2}	5000	5000	0.001	4.03×10^{-2}	4.03×10^{-2}	3.99×10^{-2}	1.001	1.001	1.000
	1×10^{-4}	1×10^{-4}	1×10^{-4}	5000	5000	0.077	6.25×10^{-2}	6.25×10^{-2}	3.91×10^{-2}	1.000	1.000	1.000
	1×10^{-4}	1×10^{-6}	1×10^{-6}	5000	5000	0.197	1.10×10^{-1}	1.10×10^{-1}	5.48×10^{-2}	1.000	1.000	1.000
Asym	1×10^{-4}	1×10^{-3}	1×10^{-4}	909	9091	0.018	1.09×10^{-1}	3.98×10^{-2}	3.91×10^{-2}	1.002	1.000	1.000
	1×10^{-4}	1×10^{-5}	1×10^{-4}	9091	909	0.064	4.25×10^{-2}	3.57×10^{-1}	3.93×10^{-2}	1.000	1.001	1.000
No sub	1×10^{-4}	–		10000		–	4.00×10^{-2}			1.000		
Sym	1×10^{-3}	1×10^{-2}	1×10^{-2}	5000	5000	0.001	6.78×10^{-4}	6.78×10^{-4}	6.09×10^{-4}	1.048	1.048	1.027
	1×10^{-3}	1×10^{-4}	1×10^{-4}	5000	5000	0.020	1.95×10^{-3}	1.95×10^{-3}	9.81×10^{-4}	1.016	1.016	1.016
	1×10^{-3}	1×10^{-6}	1×10^{-6}	5000	5000	0.024	2.29×10^{-3}	2.29×10^{-3}	1.14×10^{-3}	1.013	1.013	1.013
Asym	1×10^{-3}	1×10^{-3}	1×10^{-4}	909	9091	0.011	1.61×10^{-2}	6.65×10^{-4}	7.08×10^{-4}	1.012	1.027	1.024
	1×10^{-3}	1×10^{-5}	1×10^{-4}	9091	909	0.019	7.19×10^{-4}	4.01×10^{-2}	9.27×10^{-4}	1.025	1.003	1.017
No sub	1×10^{-3}	–		10000		–	6.09×10^{-4}			1.027		

See the caption of Table 2.2 for notational convention. The total population size is assumed to be $N = 10,000$ and the loci are assumed to be unlinked (i.e., $r = \frac{1}{2}$).

Table 2.4: One-locus genotypic match probabilities for $N = 10,000$.

	μ	$m_{\alpha\beta}$	$m_{\beta\alpha}$	N_α	N_β	F_{ST}	$\mathbb{P}_g(\{x_i, \bar{x}_i\} \equiv \{x_i', \bar{x}_i'\})$		
							Within α	Within β	Random
Sym	1×10^{-4}	1×10^{-2}	1×10^{-2}	5000	5000	0.001	6.72×10^{-2}	6.72×10^{-2}	6.65×10^{-2}
	1×10^{-4}	1×10^{-4}	1×10^{-4}	5000	5000	0.077	1.00×10^{-1}	1.00×10^{-1}	6.48×10^{-2}
	1×10^{-4}	1×10^{-6}	1×10^{-6}	5000	5000	0.197	1.65×10^{-1}	1.65×10^{-1}	8.26×10^{-2}
Asym	1×10^{-4}	1×10^{-3}	1×10^{-4}	909	9091	0.018	1.65×10^{-1}	6.64×10^{-2}	6.53×10^{-2}
	1×10^{-4}	1×10^{-5}	1×10^{-4}	9091	909	0.064	7.05×10^{-2}	4.46×10^{-1}	6.53×10^{-2}
No sub	1×10^{-4}	–		10000		–	6.67×10^{-2}		
Sym	1×10^{-3}	1×10^{-2}	1×10^{-2}	5000	5000	0.001	1.26×10^{-3}	1.26×10^{-3}	1.16×10^{-3}
	1×10^{-3}	1×10^{-4}	1×10^{-4}	5000	5000	0.020	3.67×10^{-3}	3.67×10^{-3}	1.85×10^{-3}
	1×10^{-3}	1×10^{-6}	1×10^{-6}	5000	5000	0.024	4.31×10^{-3}	4.31×10^{-3}	2.15×10^{-3}
Asym	1×10^{-3}	1×10^{-3}	1×10^{-4}	909	9091	0.011	2.82×10^{-2}	1.26×10^{-3}	1.33×10^{-3}
	1×10^{-3}	1×10^{-5}	1×10^{-4}	9091	909	0.019	1.37×10^{-3}	6.66×10^{-2}	1.68×10^{-3}
No sub	1×10^{-3}	–		10000		–	1.16×10^{-3}		

See the caption of Table 2.2 for notational convention.

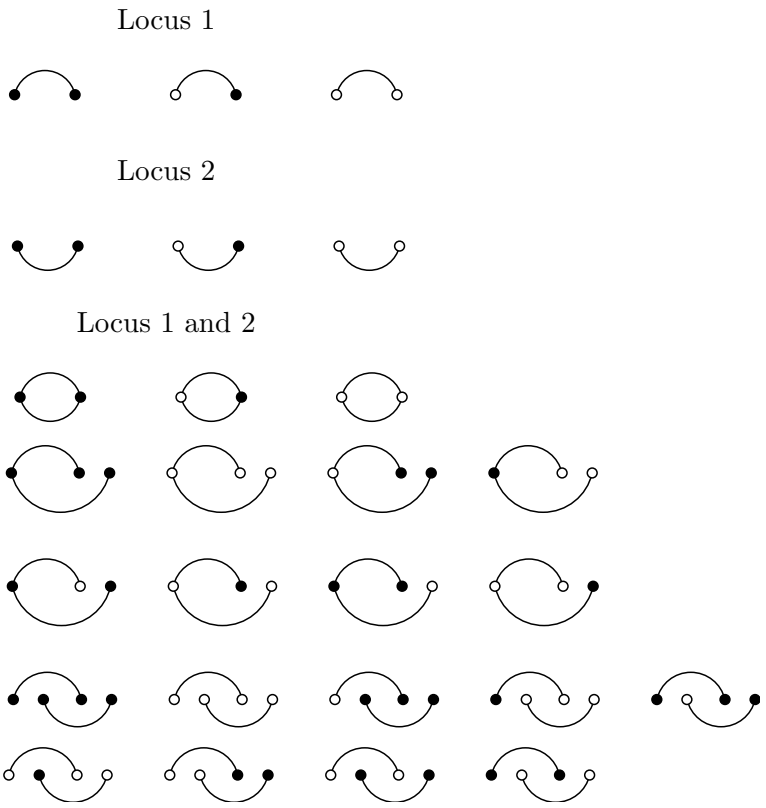

Figure 2.8: The match graphs involved in the two-locus haplotypic match probability. These probabilities form a closed system of equations comprising 26 variables and 26 equations.

$$\mathbb{P}_h(\boldsymbol{a} \equiv \boldsymbol{a}') =$$

$$\mathbb{P}_h(\boldsymbol{b} \equiv \boldsymbol{b}') =$$

$$\mathbb{P}_h(\boldsymbol{a} \equiv \boldsymbol{b}) =$$

Figure 2.9: The match graphs corresponding to the two-locus haplotypic match probability for two gametes from subpopulation α, two from subpopulation β and one from subpopulation α and one from β.

In the graphical framework, the probability of a one-locus genotypic match is as shown in Figure 2.10. The system of equations for the one-locus genotypic match case involves 18 variables (and 18 equations). Numerical values are reported in Table 2.4.

The conclusions are similar to those for the one-locus haplotypic case. One noticeable difference is that the match probabilities are smaller than for haplotypic matches.

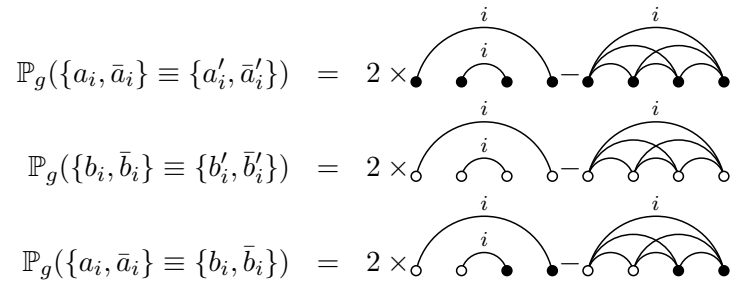

Figure 2.10: The one-locus genotypic match probabilities for two pairs of alleles at locus i.

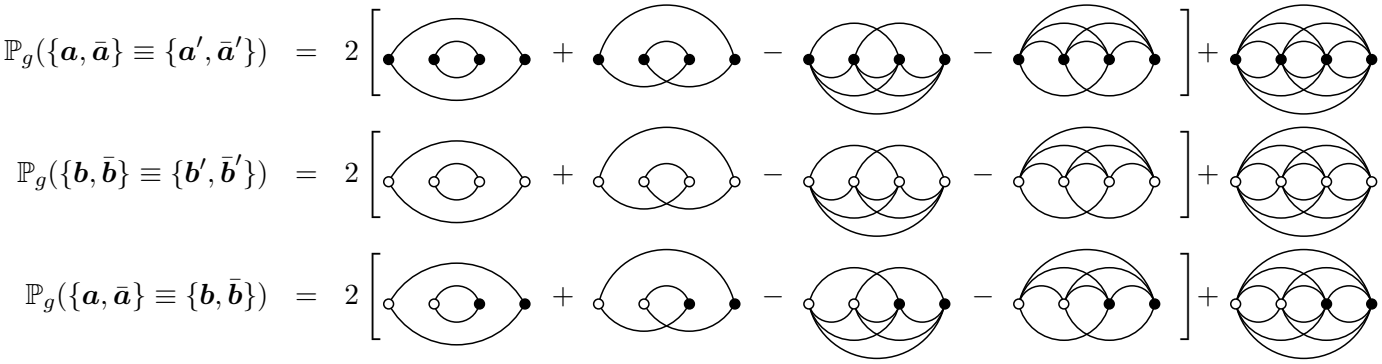

Figure 2.11: The two-locus genotypic match probabilities for two pairs of two-locus gametes.

Two-locus genotypic match probability

Consider two unordered pairs $\{\boldsymbol{x}_i, \bar{\boldsymbol{x}}_i\}$ and $\{\boldsymbol{x}_i', \bar{\boldsymbol{x}}_i'\}$ of gametes. The probability of two-locus genotypic match is denoted by $\mathbb{P}_g(\{\boldsymbol{x}_i, \bar{\boldsymbol{x}}_i\} \equiv \{\boldsymbol{x}_i', \bar{\boldsymbol{x}}_i'\})$. The match probability under the product rule is $\mathbb{P}_g(\{x_1, \bar{x}_1\} \equiv \{x_1', \bar{x}_1'\}) \times \mathbb{P}_g(\{x_2, \bar{x}_2\} \equiv \{x_2', \bar{x}_2'\})$. To quantify the departure from the product rule, we compute the following ratio:

$$R_g(\{\boldsymbol{x}_i, \bar{\boldsymbol{x}}_i\} \equiv \{\boldsymbol{x}_i', \bar{\boldsymbol{x}}_i'\}) = \frac{\mathbb{P}_g(\{\boldsymbol{x}_i, \bar{\boldsymbol{x}}_i\} \equiv \{\boldsymbol{x}_i', \bar{\boldsymbol{x}}_i'\})}{\mathbb{P}_g(\{x_1, \bar{x}_1\} \equiv \{x_1', \bar{x}_1'\})\mathbb{P}_g(\{x_2, \bar{x}_2\} \equiv \{x_2', \bar{x}_2'\})}.$$

There are four possible ways of having a two-locus genotypic match. But, again, those possibilities are not mutually exclusive and the probability of a two-locus genotypic match can be computed using the inclusion-exclusion principle. The three possible two-locus match probabilities, depending on the subpopulations of the gametes, are shown in Figure 2.11. The three ratios, R_g, can be computed depending on the origin of the gametes. The closed system of equation has 1463 variables (or equations). Numerical values are reported in Table 2.5.

Similar conclusions can be drawn as in the two-locus haplotypic case. The main difference is that genotypic match probabilities are considerably smaller (by about one order of magnitude). Moreover, for small migration rates, the ratio R_g is not as elevated as R_h. For example, in the case of symmetric migration with $m_{\alpha\beta} = m_{\beta\alpha} = 10^{-6}$, the between-subpopulation ratio $R_g(\{\boldsymbol{a}, \bar{\boldsymbol{a}}\} \equiv \{\boldsymbol{b}, \bar{\boldsymbol{b}}\})$ is equal to 1.229 for $\mu = 10^{-3}$ and 1.007 for $\mu = 10^{-4}$, while the corresponding values for the haplotypic ratio R_h are 5.113 and 1.050, respectively (see Tables 2.3 and 2.5). The ratio R_g is close to 1 for $\mu = 10^{-4}$, suggesting that in the genotypic case, the product rule is accurate in the two-locus case for small mutation rates.

Table 2.5: Two-locus genotypic match probabilities and their ratios R_g to the approximate two-locus genotypic match probabilities obtained using the product rule.

							$\mathbb{P}_g(\{\boldsymbol{x}_i, \bar{\boldsymbol{x}}_i\} \equiv \{\boldsymbol{x}'_i, \bar{\boldsymbol{x}}'_i\})$			$R_g(\{\boldsymbol{x}_i, \bar{\boldsymbol{x}}_i\} \equiv \{\boldsymbol{x}'_i, \bar{\boldsymbol{x}}'_i\})$		
	μ	$m_{\alpha\beta}$	$m_{\beta\alpha}$	N_α	N_β	F_{ST}	Within α	Within β	Random	Within α	Within β	Random
Sym	1×10^{-4}	1×10^{-2}	1×10^{-2}	5000	5000	0.001	4.51×10^{-3}	4.51×10^{-3}	4.43×10^{-3}	1.001	1.001	1.000
	1×10^{-4}	1×10^{-4}	1×10^{-4}	5000	5000	0.077	1.00×10^{-2}	1.00×10^{-2}	5.45×10^{-3}	1.001	1.001	1.001
	1×10^{-4}	1×10^{-6}	1×10^{-6}	5000	5000	0.197	2.71×10^{-2}	2.71×10^{-2}	1.36×10^{-2}	1.000	1.000	1.000
Asym	1×10^{-4}	1×10^{-3}	1×10^{-4}	909	9091	0.018	2.72×10^{-2}	4.41×10^{-3}	4.36×10^{-3}	1.002	1.000	1.000
	1×10^{-4}	1×10^{-5}	1×10^{-4}	9091	909	0.064	4.97×10^{-3}	1.99×10^{-1}	5.83×10^{-3}	1.000	1.001	1.001
No sub	1×10^{-4}	–		10000		–	4.44×10^{-3}			1.000		
Sym	1×10^{-3}	1×10^{-2}	1×10^{-2}	5000	5000	0.001	1.67×10^{-6}	1.67×10^{-6}	1.39×10^{-6}	1.049	1.049	1.030
	1×10^{-3}	1×10^{-4}	1×10^{-4}	5000	5000	0.020	1.37×10^{-5}	1.37×10^{-5}	6.84×10^{-6}	1.017	1.017	1.017
	1×10^{-3}	1×10^{-6}	1×10^{-6}	5000	5000	0.024	1.88×10^{-5}	1.88×10^{-5}	9.41×10^{-6}	1.013	1.013	1.013
Asym	1×10^{-3}	1×10^{-3}	1×10^{-4}	909	9091	0.011	8.05×10^{-4}	1.64×10^{-6}	8.03×10^{-6}	1.015	1.027	1.017
	1×10^{-3}	1×10^{-5}	1×10^{-4}	9091	909	0.019	1.92×10^{-6}	4.45×10^{-3}	3.84×10^{-5}	1.025	1.004	1.005
No sub	1×10^{-3}	–		10000		–	1.38×10^{-6}			1.027		

See the caption of Table 2.2 for notational convention. The total population size is assumed to be $N = 10,000$ and the loci are assumed to be unlinked (i.e., $r = \frac{1}{2}$).

To examine the effect of population subdivision on the match probability computation, we plot in Figure 2.12 $\mathbb{P}_h^{random}(\boldsymbol{x} \equiv \boldsymbol{x}')$ and $\mathbb{P}_h(\boldsymbol{a} \equiv \boldsymbol{a}')$ (respectively, $\mathbb{P}_g^{random}(\{\boldsymbol{x}_i, \bar{\boldsymbol{x}}_i\} \equiv \{\boldsymbol{x}'_i, \bar{\boldsymbol{x}}'_i\})$ and $\mathbb{P}_g(\{\boldsymbol{a}_i, \bar{\boldsymbol{a}}_i\} \equiv \{\boldsymbol{a}'_i, \bar{\boldsymbol{a}}'_i\})$) as a function of F_{ST} for the two-locus haplotypic (respectively, genotypic) case. The one-locus cases are qualitatively similar and therefore not shown. We compare the no subdivision case, the symmetric case, and the asymmetric case of Table 2.3 and Table 2.5 for $\mu = 1 \times 10^{-4}$. The results in our tables labeled "Within α" and "Within β" are relevant for the same-subpopulation scenario, i.e. where the two individuals are drawn from one subpopulation versus drawn from the entire population ignoring subdivision. The "Within α" results assume population α is that subpopulation and the "Within β" results assume population β is that subpopulation. "Random" are relevant for the cryptic subdivision scenario, in which two partially isolated populations are treated as a single population.

As has been shown before (e.g. Balding and Nichols (1994, 1995)), ignoring subdivision can lead to underestimation of the match probabilities for the same-subpopulation scenario, i.e. when the two individuals belong to the same subpopulation (subplots on the right of Figure 2.12). We see this effect for both the asymmetric and the symmetric case even for small F_{ST} values. On the other hand, for the cryptic subdivision scenario, the asymmetric and symmetric "Random" match probabilities agree closely with the no subdivision match probabilities. In this case, ignoring subdivision leads to significant underestimates of the match probability only when the mutation rate is higher (say, $\mu = 1 \times 10^{-3}$; plots not shown). This effect is qualitatively the same for all cases considered.

2.5 Discussion and Conclusion

Given the increasing use of DNA evidence in criminal investigation in the United States and elsewhere, it is important to determine whether the one-locus match probability equations are adequate and whether the product rule provides accurate estimates of multi-locus match probabilities. In this paper, we showed that migration can be easily incorporated into the previously proposed graphical framework for computing exact match probabilities at stationarity. We computed the match probabilities for two loci in a population with two subpopulations that exchange migrants.

One of the advantages of our approach is that it relies on an explicit population subdivision model. That is, if one has an idea of the actual subdivision, the appropriate scenario can be investigated. Our results show that, for the same-subpopulation scenario, even with small F_{ST} values for the case of two subpopulations with symmetric or asymmetric migration, ignoring subdivision can lead to important underestimates of the match probability. This result had been shown before by others e.g. Balding and Nichols (1994, 1995) for a different subdivision model. For the cryptic subdivision scenario, underestimation of the match probability occurs only for high mutation rates (μ) of 10^{-3}, not necessarily consistent with reasonable levels of heterozygosity. Similarly, in all the results, noticeable deviations from the product rule are obtained only for the largest mutation rate (μ) of 10^{-3}, and even for that mutation rate the predictions of the product rule are close to the correct values. Under the infinite-alleles mutation model, a total population size of 10,000 and a mutation rate of 10^{-3} predict a heterozygosity larger than $4N\mu/(1+4N\mu) = 0.976$—how much larger

depends on the migration rates. Under a generalized stepwise mutation model, the predicted heterozygosity is somewhat smaller, of the order of 0.9 (Di Rienzo et al., 1994). Published heterozygosities for CODIS loci are in the range of 0.75-0.9 (Budowle et al., 2001), which suggests that mutation rates at CODIS are lower than 10^{-3} and that is consistent with the observed mutation rates at other STR loci (Ellegren, 2000). Therefore, for two loci, we conclude that the population subdivision of the kind modeled in this paper does lead to underestimates of the match probabilities but does not lead to substantial deviations from the predictions of the product rule. This conclusion is consistent with established practice see e.g. Balding (2005) and Committee on DNA Forensic Science: An Update (1996). Nevertheless, the effect of more loci and more subpopulations remains an open question. Our intuition, based on the results of higher mutation rate, is that the effects might be more important for more than two loci and more than two subpopulations. Finally, the method we present in this paper is relevant, not only for the forensic sciences, but in a general context of identity-by-descent calculation.

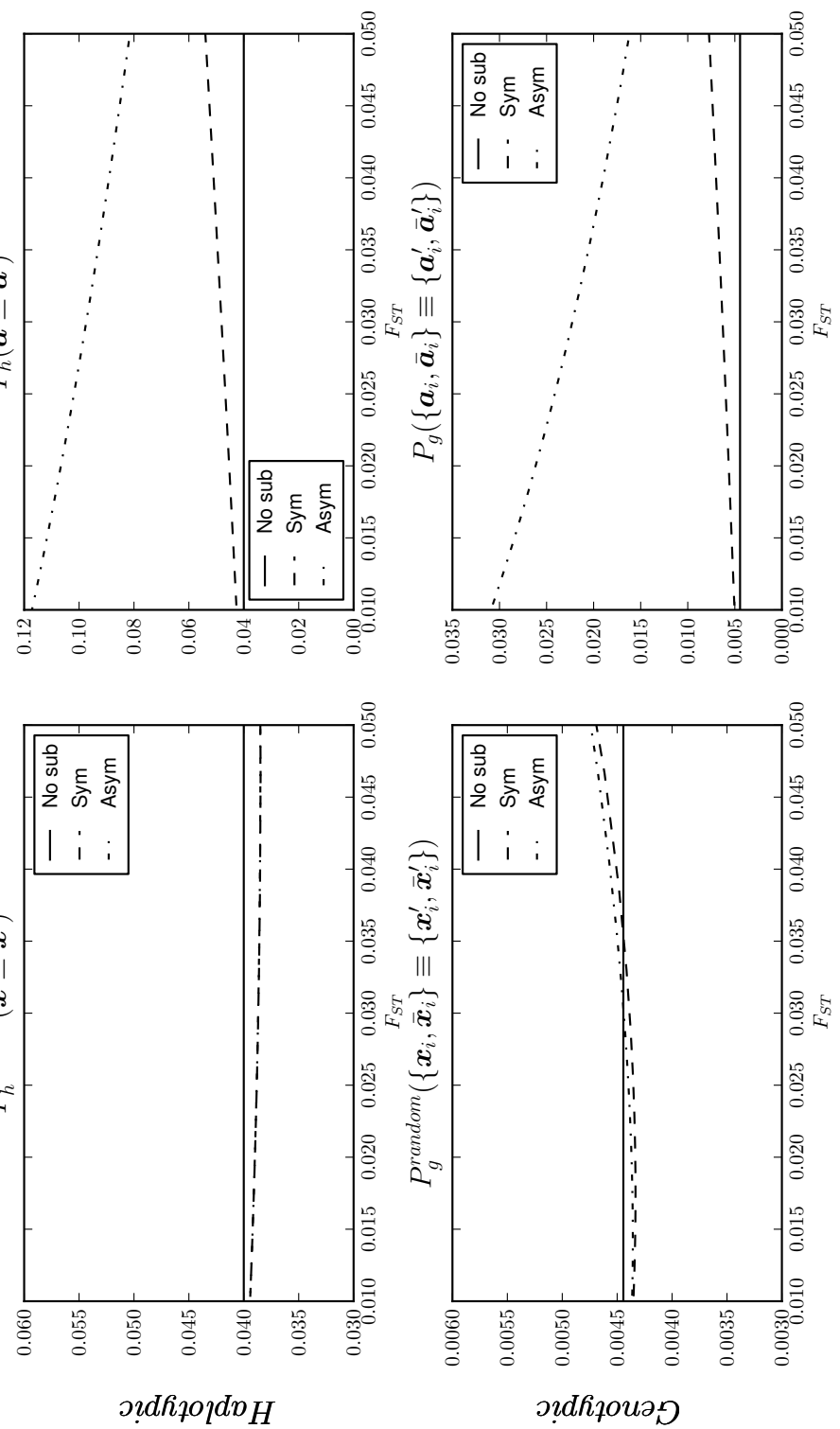

Figure 2.12: Dependency of the two-locus match probabilities on F_{ST} for F_{ST} values from 1% to 5%. The plots on the left correspond to the cryptic subdivision case where the two individuals sampled (e.g. the defendant and the crime-scene sample) are from the entire population. The plots on the right are for the same-subpopulation scenario where the two individuals sampled are from the same subpopulation, N_α. We fix the mutation rate $\mu = 1 \times 10^{-4}$ and the total population size $N = 10,000$. The top (respectively, bottom) plots are for the haplotypic (respectively, genotypic) match probabilities. The relevant column for the cryptic subdivision scenario (respectively, same-subpopulation scenario) is the "Random" (respectively, "Within α") column from Table 2.3 and Table 2.5. For the no subdivision case ("No sub") the solid lines are the same for the two scenarios for the haplotypic case (respectively, genotypic case), i.e. the "No sub" row of Table 2.3 (respectively, Table 2.5). For the asymmetric case, we fix the migration rate $m_{\beta\alpha} = 1 \times 10^{-4}$, so the subpopulation size N_α increases as F_{ST} increases from 1% to 5%. The plots for the one-locus case are similar and therefore not shown.

Chapter 3

Detecting epistasis via Markov bases

3.1 Introduction

Conditions with genetic components such as cancer, heart disease, and diabetes, are the most common causes of mortality in developed countries. Therefore, the mapping of genes involved in such complex diseases represents a major goal of human genetics. However, genetic variants associated with complex diseases are hard to detect. Indeed, only a small portion of the heritability of complex diseases can be explained by the variants identified so far. This led to several hypotheses (see e.g. Manolio et al. (2009)). One of them is that most common diseases are caused by several rare variants with low effects, rather than a few common variants with large effects (Pritchard (2001)). Another hypothesis is that the variants interact in order to produce the disease phenotype and independently only explain a small fraction of the genetic variance. In this work, we mainly focus on the interaction hypothesis, but we will also discuss the relevance of our method to the rare variant hypothesis along the way.

Recent development of methods to screen hundreds of thousands of SNPs has allowed the discovery of over 50 disease susceptibility loci with marginal effects (McCarthy et al. (2008)). Genome-wide association studies have hence proven to be fruitful in understanding complex multifactorial traits. The absence of reports of interacting loci, however, shows the need for better methods for detecting not only marginal effects of specific loci, but also interactions of loci. Although some progress in detecting interactions has been achieved in the last few years using simple log-linear models, these methods remain inefficient to detect interactions for large-scale data (Albrechtsen et al. (2007)).

Many models of interaction have been presented in the past, as for example the additive model and the multiplicative model. The former model assumes that the SNPs act independently, and a single marker approach seems to perform well. In the multiplicative model, SNPs interact in the sense that the presence of two (or more) variants have a stronger effect than the sum of the effects of each single SNP. We will discuss such models in more detail in Section 3.2.1. A complete classification of two-locus interaction models has been given in Hallgrimsdottir and Yuster (2008).

In the method described in this work, we first reduce the potential interacting SNPs to a small number by filtering all SNPs genome-wide with a single locus approach. The

loci achieving some threshold are then further examined for interactions. Such a two-stage approach has been suggested in Marchini et al. (2005). For some models of interaction, they show that the two-stage approach outperforms the single-locus search and performs at least as well as when testing for interaction within all subsets of k SNPs.

Single locus methods consider each SNP individually and test for association based on differences in genotypic frequencies between case and control individuals. Widely used methods for the single-locus search are the χ^2 goodness-of-fit test or Fisher's exact test together with a Bonferroni correction of the p-values to account for the large number of tests performed. We suggest using Fisher's exact test as a first stage to rank the SNPs by their p-value and select a subset of SNPs, which is then further analyzed. Under the rare variant hypothesis the resulting contingency tables are sparse and it is desirable to test for interactions within the selected subset using an exact test. We suggest using Markov bases for this purpose.

In Section 3.2, we define three models of interaction and present our algorithm for detecting epistasis using Markov bases in hypothesis testing. In Section 3.3, we test our method on simulated data and make a comparison to logistic regression and BEAM, a Bayesian approach (Zhang and Liu (2007)). Finally, we run our algorithm on a genome-wide dataset from dogs (Cadieu et al. (2009)) to test for epistasis related to canine hair length.

3.2 Method

3.2.1 Models of interaction

In this work, we mainly study the interaction between two SNPs and a binary phenotype, as for example the disease status of an individual. However, our method can be easily generalized for studying interaction between three or more SNPs and a phenotype with three or more states. We show a generalization in Section 3.3.4, where we analyze a genome-wide dataset from dogs and, inter alia, test for interaction between three SNPs and a binary hair length phenotype (short hair versus long hair).

The binary phenotype is denoted by D, taking values 0 and 1. We assume that the SNPs are polymorphic with only two possible nucleotides. The two SNPs are denoted by X and Y, each with genotypes taking values 0, 1 and 2 representing the number of minor alleles. We investigate three different models of interaction: a control model, an additive model, and a multiplicative model. The parameterization is given in the following tables showing the odds of having a specific phenotype

$$\frac{\mathbb{P}(D = 1|\text{genotype})}{\mathbb{P}(D = 0|\text{genotype})}.$$

- **Control model:**

		Y		
		0	1	2
	0	ϵ	ϵ	ϵ
X	1	ϵ	ϵ	ϵ
	2	ϵ	ϵ	ϵ

- **Additive model:**

		Y		
		0	1	2
	0	ϵ	$\epsilon\beta$	$\epsilon\beta^2$
X	1	$\epsilon\alpha$	$\epsilon\alpha\beta$	$\epsilon\alpha\beta^2$
	2	$\epsilon\alpha^2$	$\epsilon\alpha^2\beta$	$\epsilon\alpha^2\beta^2$

- **Multiplicative model:**

		Y		
		0	1	2
	0	ϵ	$\epsilon\beta$	$\epsilon\beta^2$
X	1	$\epsilon\alpha$	$\epsilon\alpha\beta\delta$	$\epsilon\alpha\beta^2\delta^2$
	2	$\epsilon\alpha^2$	$\epsilon\alpha^2\beta\delta^2$	$\epsilon\alpha^2\beta^2\delta^4$

These three models can also be expressed as log-linear models. We denote the state of X by i, the state of Y by j, and the state of D by k. If n_{ijk} describes the expected cell counts in a $3 \times 3 \times 2$ contingency table, then the three models can be expressed in the following way, where the γ terms represent the effects the variables have on the cell counts (e.g. γ_i^X represents the main effect for X), and α, β, δ, and ϵ are defined by the odds of having a specific phenotype shown in the above tables:

$$\begin{aligned}
&\textbf{Control model:} && \log(n_{ijk}) = \gamma + \gamma_i^X + \gamma_j^Y + \gamma_{ij}^{XY} + k\log(\epsilon) \\
&\textbf{Additive model:} && \log(n_{ijk}) = \gamma + \gamma_i^X + \gamma_j^Y + \gamma_{ij}^{XY} + k\log(\epsilon) + ik\log\alpha \\
&&& \quad + jk\log\beta \\
&\textbf{Multiplicative model:} && \log(n_{ijk}) = \gamma + \gamma_i^X + \gamma_j^Y + \gamma_{ij}^{XY} + k\log(\epsilon) + ik\log\alpha \\
&&& \quad + jk\log\beta + ijk\log\delta
\end{aligned}$$

Note that in the additive model the interaction effect for SNP X (SNP Y) and the disease status is additive with respect to the number of causative SNPs i (j), whereas in the multiplicative model there is an additional 3-way interaction effect between SNPs X, Y, and the disease status, which is multiplicative in the number of causative SNPs i, j. From the representation as log-linear models we can deduce the nesting relationship shown on the Venn diagram in Figure 3.1. Note that the additive model corresponds to the intersection of the no 3-way interaction model ($\log(n_{ijk}) = \gamma + \gamma_i^X + \gamma_j^Y + \gamma_k^D + \gamma_{ij}^{XY} + \gamma_{ik}^{XD} + \gamma_{jk}^{YD}$) with the multiplicative model, and the control model is nested within the additive model.

In a biological context, interaction between markers (or SNPs) is usually used as a synonym for *epistasis*. Cordell (2002) gives a broad definition: "Epistasis refers to departure from 'independence' of the effects of different genetic loci in the way they combine to cause disease". Epistasis is for example the result of a multiplicative effect between two markers (i.e. $\log(\delta) \neq 0$ in the multiplicative model).

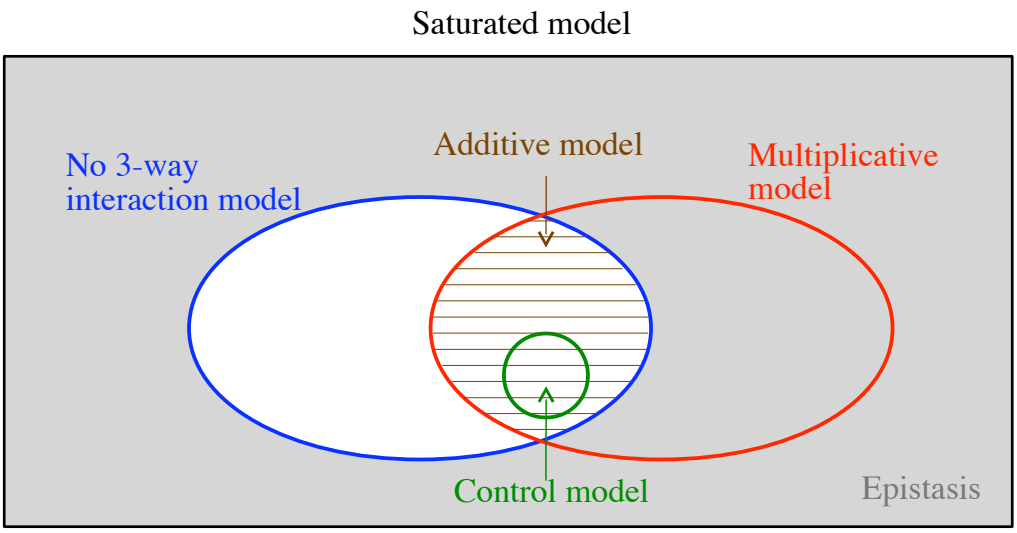

Figure 3.1: Nesting relationship of the control model, the additive model, and the multiplicative model. The intersection of the no 3-way interaction model with the multiplicative model corresponds to the additive model. The shading indicates the presence of epistasis.

In contrast, in a mathematical context interaction is used as synonym for *correlation*. Two markers are said to be interacting if they are correlated, i.e.

$$\mathbb{P}(\text{marker } 1 = i, \text{marker } 2 = j) \neq \mathbb{P}(\text{marker } 1 = i)\mathbb{P}(\text{marker } 2 = j).$$

In general, in association studies the goal is to find a set of markers that are correlated with a specific phenotype. However, the markers can be correlated with each other as well. In what follows, we will use the term interaction as synonym for correlation and the term epistasis with respect to a specific phenotype synonymously to the presence of a k-way interaction ($k \geq 3$) between $k-1$ SNPs and a specific phenotype. The epistatic models are indicated by the shading in Figure 3.1.

3.2.2 Algorithm

The χ^2 goodness-of-fit-test is the most widely used test for detecting interaction within contingency tables. Under independence the χ^2 statistic is asymptotically χ^2 distributed. However, this approximation is problematic when some cell counts are small, which is often the case in contingency tables resulting from association studies and particularly problematic under the rare variant hypothesis. The other widely used test is Fisher's exact test. As its name suggests, it has the advantage of being exact. But it is a permutation test and therefore computationally more intensive. For tables with large total counts or tables of higher dimension, enumerating all possible tables with given margins is not feasible.

Diaconis and Sturmfels (1998) describe an extended version of Fisher's exact test using Markov bases. A Markov basis for testing a specific interaction model is a set of moves connecting all contingency tables with the same sufficient statistics. So, a Markov basis allows constructing a Markov chain on the set of contingency tables with given margins and

Model	Minimal sufficient statistics	Expected counts
(X, Y, D)	$(n_{i..})$, $(n_{.j.})$, $(n_{..k})$	$\hat{n}_{ijk} = \frac{n_{i..}n_{.j.}n_{..k}}{(n_{...})^2}$
(XY, D)	$(n_{ij.})$, $(n_{..k})$	$\hat{n}_{ijk} = \frac{n_{ij.}n_{..k}}{(n_{...})}$
(XD, Y)	$(n_{i.k})$, $(n_{.j.})$	$\hat{n}_{ijk} = \frac{n_{i.k}n_{.j.}}{(n_{...})}$
(X, YD)	$(n_{i..})$, $(n_{.jk})$	$\hat{n}_{ijk} = \frac{n_{.jk}n_{i..}}{(n_{...})}$
(XY, YD)	$(n_{ij.})$, $(n_{.jk})$	$\hat{n}_{ijk} = \frac{n_{ij.}n_{.jk}}{(n_{.j.})}$
(XY, XD)	$(n_{ij.})$, $(n_{i.k})$	$\hat{n}_{ijk} = \frac{n_{ij.}n_{i.k}}{(n_{i..})}$
(XD, YD)	$(n_{i.k})$, $(n_{.jk})$	$\hat{n}_{ijk} = \frac{n_{i.k}n_{j.k}}{(n_{..k})}$
(XY, XD, YD)	$(n_{ij.})$, $(n_{i.k})$, $(n_{.jk})$	Iterative proportional fitting

Table 3.1: Standard interaction models for three-dimensional contingency tables.

computing the p-value of a given contingency table using the resulting posterior distribution. Such a test can be used for analyzing multidimensional tables with large total counts. In addition, it has been shown in Diaconis and Sturmfels (1998) that the resulting posterior distribution is a good approximation of the exact distribution of the χ^2-statistic even for very sparse contingency tables, leading to a substantially more accurate interaction test than the χ^2-test for sparse tables. Useful properties of Markov bases can be found in (Drton et al. (2009)).

The Markov basis of the null model can be computed using the software `4ti2`* and an example is given in the appendix. Then a Markov chain is started in the observed $3 \times 3 \times 2$ data table using the elements of the Markov basis as moves in the Metropolis-Hastings steps. At each step the χ^2 statistic is computed. Its posterior distribution is an approximation of the exact distribution of the χ^2 statistic.

Interaction tests with the extended version of Fisher's exact test

In this subsection we present various hypotheses that can easily be tested using Markov bases and discuss a hypothesis that is particularly interesting for association studies. The corresponding Markov basis can be found in the appendix. For simplicity we constrain this discussion to the case of two SNPs and a binary phenotype.

Table 3.2.2 consists of the standard log-linear models on three variables. Their fit to a given data table can be computed using the extended version of Fisher's exact test. We use the notation presented in Bishop et al. (1975) to denote the different models. Interaction is assumed between the variables not separated by commas in the model. So the model (X, Y, D) in Table 3.2.2 represents the independence model, the model (XY, XD, YD) the no 3-way interaction model and the other models are intermediate models. For association studies the no 3-way interaction model (XY, XD, YD) is particularly interesting and will be used as null model in our testing procedure.

Performing the extended version of Fisher's exact test involves sampling from the space of contingency tables with fixed minimal sufficient statistics and computing the χ^2 statistic. So, the minimal sufficient statistics and the expected counts for each cell of the table need

*http://www.4ti2.de/

		Phenotype status:		**Total:**
		0	1	
Haplotype:	00	n_{000}	n_{001}	$n_{00.}$
	01	n_{010}	n_{011}	$n_{01.}$
	10	n_{100}	n_{101}	$n_{10.}$
	11	n_{110}	n_{111}	$n_{11.}$
Total:		$n_{..0}$	$n_{..1}$	$n_{...}$

Table 3.2: Testing for association between haplotypes and phenotype.

to be calculated. These are given in Table 3.2.2. If a loop is present in the model configuration as for example in the no 3-way interaction model (this model can be rewritten as (XY, YD, DX)), then there is no closed-form estimator for the cell counts (see Bishop et al. (1975)). But in this case, estimates can be achieved by iterative proportional fitting (i.e. Fienberg (1970)).

It is important to note that testing for epistasis necessarily implies working with multidimensional contingency tables and is not possible in the collapsed two-dimensional table shown above. In this table, the two SNPs are treated like a single variable and we consider the haplotype and not the SNPs separately. The sufficient statistics for the model described in Table 3.2.2 are the row and column sums $(n_{ij.})$ and $(n_{..k})$. So testing for association in this collapsed table is the same as using (XY, D) as null model. In this case, the null hypothesis would be rejected even in the presence of marginal effects only, showing that testing for epistasis in Table 3.2.2 is impossible.

Hypothesis testing with the extended version of Fisher's exact test

Our goal is to detect epistasis when present. According to the definition of epistasis in Section 3.2.1 and as shown in Figure 3.1, epistasis is present with regard to two SNPs and a specific phenotype, when a 3-way interaction is found. So we suggest using as null hypothesis the no 3-way interaction model and testing this hypothesis with the extended version of Fisher's exact test. The corresponding Markov basis consists of 15 moves and is given in the appendix. It can be used to compute the posterior distribution of the χ^2 statistic and approximate the exact p-value of the data table. If the p-value is lower than some threshold, we reject the null hypothesis of no epistasis.

Although in this work we focus merely on epistasis, it is worth noting that one can easily build tests for different types of interaction using Markov bases. If one is interested in detecting whether the epistatic effect is of multiplicative nature, one can perform the extended version of Fisher's exact test on the contingency tables, which have been classified as epistatic, using the multiplicative model as null hypothesis. In this case, the corresponding Markov basis consists of 49 moves. Similarly, if one is interested in detecting additive effects, one can use the additive model as null hypothesis and test the contingency tables, which have been classified as non-epistatic. In this case, the corresponding Markov basis consists

of 156 moves. The Markov bases for these tests can be found on our website[†].

3.3 Results

In this section, we first conduct a simulation study to evaluate the performance of the suggested method. We then compare our method to a two-stage logistic regression approach and to BEAM (Zhang and Liu (2007)). Logistic regression is a widely used method for detecting epistasis within a selection of SNPs. BEAM is a purely Bayesian method for detecting epistatic interactions on a genome-wide scale. We end this section by applying our method to a genome-wide data set consisting of 685 dogs with the goal of finding epistasis associated with canine hair length.

3.3.1 Simulation study

We simulated a total of 50 potential association studies with 400 cases and 400 controls for three different minor allele frequencies of the causative SNPs and the three models of interaction presented in Section 3.2.1. We chose as minor allele frequencies (MAF) 0.1, 0.25 and 0.4. The parameters for the three models of interaction were determined numerically fixing the marginal effect measured by the effect size

$$\lambda_i := \frac{p(D=1|g_i=1)}{p(D=0|g_i=1)} \frac{p(D=0|g_i=0)}{p(D=1|g_i=0)} - 1$$

and the prevalence

$$\pi := \sum_{g_1,g_2} p(D|g_1,g_2)p(g_1,g_2).$$

For our simulations, we used an effect size of $\lambda_1 = \lambda_2 = 1$ and a sample prevalence of $\pi = 0.5$. Choosing in addition $\alpha = \beta$ in the additive model, and $\alpha = \beta$ and $\delta = 3\alpha$ in the multiplicative model, determines all parameters of the interaction models and one can solve for α, β, δ and ϵ numerically.

The simulations were performed using HAP-SAMPLE (Wright et al. (2007)) and were restricted to the SNPs typed with the Affy CHIP on chromosome 9 and chromosome 13 of the Phase I/II HapMap data[‡], resulting in about 10,000 SNPs per individual. On each of the two chromosomes we selected one SNP to be causative. The causative SNPs were chosen consistent with the minor allele frequencies and far apart from any other marker (at least 20,000bp apart). Note that HAP-SAMPLE generates the cases and controls by resampling from HapMap. This means that the simulated data show linkage disequilibrium and allele frequencies similar to real data.

As suggested in Marchini et al. (2005), we took a two-stage approach for finding interacting SNPs. In the first step, we ranked all SNPs according to their p-value in Fisher's exact test on the 2x3 genotype table and selected the ten SNPs with the lowest marginal p-values. Within this subset, we then tested for interaction using the extended version of Fisher's

[†]http://www.carolineuhler.com/epistasis.htm
[‡]http://hapmap.ncbi.nlm.nih.gov/

exact test with the no 3-way interaction model as null hypothesis. We generated three Markov chains with 40,000 iterations each and different starting values, and used the tools described in Gilks et al. (1995) to assess convergence of the chains. This included analyzing the Gelman-Rubin statistic and the autocorrelations. After discarding an initial burn-in of 10,000 iterations, we combined the remaining samples of the three chains to generate the posterior distribution of the χ^2 statistic.

In Figure 3.2 (left), we report the rejection rate of the no 3-way interaction hypothesis for each of the three minor allele frequencies. Per point in the figure we simulated 50 potential association studies. The power of our two-stage testing procedure corresponds to the curve under the multiplicative model. The higher the minor allele frequency, the more accurately we can detect epistasis. Under the additive model and the control model, no epistasis is present. We never rejected the null hypothesis under the control model and only once under the additive model, resulting in a high specificity of the testing procedure.

We also analyze the performance of each step separately. Figure 3.2 (middle) shows the performance of the first step and reports the proportion of 50 association studies, in which the two causative SNPs were ranked among the ten SNPs with the lowest p-values. Because Fisher's exact test measures marginal association, the curves under the additive model and the multiplicative model are similar.

Figure 3.2 (right) shows the performance of the second step in our method and reports the proportion of 50 association studies, in which the null hypothesis of no 3-way interaction was rejected using only the extended version of Fisher's exact test on the 50 causative SNP pairs.

3.3.2 Comparison to logistic regression

For validation, we compare the performance of our method to logistic regression via ROC curves. Logistic regression is probably the most widely used method for detecting epistasis within a selection of SNPs nowadays. We base the comparison on the simulated association studies presented in the previous section using only the simulations under the multiplicative model. The structure of interaction within this model should favor logistic regression as logistic regression tests for exactly this kind of interaction.

As before, for each minor allele frequency and each of the 50 simulation studies, we first filtered all SNPs with Fisher's exact test and chose the ten SNPs with the lowest p-values for further analysis. Both causative SNPs are within the ten filtered SNPs for 19 (46) [45] out of the 50 simulation studies for MAF=0.1 (MAF=0.25) [MAF=0.4]. We then ran the extended version of Fisher's exact test and logistic regression on all possible pairs of SNPs in the subsets consisting of the ten filtered SNPs. This results in $50 \cdot \binom{10}{2}$ tests per minor allele frequency with 19 (46) [45] true positives for MAF=0.1 (MAF=0.25) [MAF=0.4].

Because both methods, logistic regression and our method, require filtering all SNPs first, we compare the methods only based on the ten filtered SNPs. The ROC curves comparing the second stage of our method to logistic regression are plotted in Figure 3.3 showing that our method performs substantially better than logistic regression for MAF=0.1 with an area under the ROC curve of 0.861 compared to 0.773 for logistic regression. For MAF=0.25 and MAF=0.4 both methods have nearly perfect ROC curves with areas 0.9986 [0.99994] for our method compared to 0.9993 [0.99997] for logistic regression for MAF=0.25 [MAF=0.4].

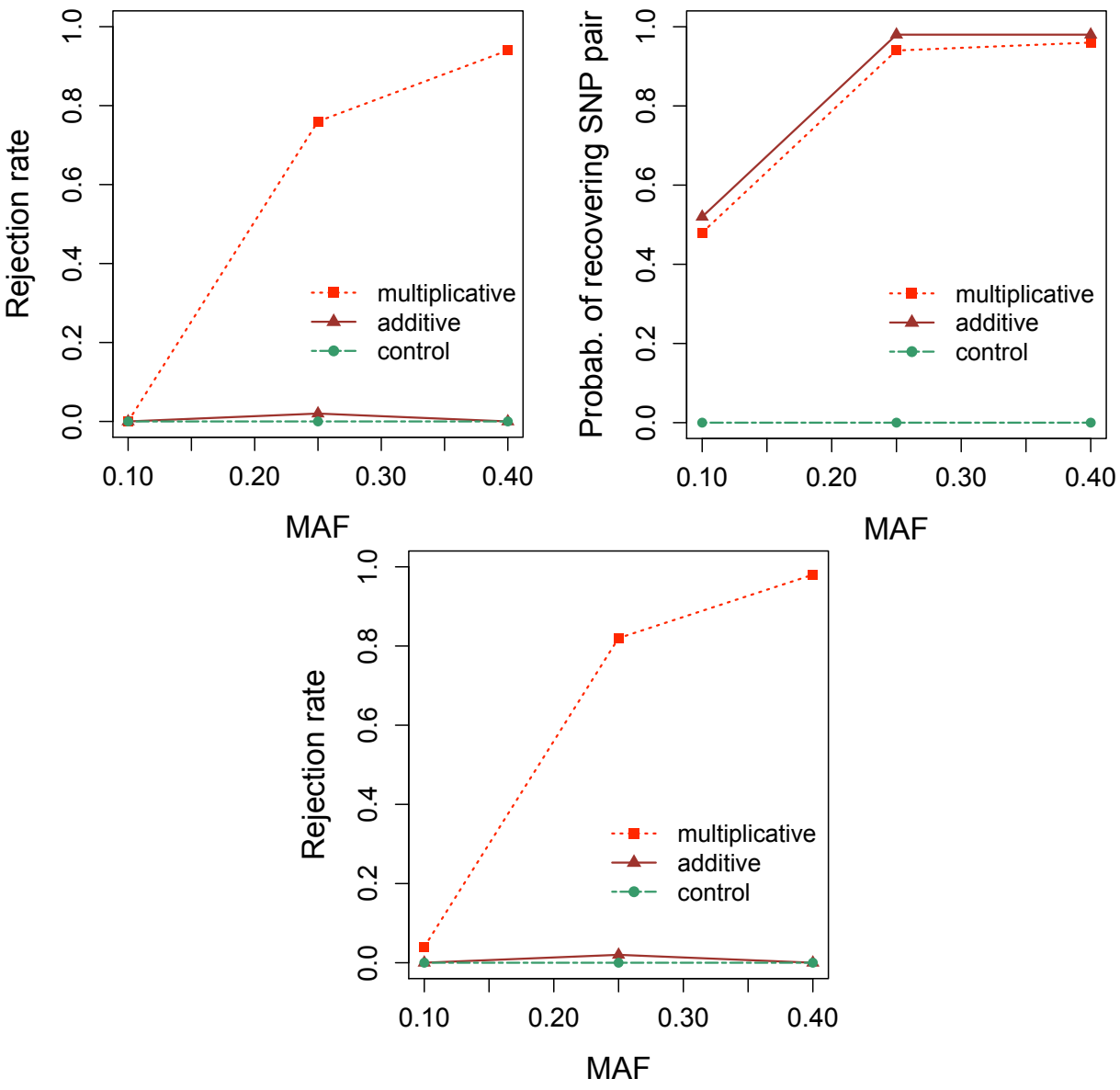

Figure 3.2: Rejection rate of the no 3-way interaction test in the two-stage approach on 50 simulated association studies for MAF=0.1, MAF=0.25, and MAF=0.4 (top left). Proportion of 50 association studies, in which the two causative SNPs were ranked among the ten SNPs with the lowest p-values by Fisher's exact test (top right). Rejection rate of the no 3-way interaction hypothesis using only the extended version of Fisher's exact test on the 50 causative SNP pairs (bottom).

3.3.3 Comparison to BEAM

We also compare our method to BEAM, a Bayesian approach for detecting epistatic interactions in association studies (Zhang and Liu (2007)). We chose to compare our method to BEAM, because the authors show it is more powerful than a variety of other approaches including the stepwise logistic regression approach, and it is one of the few recent methods

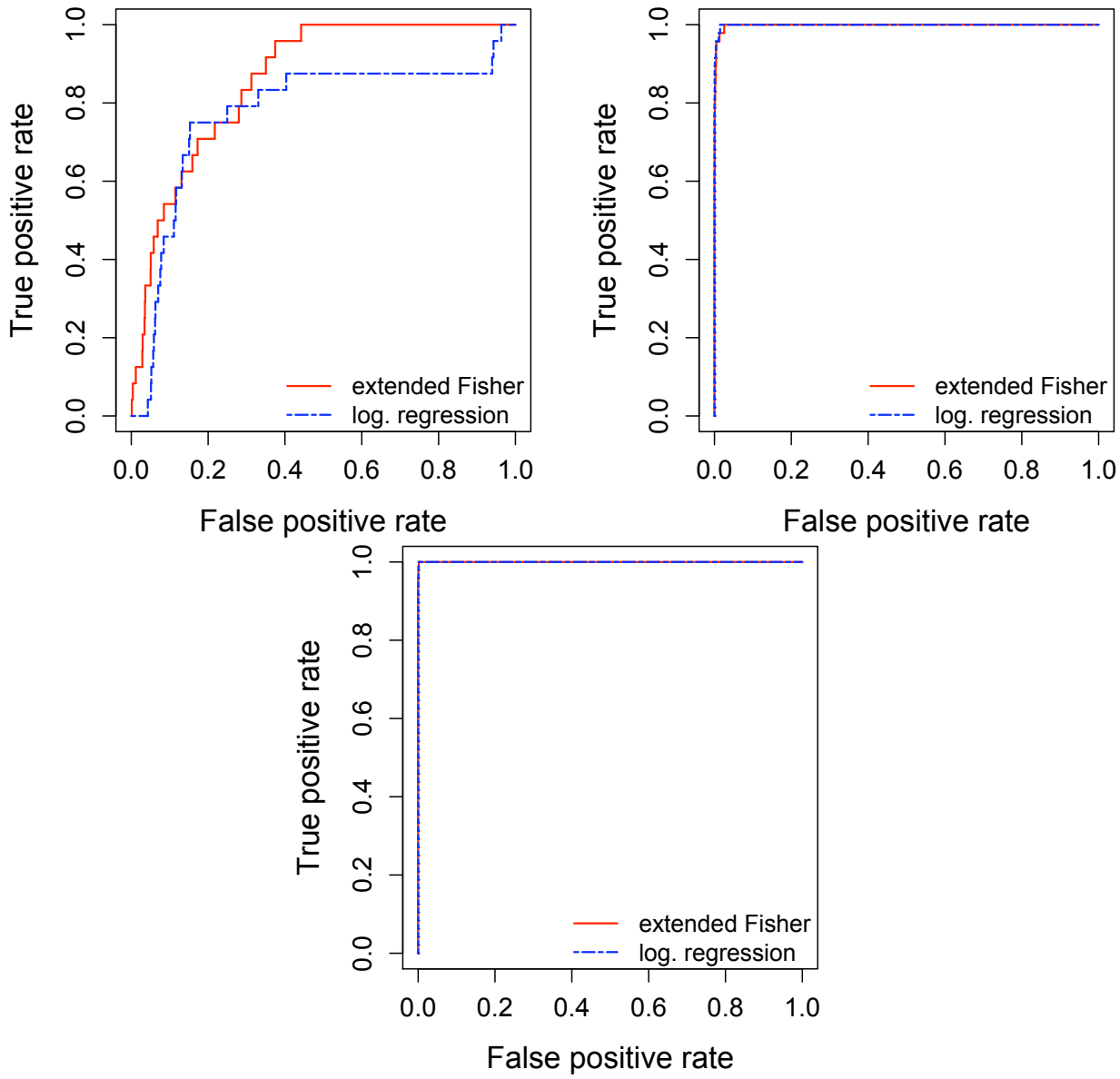

Figure 3.3: ROC curves of the extended version of Fisher's exact test and logistic regression for MAF=0.1 (top left), MAF=0.25 (top right), and MAF=0.4 (bottom) based on the ten filtered SNPs.

that can handle genome-wide data.

In this method, all SNPs are divided into three groups, namely, SNPs that are not associated with the disease, SNPs that contribute to the disease risk only through main effects, and SNPs that interact to cause the disease. BEAM outputs the posterior probabilities for each SNP to belong to these three groups. The authors propose to use the results in a frequentist hypothesis-testing framework calculating the so called B-statistc and testing for association between each SNP or set of SNPs and the disease phenotype. BEAM was designed to increase the power to detect any association with the disease, and not to separate

main effects from epistasis. Therefore, BEAM outputs SNPs that interact marginally or through a k-way interaction with the disease. This does not match our definition of epistasis since the presence of marginal effects only, already gives rise to a significant result using BEAM.

We compare our method to BEAM using the B-statistic. BEAM reports this statistic only for the pairs of SNPs which have a non-zero posterior probability of belonging to the third group. In addition, the B-statistic is automatically set to zero for the SNP pairs where any of the SNPs is found to be interacting marginally with the disease. We force BEAM to include the marginal effects into the B-statistic by choosing a significance level of zero for marginal effects. This should favor BEAM in terms of sensitivity.

We ran BEAM with the default parameters on our simulated datasets for the multiplicative model. Due to the long running time of BEAM, we based the comparison only on 1,000 SNPs out of the 10,000 SNPs simulated for the analysis in Section 3.3.1. BEAM takes about 10.6 hours for the analysis of one dataset with 10,000 SNPs and 400 cases and controls, whereas the same analysis with our method takes about 0.7 hours on an Intel Core 2.2 GHz laptop with 2 Gb memory.

In contrast to BEAM, our method is a stepwise approach, which makes a comparison via ROC curves difficult. We therefore compare the performance of all three tests by plotting for a fixed number x of SNP pairs the proportion of simulation studies for which the interacting SNP pair belongs to the x SNP pairs with the lowest p-values. The resulting curves are shown in Figure 3.4. Although the marginal effects were not extracted, BEAM has a very high false negative rate, attributing a p-value of 1 to the majority of SNPs, interacting and not interacting SNPs.

3.3.4 Genome-wide association study of hair length in dogs

We demonstrate the potential of our Markov basis method in genome-wide association studies by analyzing a hair length dataset consisting of 685 dogs from 65 breeds and containing $40,842$ SNPs (Cadieu et al. (2009)).

The individuals in Cadieu et al. (2009) were divided into two groups for the hair length phenotype: 319 dogs from 31 breeds with long hair as cases and 364 from 34 breeds with short hair as controls. In the original study, it is shown that the long versus short hair phenotype is associated with a mutation (Cys95Phe) that changes exon one in the *fibroblast growth factor-5* (*FGF5* gene). Indeed, the SNP with the lowest p-value using Fisher's exact test is located on chromosome 32 at position $7,100,913$ for the Canmap dataset, i.e. about 300Kb apart from *FGF5*.

We ranked the $40,842$ SNPs by their p-value using Fisher's exact test and selected the 20 lowest ranked SNPs (about 0.05%) to test for 3-way interaction. Note that all 20 SNPs are significantly correlated (p-value < 0.05) with the phenotype. We found a significant p-value (< 0.05) for four out of the $\binom{20}{2}$ pairs. These pairs together with their p-values are listed in Table 3.3.

The pairs include six distinct SNPs located on five different chromosomes and the two SNPs lying on the same chromosome are not significantly interacting (p-value of 0.54). This means that a false positive correlation due to hitchhiking effects can likley be avoided. Hitchhiking effects are known to extend across long stretches of chromosomes in particular

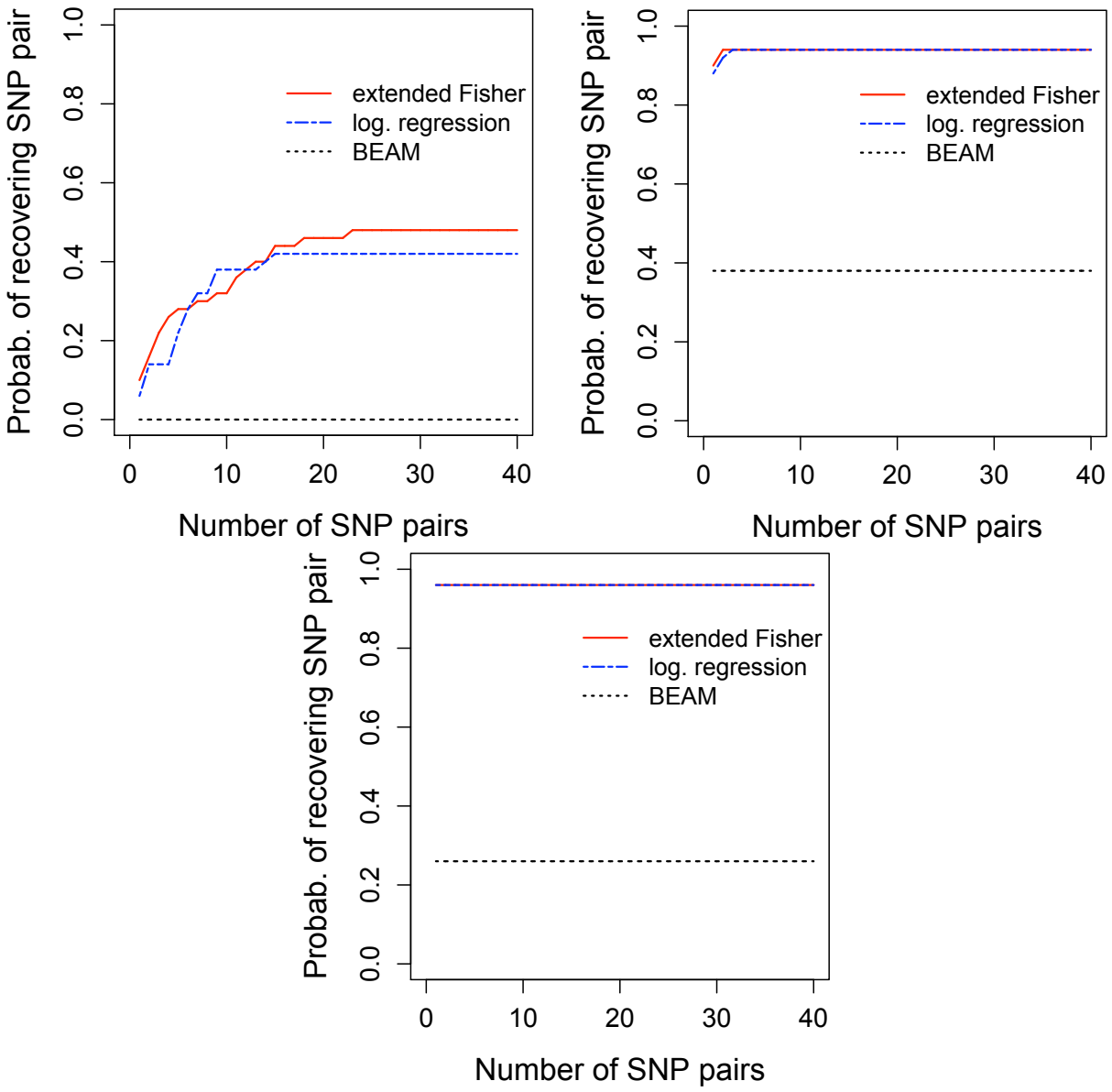

Figure 3.4: Proportion of simulation studies for which the interacting SNP pair belongs to the x SNP pairs with the lowest p-values for MAF=0.1 (top left), MAF=0.25 (top right), and MAF=0.4 (bottom).

in domesticated species (Sutter et al. (2004); Wayne and Ostrander (2007); Mather et al. (2007)) consistent with the prediction of Smith and Haigh (1974).

In order to identify potential pathways we first considered genes, which are close to the six SNPs we identified as interacting. To do so, we used the dog genome available through the ncbi website[§]. Most of the genes we report here have been annotated automatically. Our strategy was to consider the gene containing the candidate SNP (if any) and the immediate left and right neighboring gene, resulting in a total of two or three genes per SNP.

[§]http://www.ncbi.nlm.nih.gov/genome/guide/dog/, build 2.1

chromosome and location of SNPs	p-value	potential relevant genes
chr30.18465869, chr26.6171079	0	*FGF7*-?
chr15.44092912, chr23.49871523	0	*IGF1*-*P2RY1*
chr24.26359293, chr15.43667654	2e-04	*ASIP*-?
chr15.43667654, chr23.49871523	1e-04	?-*P2RY1*

Table 3.3: Pairs of SNPs, which significantly interact with the hair length phenotype for the Canmap dataset. Question marks indicate that we were not able to identify a closeby gene which is functionally related to hair growth.

Among the six significantly interacting SNPs, four are located close to genes that have been shown to be linked to hair growth in other organisms. This is not surprising, since these SNPs also have a significant marginal association with hair growth. We here report the function of these candidate genes. The two other SNPs are located close to genes that we were not able to identify as functionally related to hair growth.

First, the SNP chr30.18465869 is located close to (about 80Kb) *fibroblast growth factor 7* (*FGF7* also called *keratinocyte growth factor*, *KGF*), i.e. it belongs to the same family as the gene reported in the original study (but on a different chromosome). The FGF family members are involved in a variety of biological processes including hair development reported in human, mouse, rat and chicken (GO:0031069, Ashburner et al. (2000)).

Secondly, chr15.44092912 is located between two genes, and about 200Kb from the *insulin-like growth factor 1* gene (*IGF1*). *IGF1* has been reported to be associated with the hair growth cycle and the differentiation of the hair shaft in mice (Weger and Schlake (2005)).

Thirdly, chr23.49871523 is located about 430Kb from the *purinergic receptor P2Y1* (*P2RY1*). The purinergic receptors have been shown to be part of a signaling system for proliferation and differentiation in human anagen hair follicles (Greig et al. (2008)).

Finally, chr24.26359293 is located inside the agouti-signaling protein (gene *ASIP*), a gene known to affect coat color in dogs and other mammals. The link to hair growth is not obvious but this gene is expressed during four to seven days of hair growth in mice (Wolff et al. (2007)).

According to our analysis, *IGF1* and *P2RY1* are significantly interacting. All other pairs of interacting SNPs involve at least one SNP for which we were not able to identify a closeby candidate gene related to hair growth (see Table 3.3). *IGF1* has a tyrosine kinase receptor and *P2RY1* is a G-protein coupled receptor. One possibility is that these receptors cross-talk as has been shown previously for these types of receptors in order to control mitogenic signals (Dikic and Blaukat (1999)). However, a functional assay would be necessary to establish that any of the statistical interactions we found are also biologically meaningful.

We also considered all triplets of SNPs among the 20 preselected SNPs and tested for 4-way interaction. However, we did not find any evidence for interaction among the $\binom{20}{3}$ triplets.

3.4 Discussion

In this work, we proposed a Markov basis approach for detecting epistasis in genome-wide association studies. The use of different Markov bases allows to easily test for different types of interaction and epistasis involving two or more SNPs. These Markov bases need to be computed only once and can be downloaded from our website[¶] for the tests presented in this work.

The use of an exact test is of particular relevance for disease mapping studies where the contingency tables are often sparse. One example where there has been also functional validation, is a deletion associated with Crohn's disease McCarroll et al. (2008) This deletion was found to have a population frequency of 0.07, and a frequency of 0.11 in the cases Manolio et al. (2009). So within 400 controls and under Hardy-Weinberg equilibrium, we would expect only 2 individuals to be homozygote for this deletion. This shows that also for a moderate number of cases and controls the resulting tables for disease association studies are likely to be sparse. The sparsity is even more pronounced for rare variants, defined as variants with a MAF smaller than 0.005. Current genome wide association studies are still missing these rare variants, but advances in sequencing technologies should allow to sequence these variants and apporpriate statistical methods will then be necessary.

We tested our method in simulation studies and showed that it outperforms a stepwise logistic regression approach and BEAM for the multiplicative interaction model. Logistic regression has the advantage of a very short running time (3 seconds compared to 39 minutes using our method for the analysis of one dataset with 10,000 SNPs and 400 cases and controls not including the filtering step, which takes about 1 minute for both methods on an Intel Core 2.2 GHz laptop with 2 Gb memory). However, especially for a minor allele frequency of 0.1, logistic regression performs worse than our method, even when simulating epistasis under a multiplicative model, which should favor logistic regression. This difference arises because our method approximates the exact p-value well for all sample sizes while the performance of logistic regression increases with larger sample size. 400 cases and 400 controls are not sufficient to get a good performance using logistic regression for a minor allele frequency of 0.1 and it is expected to do even worse for rare variants. Another advantage of our method compared to logistic regression is that it is not geared towards testing for multiplicative interaction only, but should be able to detect epistasis regardless of the interaction model chosen. It would be interesting to compare these two methods on data sets generated by other interaction models.

BEAM on the other hand, has the advantage of not needing to filter the large number of SNPs first. However, it runs about 15 times slower than our method for our simulations and has a very high false negative rate. The difference between our results and what the authors of BEAM have found might be due to linkage disequilibrium in our data. BEAM handles linkage disequilibrium with a first order Markov chain, which will be improved in future versions (Yu Zhang, personal communication). But as of today, we conclude that this method is impractical for whole genome association studies, since linkage disequilibrium is present in most real datasets.

The limitation of our method is the need for a filtering step to reduce the number of

[¶]http://www.carolineuhler.com/epistasis.htm

SNPs to a small subset. Especially if the marginal association of the interacting SNPs with the disease is small, these SNPs might not be caught by the filter. However, in our simulations using Fisher's exact test as a filter seems to perform well. Another possibility is to incorporate biological information such as existing pathways (Emily et al. (2009)) to choose a subset of possibly interacting SNPs.

We demonstrated the potential of the proposed two-stage method in genome-wide association studies by analyzing a hair length dataset consisting of 685 dogs and containing $40,842$ SNPs using the extended version of Fisher's exact test. In this dataset, we found a significant epistatic effect for four SNP pairs. These SNPs lie on different chromosomes, reducing the risk of a false positive correlation due to linkage effects. The dataset includes dogs from 65 distinct breeds. Although linkage disequilibrium has been shown to extend over several megabases within breeds, linkage disquelibrium extends only over tens of kilobases between breeds and drops faster than in human populations (Sutter et al. (2004), Karlsson et al. (2007), Lindblad-Toh et al. (2005)), suggesting that it is possible to do fine-mapping between breeds. These observations are consistent with two bottlenecks, the first associated with the domestication from wolves and the second associated with the intense selection to create the breeds. Other studies have successfully employed the extensive variation between breeds to map genes affecting size and behavior (Jones et al. (2008); Cadieu et al. (2009)). The validity of this approach rests on the assumption that the breeds used are random samples of unrelated breeds or that related breeds make up a small part of our sample (Jones et al. (2008); Goddard and Hayes (2009)). This is rarely the case and false positive results may therefore have arisen from population structure. A second independent dataset would be useful to confirm our findings. Finally, a functional assay would be necessary to establish if the interactions we found are also biologically meaningful.

Appendix: Markov basis

The Markov basis corresponding to the no 3-way interaction model on a $3 \times 3 \times 2$ table is given below. The tables are reported as vectors

$$(n_{111}, n_{211}, n_{311}, n_{121}, n_{221}, n_{321}, n_{131}, n_{231}, n_{331}, n_{112}, n_{212}, n_{312}, n_{122}, n_{222}, n_{322}, n_{132}, n_{232}, n_{332}).$$

$$
\begin{array}{lcrrrrrrrrrrrrrrrrrr}
f_1 &=& (0 & 0 & 0 & 1 & 0 & -1 & -1 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 1 & 1 & 0 & -1) \\
f_2 &=& (0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 1 & -1) \\
f_3 &=& (1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 & -1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & -1) \\
f_4 &=& (0 & 1 & -1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 1 & -1) \\
f_5 &=& (0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 1 & -1 & 0) \\
f_6 &=& (1 & -1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 1 & -1 & 0) \\
f_7 &=& (1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 1 & -1 & 0 & 0 & 0 & 0) \\
f_8 &=& (1 & 0 & -1 & -1 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 1 & 1 & 0 & -1 & 0 & 0 & 0) \\
f_9 &=& (0 & 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 1 & -1 & 0 & 0 & 0) \\
f_{10} &=& (0 & 1 & -1 & -1 & 0 & 1 & 1 & -1 & 0 & 0 & -1 & 1 & 1 & 0 & -1 & -1 & 1 & 0) \\
f_{11} &=& (1 & 0 & -1 & 0 & -1 & 1 & -1 & 1 & 0 & -1 & 0 & 1 & 0 & 1 & -1 & 1 & -1 & 0) \\
f_{12} &=& (-1 & 1 & 0 & 1 & 0 & -1 & 0 & -1 & 1 & 1 & -1 & 0 & -1 & 0 & 1 & 0 & 1 & -1) \\
f_{13} &=& (1 & -1 & 0 & 0 & 1 & -1 & -1 & 0 & 1 & -1 & 1 & 0 & 0 & -1 & 1 & 1 & 0 & -1) \\
f_{14} &=& (1 & 0 & -1 & -1 & 1 & 0 & 0 & -1 & 1 & -1 & 0 & 1 & 1 & -1 & 0 & 0 & 1 & -1) \\
f_{15} &=& (0 & 1 & -1 & 1 & -1 & 0 & -1 & 0 & 1 & 0 & -1 & 1 & -1 & 1 & 0 & 1 & 0 & -1)
\end{array}
$$

Chapter 4

Characterizing Neanderthal admixture using the joint derived SFS with humans

Genealogy. *An account of one's descent from an ancestor who did not particularly care to trace his own.* The Devil's Dictionary, Ambrose Bierce, 1911.

4.1 Introduction

Neanderthal history Maybe more than any other member of the *homo* genus, the Neanderthals fascinate. Much of it is probably owing to the first characterization of the Neanderthals as inferior to humans. They were always thought to be devoid of symbolic thinking, some wild brutal form of human. For example, in 1864, Hugh Falconer, an eminent paleoanthropologist, attributed the Forbe's Quarry Neanderthal cranium to "a very low type of humanity - very low and savage, and of extreme antiquity" at the British Association for the Advancement of Science meeting in Bath (Patou-Mathis, 2006). Even today, the word "neanderthal" is used as an adjective to suggest a person that is uncivilized, unintelligent, brutal, or even old-fashioned and conservative*. The idea that modern humans effectively replaced them while dispersing out of Africa has contributed to such mythology. It is perhaps for this reason that the debate about whether the (inferior) Neanderthals admixed with the (superior) humans continues to attract the attention and to fuel the imagination of a wide public (Wade, 2010; Curry, 2010; Rincon, 2010; Finlayson, 2010; Dessibourg, 2010).

Neanderthals are believed to be the closest evolutionary relative of modern humans. But how exactly the Neanderthals disappeared, and the nature of their relationship to modern humans, remains a widely disputed topic. The first Neanderthal, a 2-3 year old child, was found in Engis Cave, Belgium, in 1829-1830 but was not attributed to the Neanderthals until much later. The type specimen, a skullcap, was discovered 20 years later in the Neander Valley in 1856. Several fossils have been found in Eurasia since these first discoveries from sites ranging from Europe to Siberia (see Figure 4.1).

---

* http://oxforddictionaries.com/view/entry/m_en_gb0550940#m_en_gb0550940
http://dictionary.cambridge.org/dictionary/british/neanderthal_2

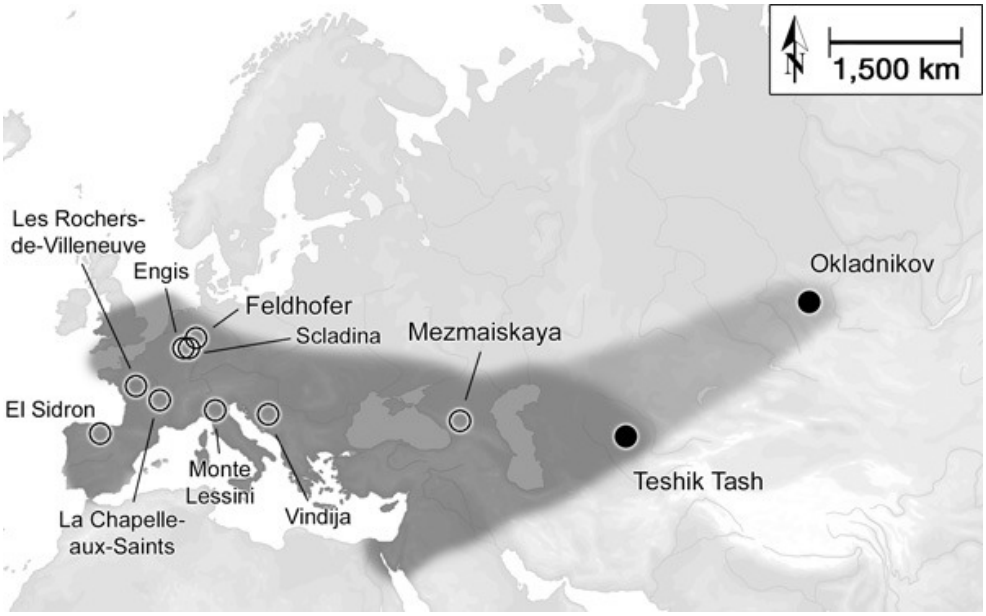

Figure 4.1: Range of Neanderthal fossils across Europe and Middle East. Figure provided by Johannes Krause, (see Figure 1 of Krause et al. (2007)).

It is hard to get a precise picture of the extent to which Neanderthals and anatomically modern humans coexisted or alternated in space and time. Indeed, authors do not always agree, with lines of evidence stemming from both fossil remains and archaeological study. According to Patou-Mathis (2006), the last Neanderthal fossils were found in Europe and are of the order of 30,000 years old. The most recent fossil was discovered in the Vindija cave, Croatia and is estimated to about 28,000 years (Smith et al. (1999) but see also Joris and Street (2008)). In the Middle East on the other hand, the youngest known discovery is found in Amud in a layer about 45,000 years old (e.g. Patou-Mathis (2006)). Industry associated with Neanderthals is dated to between 46,000 and 100,000 years in the Tabun cave (Israel) (e.g. Conroy (2005)).

Some authors believe that the first anatomically modern humans were in Europe as early as 40,000 years (e.g. Patou-Mathis (2006)). If those dates are correct, Neanderthals and anatomically modern humans co-inhabited in Europe for around 12,000 years.

In the Middle East, the coexistence seems to have lasted longer. Anatomically modern humans predate Neanderthals according to the fossil record (Patou-Mathis (2006) and Conroy (2005)). At Qafzeh, dental samples believed to belong to anatomically modern humans have been dated to 90,000 years (Conroy (2005) and citation therein). This would mean an overlap of around 45,000 years.

The numbers differ by up to an order of magnitude between authors, but in most cases, they all agree that Neanderthals and humans coexisted for at least 2,000 years in Europe, and in the Middle East around 40,000 years ago (but see also Joris and Street (2008)). The discrepancies between authors is due both to the difficulty of attributing fossils to a specific lineage and also of dating the specimens.

Most paleoanthropologists agree that classic Neanderthals have unique features (Conroy,

2005; Harvati et al., 2004), with the most often cited being the double arched bony browridge, the high wide and voluminous nose, high rounded orbits, large front teeth and short limbs. In favor of admixture, some authors have suggested certain remains to be hybrids between modern humans and Neanderthals (Duarte et al. (1999) but see Tattersall and Schwartz (1999)).

In the last decades, several teams have demonstrated that Neanderthals used sophisticated tools (Eren et al., 2008), controlled fire (Albert et al., 2010; Roebroeks and Villa, 2011), buried their dead (Solecki, 1975) and made symbolic or ornamental objects (Zilhao et al., 2010). These later studies have served to shed a more "human" light on Neanderthals. Despite this wealth of anthropological knowledge however, there is insufficient evidence to resolve the admixture question (see Herrera et al. (2009) for a review).

Evidence from genetic data The late 20th century has seen the emergence of sequencing technologies that have boosted the production of genomic data. This development has had a huge impact on the field of ancient DNA.

Thanks to those advances, the first short fragment of Neanderthal mtDNA was published in 1997 (Krings et al., 1997). Several more fragments of Neanderthal DNA were published in the following years (Krings et al., 1999; Ovchinnikov et al., 2000; Krings et al., 2000; Schmitz et al., 2002; Serre et al., 2004; Lalueza-Fox et al., 2005; Orlando et al., 2006; Caramelli et al., 2006; Lalueza-Fox et al., 2006; Krause et al., 2007) and in 2008 (Green et al., 2008), the first complete mtDNA sequence was obtained. Finally in 2009, five more complete mtDNA were sequenced (Briggs et al., 2009). All studies based on mtDNA have suggested that the Neanderthals fall outside of human diversity. However, mtDNA is only a single locus, and some have argued that a small amount of admixture could not be excluded. While Nordborg (1998) concludes that one can only reject a model with random mating between Neanderthals and modern humans, Serre et al. (2004) exclude an admixture proportion higher than 25% and Blum and Rosenberg (2007) found an upper bound of 5% for the amount of admixture. Those three studies are based on different mtDNA data, and assume different demographic models. Others found no evidence for admixture. Indeed, Currat and Excoffier (2004) argue, also based on mtDNA, that admixture may be lower than 0.1% if there is range expansion.

Prior to the sequencing of the Neanderthal genome, several studies used human (nuclear) DNA and made inference about admixture proportions. Some made use of multiple loci. For example Wall et al. (2009) use a simulation approach to get maximum likelihood estimates for several demographic scenarios. In particular, they conclude that 14% admixture could explain the data, and that a model without admixture could not fit the data significantly better. Eswaran et al. (2005), using a simulation approach, conclude that up to 80% of human loci may have been influenced by archaic admixture. But others, such as Fagundes et al. (2007), also analyzing several loci from modern humans, find no sign of interbreeding. Some specific loci exhibit patterns suggestive of an introgression event such as the tau MAPT locus (Baker et al., 1999), or the *microcephalin* and *ASPM* gene (Evans et al. (2005), Mekel-Bobrov et al. (2005) but see Currat et al. (2006)). In other words, similar to the fossil based studies, genetic based studies often arrive at quite different conclusions.

Large quantities of Neanderthal nuclear genome were released in 2006 by Noonan et al. (2006) and Green et al. (2006). Unfortunately, these studies were not able to answer the

question unequivocally. In particular they produced conflicting results with respect to the admixture question. The data of Noonan et al. (2006) is compatible with 0% to 39% of admixture while the data of Green et al. (2006) is compatible with 81% to 100% of admixture, see Wall and Kim (2007). This discrepancy was probably due to some amount of contamination in Green et al.'s dataset (Wall and Kim, 2007).

Last year, a draft of a complete Neanderthal genome was published (Green et al., 2010), with evidence for admixture between Neanderthals and non-Africans being strongly argued for. The conclusion is based on a statistic (D), that compares two human genomes to the Neanderthal, using chimp as an outgroup. The authors show that the Neanderthal is closer to non-Africans than to Africans. They measure around $1-4\%$ admixture, assuming admixture happened 50,000 to 80,000 years ago.

The goal of the current work is to estimate the time of admixture and the amount of admixture under a simplistic demographic scenario with a single event of admixture at some time, t_{admix}, in the past, and with f describing the admixture proportion. We perform a re-analysis of the draft Neanderthal genome aligned to two Africans and two European complete genomes. The use of additional human data should afford greater power to estimate demographic parameters than the Dstat (Durand et al., 2011) used in Green et al. (2010).

In the Methods section (section 4.2), we first describe the derived SFS, followed by the inference scheme and finally we describe the likelihood function. In the Results section (section 4.3) we analyze the Neanderthal genome data with the high coverage 1,000 genome trios, show the theoretical expectations of the spectrum with admixture, and estimate the maximum likelihood parameters. We conclude with section 4.4, discussing several future directions but also caveats.

Throughout this work, we will fix some demographic features to estimated values in published articles. In particular we will follow closely Wall et al. (2009), in order to simplify the problem. Note that this has a certain circularity since Wall et al. used also African and non-African data to estimate their demographic parameters, and that those estimated parameters depend in turn on the admixture proportion.

4.2 Method

4.2.1 Site frequency spectrum

The site frequency spectrum is a summary of single nucleotide polymorphism (SNP) data. It contains information about past demographic or selection events in a population. Assuming a sample of n chromosomes, we can denote by X_i the number of sites where the derived allele is present in i copies among n. For example the number of sites per locus where the derived allele is present in 1 copy (singletons) will be denoted X_1, the number of sites where the derived allele is present in 2 copies, X_2 etc. We define the site frequency spectrum SFS as the set of means over loci of $X_1, X_2, ..., X_{n-1}$, $SFS = (E(X_1), E(X_2), ..., E(X_{n-1}))$, i.e. we have $SFS_i = E(X_i)$.

The higher the mutation rate the higher the value of $E(X_i)$ for all i. One intuitive way to see that is to think in terms of coalescent theory and of the tree that describes the sample (see e.g. Wakeley (2008)). The SFS_i is proportional to the length of branches with i descendants

in a sample. If we assume an infinite sites model, then the number of mutations is going to be thrown on the tree following a Poisson process with a rate proportional to the branch lengths of the tree and to the rescaled mutation rate, θ. If we denote by μ the mutation rate and N_e the effective population size, then we have $\theta = 4N_e\mu$, the rescaled mutation rate. The rescaled mutation rate is then a scaling factor for the SFS. For example, under the standard model of constant population size and no selection, the SFS is defined by $SFS_i = \frac{\theta}{i}$ (e.g. Griffiths (2003)).

This definition generalizes easily to M populations. In this case, we look at the number of sites with i_1 derived alleles in P_1 among n_1 chromosomes and i_2 derived alleles in P_2 among n_2 chromosomes, that we can denote X_{i_1,i_2}. Then the SFS is characterized by $SFS_{i_1,i_2} = E(X_{i_1,i_2})$.

4.2.2 Theory for the derived SFS

Without admixture

In the absence of any admixture, Chen et al. (2007) published several results for the joint SFS between two populations, P_1 and P_2. In particular they derived several results when there is only one sample available in P_1.

We denote by $SFS^{der} = (SFS_{i_1=1,i_2}) = (SFS_{i_1=1,1}, SFS_{i_1=1,2}, ..., SFS_{i_1=1,n_2})$, respectively $SFS^{anc} = (SFS_{i_1=0,i_2})$, the joint SFS between P_1 and P_2 when the one sample in P_1 has the derived (respectively ancestral) allele. We call those spectra the derived SFS (SFS^{der}) and the ancestral SFS (SFS^{anc}).

Let us assume that the populations P_1 and P_2 diverged at time $\mathcal{T}$ generations in the past. Then for the derived spectrum they obtained a very simple formula:

$$SFS_{i_1=1,i_2} = \frac{\theta e^{-t(\mathcal{T})}}{n_2 + 1} \tag{4.2.1}$$

where $0 < i_2 < n_2$ and the rescaled time $t(\mathcal{T}) = \int_0^{\mathcal{T}} \frac{dt'}{2N(t')}$. In other words, the joint derived SFS is constant for all i_2. This result is useful because it is robust to population size changes in the two populations, i.e., the derived SFS will always be flat no matter the history of the population size is. The only thing that changes is the expected number of sites for each category.

The expression for SFS^{anc} is more complex and we are not reproducing it here since we are not going to use the SFS^{anc} for inference. From now on we will focus upon the SFS^{der}.

With admixture

In this paragraph, we follow Gutenkunst et al.'s notation. The effect of admixture on the SFS^{der} is not obvious. One way to get the expected values is to take a diffusion approach, which is a continuous approximation to a Wright-Fisher population genetics model in discrete generations (e.g. Durret (2010)). For two populations, we can denote by $\phi(x_1, x_2, t)$ the joint density of derived allele at frequencies x_1 and x_2 at time t in P_1 and P_2. In order to find ϕ for a particular demographic scenario without selection or migration between populations, one can solve:

$$\frac{\partial}{\partial t}\phi = \frac{1}{2}\left(\frac{\partial^2}{\partial x_1^2}\frac{x_1(1-x_1)}{\nu_1} + \frac{\partial^2}{\partial x_2^2}\frac{x_2(1-x_2)}{\nu_2}\right)\phi \tag{4.2.2}$$

where $\nu_i = \frac{N_i}{N_{ref}}$ the relative effective size of population i with respect to a reference population size N_{ref}. We can pick $N_{ref} = N_e$, the effective population size of the first population for example. We will assume one admixture event. If the admixture event is from $1 \to 2$ forward in time, this corresponds to creating a population P_3 made of a mixture of P_1 and P_2 and then removing P_2. We denote by f the proportion of P_3 coming from P_1 and by t_{admix} the time of admixture. We have the population creation $\phi(x_1, x_2, x_3, t_{admix}) = \phi(x_1, x_2, t_{admix})\delta(x_3 - [fx_1 + (1-f)x_2])$, where δ is the Dirac delta function. At the same time, we remove the population $\phi(x_1, x_3, t_{admix}) = \int_0^1 \phi(x_1, x_2, x_3, t_{admix})dx_2$.

The joint density ϕ at present is obtained by solving equation 4.2.2 and intergating up to t_{admix}, then transforming ϕ as mentioned above for the admixture event and finally solving again equation 4.2.2 and integrating up to present. To get the joint SFS at present, t_{pres}, we integrate the density ϕ at time t_{pres} over all possible frequencies. This allows to go from population frequencies, to sample frequencies. Assuming a binomial sampling of the alleles and n_1 samples in P_1 and n_2 samples in P_2, we have:

$$SFS_{i_1,i_2} = \int_0^1\int_0^1 \binom{n_1}{i_1} x_1^{i_1}(1-x_1)^{n_1-i_1} \binom{n_1}{i_2} x_2^{i_2}(1-x_2)^{n_2-i_2}\phi(x_1, x_2, t_{pres})dx_1dx_2.$$

The derived SFS, as introduced above, is

$$SFS^{der} = SFS_{i_1=1,i_2} = \int_0^1\int_0^1 x_1 \binom{n_1}{i_2} x_2^{i_2}(1-x_2)^{n_2-i_2}\phi(x_1, x_2, t_{pres})dx_1dx_2.$$

Gutenkunst et al. (2009) implemented in $\partial a \partial i$ a method to numerically solve equation 4.2.2 and compute the joint SFS for up to three populations for an arbitrary number of samples per population, and for an arbitrary set of demographic events. We will use their method throughout and present results with v.1.5.2 of $\partial a \partial i$.

4.2.3 Inferring demographic parameters from the SFS^{der}

A likelihood approach

We will use $\partial a \partial i$ to find the maximum likelihood estimates (MLEs) for a particular demographic scenario. The program computes the likelihood of the demographic parameters given the observed SFS, under a particular demographic scenario. Since Gutenkunst et al. found the likelihood calculation to be deterministic and numerically smooth, they use an optimization algorithm to find the MLEs for the parameters.

We can denote Λ the parameters of the demographic scenario, SFS^{obs}_{1,i_2} the observed derived SFS and $SFS_{1,i_2} = SFS_{1,i_2}(\Lambda)$ the expected derived SFS. If all the sites are independent (i.e. if there is no linkage) each entry of the SFS is an independent Poisson variable. Then the likelihood of the parameters for the two populations as discussed above is

$$\ell(\Lambda) = \log\left(\prod_{i_2=1}^{n_2-1} \frac{e^{-SFS_{1,i_2}(\Lambda)} SFS_{1,i_2}(\Lambda)^{SFS^{obs}_{1,i_2}}}{SFS^{obs}_{1,i_2}!}\right).$$

In this work, we want to estimate two parameters $\Lambda = (f, t_{admix})$, the admixture proportion f and the time of admixture t_{admix}. To do so, we will compute the MLEs of those parameters:

$$(f^{mle}, t^{mle}_{admix}) = \underset{f,t_{admix}}{\text{argmax}}(\ell(\Lambda)).$$

Since the SNPs are linked in the data that we use (genome wide dataset), the likelihood is a composite likelihood. As discussed in Wiuf (2006), composite likelihood estimators are consistent estimators for a wide range of neutral coalescent models.

Interval of confidence (CI)

The MLEs are not normally distributed around the true value of the sample and we cannot use the Fisher information for example to compute the confidence interval of the MLEs. Moreover, as discussed below, standard likelihood ratio test statistics will not necessarily follow a χ distribution even if the number of samples is high. We discuss below how we will compute the confidence interval under these circumstances.

In order to compute confidence intervals, we could use a parametric approach and simulate data with recombination that mimics the real data, using for example ms (Hudson (2002)). The problem with such an approach is, because we use genome wide data, that it is hard to separate the data into unlinked loci. Instead, we follow the approach used in Green et al. (2010) and use a nonparametric approach. That is, block jackknife (e.g. Kuensch (1989)) to estimate the variance of the MLEs. We follow closely Green et al. (2010)'s strategy. We cut the genomes in $M = 100$ blocks. We replicate the inference step a 100 times each time removing one block. That is, for a parameter $\lambda \in \Lambda$, we get 100 maximum likelihood estimates λ_{-1}, λ_{-2}...λ_{-100} and denote $\widehat{\lambda_J}$ the mean of those estimates. To compute the standard deviation of a parameter λ we can then use the following formula:

$$\sigma_{\hat{\lambda}} = \left[\frac{M-1}{M}\sum_{i=1}^{M}(\widehat{\lambda_{-i}} - \widehat{\lambda_J})^2\right]^{1/2}.$$

4.2.4 Demographic scenarios for early human evolution

Simple model

In this work, we investigate several very simplistic models of recent human evolution, hoping that they capture the main features well enough to produce a reasonable estimate of admixture proportion and time of admixture between Neanderthals and humans.

We will express time in two units: the (unlabeled) coalescent units, and the number of years. The correspondence between the two is $0.1 = 50,000$[years]. Indeed we assume a generation time of 25[years/generation], and a constant effective population size $N_{ref} = N_e = 10,000$ for all populations. The coalescent time is the number of generations divided by $2N_e$.

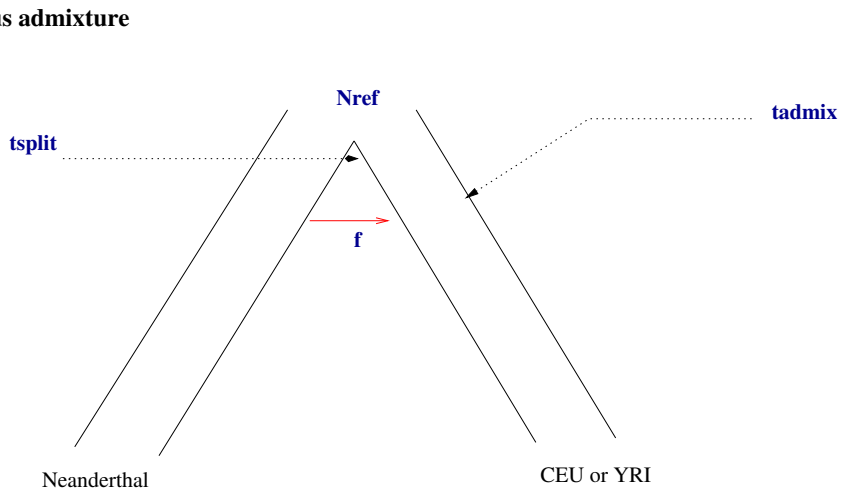

Figure 4.2: **Simple model**. Model with a population split and one admixture event. The population size is fixed to $N_{ref} = 10,000$. The split time is fixed to $t_{split} = 0.8 = 400,000$[years]. The goal is to estimate the admixture time t_{admix} and the admixture proportion f.

The first model we will consider, is also the simplest one. We assume that two populations split at some time t_{split} in the past and that there is a single event of admixture from Neanderthals into humans going forward in time. The admixture happens at time t_{admix} and the admixture proportion is f. Such a model is shown in Figure 4.2. In order to estimate $\Lambda = (t_{admix}, f)$ we will fix $t_{split} = 400,000$[years] following Wall et al. (2009). We refer to this model as the simple model from now on.

Adding a bottleneck

The second model will be considered for the European population only. It follows closely Wall et al. (2009) who, as others before (Schaffner et al., 2005; Voight et al., 2005), found evidence for a bottleneck in Europeans. As before, the two populations split at t_{split}, and there is a single bottleneck in the human population with an event of admixture as above. See Figure 4.3 for notation. These authors fix the split time between Neanderthals and modern humans to $t_{split} = 0.8 = 400,000$[years]. They impose a bottleneck of fixed duration of 1,000 years. They estimate the start of the bottleneck to $t_B^{start} = 0.072 = 36,000$[years] and the strength of the bottleneck to 0.005 fold decrease, $\nu_B = \frac{N_{Bottleneck}}{N_{ref}} = 0.005$. Note that the rationale is that only the ratio of the bottleneck duration and the bottleneck strength matters, so co-estimating those variables is hard because they are strongly correlated. Here we set the bottleneck parameters to the fixed or estimated parameters in Wall et al. (2009), i.e. we fix $t_B^{start} = 0.072$, $t_B^{dur} == 0.002$, $\nu_2^B = 0.005$ and, as before, we keep $t_{split} = 0.8$. From now on, we refer to this model as the bottleneck model.

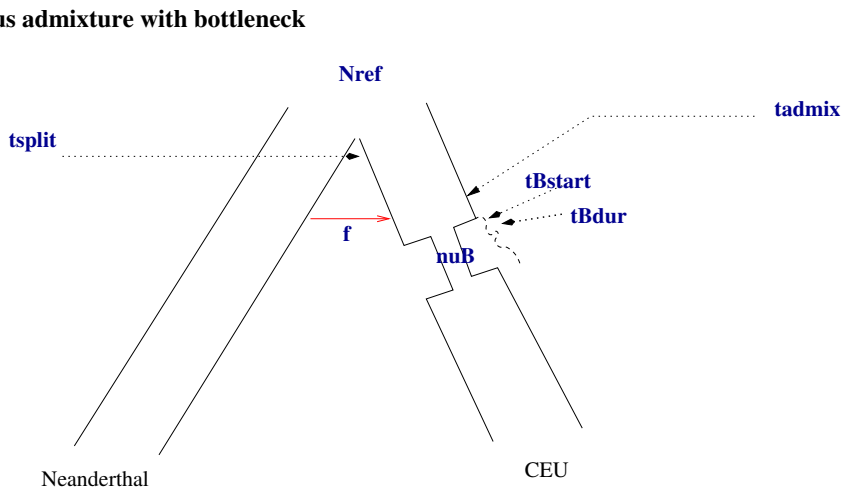

Figure 4.3: **Bottleneck model**. Simple model with a bottleneck event. The population size is fixed to $N_{ref} = 10000$. The times are fixed to $t_{split} = 0.8 = 400,000$[years], $t_{B_{start}} = 0.072 = 36,000$[years], $t_{B_{dur}} = 1000$[years] and the bottleneck parameter $\nu_B = 0.005$. The goal is to estimate the admixture time t_{admix} and the admixture proportion f.

Adding growth

The third model will be considered for the African population only. Indeed, Wall et al. (2009) found evidence for growth in the African population but not the European population. Others before them had found evidence of growth (Pluzhnikov et al., 2002; Adams and Hudson, 2004). They assume that at a time t_g the population started growing until it reached a population 100 fold larger at present and that the growth happens in an exponential manner. Similar to the other two models, we will fix the parameters they found for the growth and estimate the admixture proportion and admixture time for the Africans. So we fix the split time $t_{split} = 0.8$ and for growth in the African population the parameters are fixed to the MLEs of Wall et al. (2009), i.e., assuming the present day population is 100 times bigger than the ancestral population, $t_g = 0.07 = 35,000$[years], (see Figure 4.4). From now on we refer to this model as the growth model.

Comparing demographic models

Ideally, we would want to compare the different demographic models. For nested models, one way is to do a likelihood ratio test (LRT). For proper likelihoods, the likelihood test ratio statistics follows a χ^2 distribution.

But for composite likelihoods, the distribution of the likelihood ratio is not known. In some cases, it has been shown that a LRT is anti-conservative, rejecting the null model more often than it should (Bustamante et al. (2001)). The intuition behind this result is that linkage reduces the number of independent observations, increasing the variance. If linkage is ignored, we underestimate variance for the site frequencies. This can lead to falsely rejecting the null hypothesis, and hence render the test anti-conservative, in the cases where the alternative hypothesis leads to an increase in variance as well.

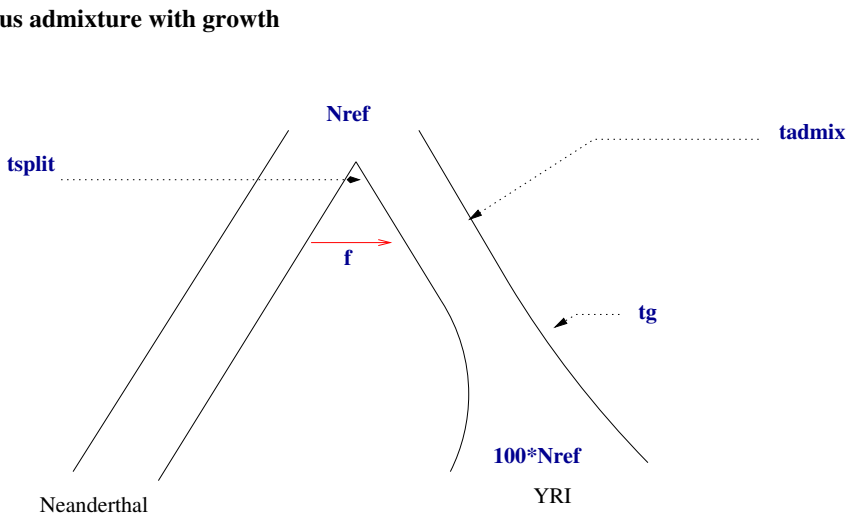

Figure 4.4: **Growth model**. Simple model with a growth event. The population size is fixed to $N_{ref} = 10,000$. The times are fixed to $t_{split} = 0.8 = 400,000$[years] and $t_g = 0.07 = 35,000$[years] and the population size of YRI at present is assumed to be 100 times bigger than the ancestral population. Moreover, we assume an exponential growth. Again, the goal is to estimate the admixture time t_{admix} and the admixture proportion f.

One way to go around this is to compute the distribution of the test statistics by simulating under the null model and computing the LRT for each replicate. In our case, one would need to know the how to adjacent sites are linked for the whole genome. More precisely one would need to separate the genome into unlinked loci, and model some recombination within locus. We will not attempt this here.

For non-nested models, the situation is similar. For proper likelihood one can compute the Akaike Information Criterion (AIC) (Akaike, 1974). Models with small AIC are preferred. The AIC is a difference between the maximum log-likelihood and the number of parameters. Unfortunately, we do not know if the properties of the AIC remain valid for composite likelihood and therefore do not attempt to use it.

To conclude, in this work, we will compare models only qualitatively, by comparing the residuals between the best fit model and the data for example.

4.2.5 Effects of sequencing error on the Neanderthal genome

Since we can compute the expected spectra, we can also look at the effect of errors in the DNA sequence. The goal is to see what is the qualitative effect of a certain amount of error. Since the Neanderthal genome data we use in this work is 1-fold coverage (see below), it is certainly incorrect to assume that the Neanderthal is errorless. We consider here the simplest possible error: uniform probability of error on the Neanderthal sequence. Indeed, the Neanderthal is likely to be damaged through deamination related to degradation of the sample as well as sequencing error. Note that uniformity is a simplifying assumption since it has been shown that the deamination is not uniform through the read and mostly cytosines are affected, see for example Briggs et al. (2007).

The effect of sequencing error is to wrongly attribute a site to the derived SFS instead of the ancestral SFS and vice-versa. We denote SFS_ϵ^{der} and SFS_ϵ^{anc} the derived and ancestral SFS when error is taken into account, and ϵ the probability of error:

$$SFS_\epsilon^{anc} = \epsilon \cdot SFS^{der} + (1-\epsilon) \cdot SFS^{anc} \tag{4.2.3}$$

and

$$SFS_\epsilon^{der} = \epsilon \cdot SFS^{anc} + (1-\epsilon) \cdot SFS^{der}. \tag{4.2.4}$$

4.2.6 Data

Neanderthal genome

We use the Neanderthal genome published in Green et al. (2010). We use the data from the three bones found in the Vindija cave and labeled Vi33.16, Vi33.25 and Vi33.26. The bones Vi33.16 and Vi33.26 were carbon dated to $38,310 \pm 2,130$[years] and $44,450 \pm 550$[years] respectively. The bone Vi33.25 was found in a layer older than the other two, but was not carbon dated. The complete mtDNA genome for Vi33.16 and Vi33.25 were sequenced Green et al. (2008); Briggs et al. (2009), and found to be different at 10 positions. The bones Vi33.16 and Vi33.26 have indistinguishable mtDNA but the age of the bones and the difference in nucleotide diversity between bones versus within bones support the hypothesis that they are from two different individuals (Green et al., 2008). Therefore, it is likely that the Neanderthal genome we use is a composite genome of three individuals.

We only use the autosomal chromosomes (the three Neanderthal individuals are thought to be female, see Green et al. (2010)). We use a similar filter as in the original paper. We keep only reads that have a mapping quality above 30, a base quality above 20 and a total coverage less than six.

Trios of the 1,000 genome project

The analysis of the human samples of the current work could be done with various human datasets of African or European origin. We choose to work with the data published by the 1,000 genome project, see Durbin et al. (2010). In particular we choose to work with the two family trios. Each family is comprised of a daughter, the mother and the father. We use the parents of each family, that is, we have for each population 4 chromosomes of two unrelated individuals. The two populations are Utah residents with Northern and Western European ancestry from the CEPH collection (CEU) and Yoruba, from Ibadan, Nigeria (YRI).

In the 1,000 genome project, the trios have the advantage of being very high coverage. The mean mapped depth is 43.14 in CEU and 40.05 in YRI. We used the vcf files that were made available through the website ((http://www.1000genomes.org/) and more specifically the data released on October 2010[†]. The vcf file for each population includes only the sites that are polymorphic when the individuals of the trios and the reference human sequence are considered together.

[†]ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/release/2010_07/trio/SNPs/

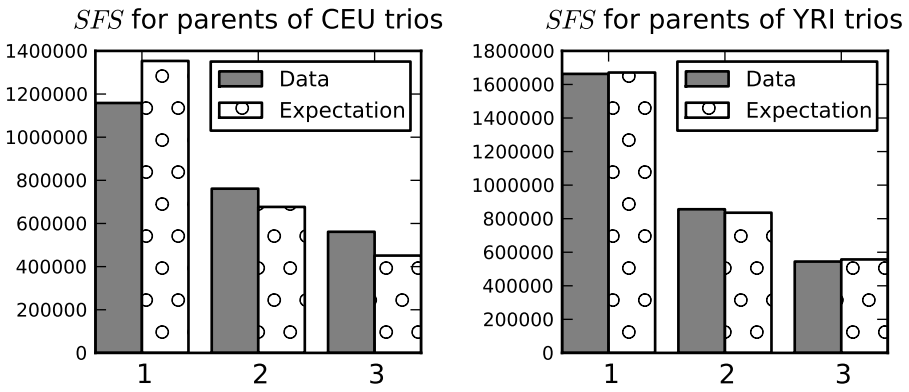

Figure 4.5: $SFSs$ or total number of sites for the CEU and YRI population (right and left respectively). In bubble pattern is the theoretical spectrum for four chromosomes under a standard neutral model ($\propto \frac{1}{i}$ where i is the number of derived alleles). In grey is the observed spectrum. We see that the observed spectra agree closely with the neutral spectrum especially for the YRI.

4.3 Results

4.3.1 SFS in CEU and YRI

Because of the high coverage in the trio data (more than 40 fold), we assume that the SNP calling process is errorless. We limit ourselves to the sites in the vcf files and therefore we exclude the non-polymorphic sites for both CEU and YRI population. We found 2,481,331 polymorphic sites for the CEU and 3,063,614 polymorphic sites for the YRI population.

Figure 4.5 gives the SFS for the parents of the CEU and YRI. From now on we will refer to the CEU parents as CEU and the YRI parents as YRI. Encouragingly, the observed $SFSs$ agree closely with the standard neutral model.

4.3.2 Joint site frequency spectrum N×CEU and N×YRI

We filter the Neanderthal genome and keep sites that have a map quality above 30, a base quality above 20 and a total coverage less than or equal to 6. Indeed, high coverage sites are more likely to result from mapping artefacts. We are left with about 1.5×10^9 base pairs. We then align the Neanderthal to the CEU dataset and the YRI dataset, lifting their pre-existing alignment to to human reference genome.

We want to calculate the SFS of N×CEU and N×YRI. The Neanderthal genome is low coverage (about 1 fold) and we verify that this is the case also for sites where the YRI or CEU are polymorphic. We plot the coverage in Figure 4.6. The distribution of coverage is very similar for the two populations, as expected. Because it is low coverage, it is hard to get genotype information with confidence. Instead, we sample one read per Neanderthal, a proxy for one Neanderthal chromosome. Damage and sequencing error are a serious problem in this context. We do not try to address in the present work. We assume that the one chromosome we sample is errorless. Initially, we ignore error in the Neanderthal sequence.

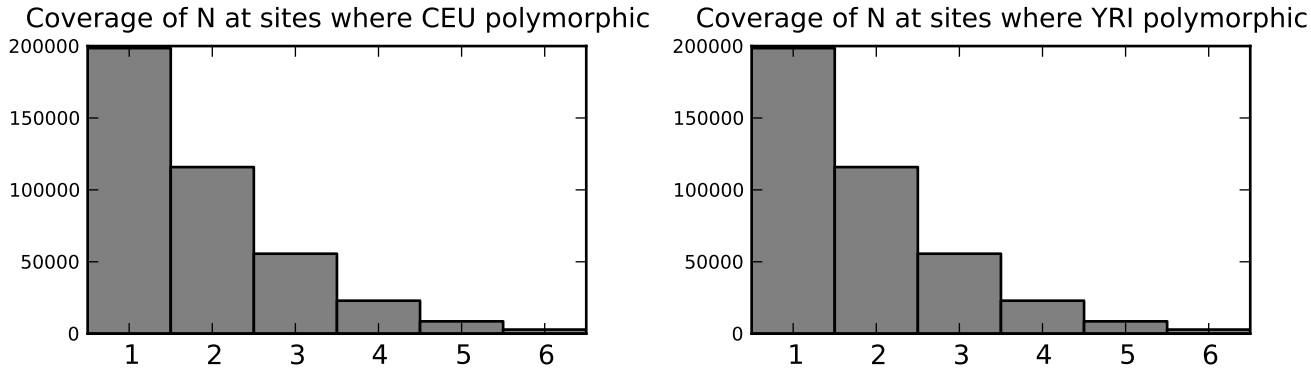

Figure 4.6: Distribution of the coverage of the Neanderthal for sites polymorphic for CEU on the left and polymorphic for YRI on the right. Note that more than 90% of the sites have up to 3 fold coverage.

We build two spectra of size 1×3. The SFS^{der} is shown in Figure 4.7.

4.3.3 Derived spectrum: expectations

We compute the numerical SFS^{der} for different values of t_{admix} and f using $\partial a \partial i$ in order to build some intuition of the effect of the different demographic events on the spectrum. The parameter θ is a scaling factor and we fix it to $\theta = 1$. We consider below that the admixture has to have happened earlier than $45,000$ years to be in agreement with the archeological record (see Introduction), i.e. at least approximately $0.09 = 45,000$[years] ago and more recently than the split between the Neanderthal and human populations.

The human and Neanderthal samples do not have the same age. This should be, in principle, taken into account. But since we have only one Neanderthal chromosome and we consider only the segregating sites in humans, the age of the Neanderthal sample has no impact on the observed spectrum for the demographic models considered. In fact, the missing branch length only contributes to sites derived in Neanderthal but not polymorphic in humans.

Note that it is unlikely that the admixture proportion is higher than 25% based on mitochondrial data results (e.g. Nordborg (1998), Serre et al. (2004)). We plot various spectra for a reasonable range for the parameters, i.e. $0 \leq t_{admix} \leq 0.4$ and $0 \leq f \leq 0.2$ in the following subsections.

Simple model

The simple model is shown in Figure 4.2. The expected spectra are shown in Figure 4.8.

As expected, without admixture the SFS^{der} is constant. The effect of having an admixture event is an overall increase in the expected number of SNPs derived in the Neanderthal and polymorphic in humans.

For higher f or smaller t_{admix}, the excess in singletons, doubletons or tripletons is higher. The effect on f is quite intuitive, the higher f the more likely it is that the Neanderthal

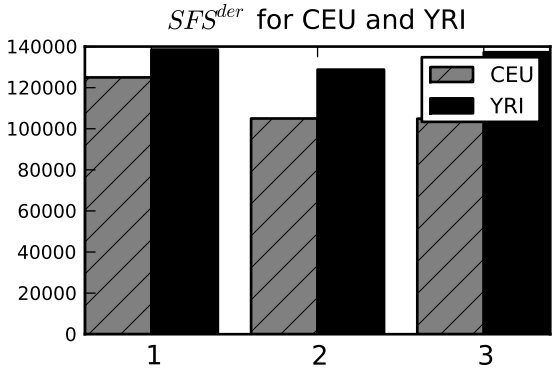

Figure 4.7: The SFS^{der} for CEU and YRI for sites where the Neanderthal has a coverage up to 6, a base quality of 20 and a map quality of 30. The spectrum is plotted for sites with 1, 2 or 3 derived alleles from left to right in each human population.

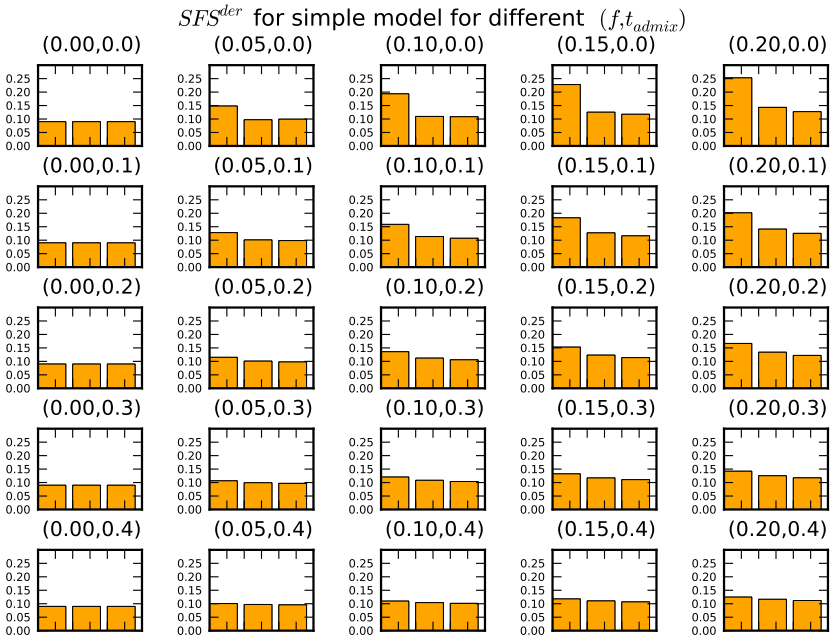

Figure 4.8: Theoretical SFSs for the simple model shown in Figure 4.2 when different values of the t_{admix} and f for $\theta = 1$.

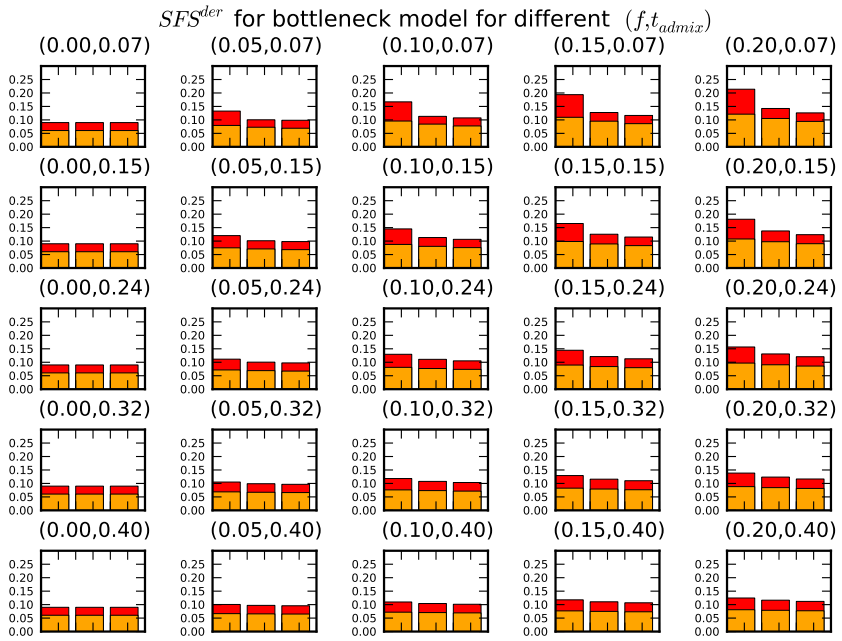

Figure 4.9: Theoretical $SFSs$ when the admixture happens before the bottleneck ($0.072 < t_{admix} < 0.8$). In red is the excess of SNPs of the simple model, in orange is the overlap between the simple model and the bottleneck model.

and the humans are in the same population before all four humans coalesce. The effect of t_{admix} is a bit less intuitive perhaps. It is easier to think in terms of the singletons. The older the time of admixture the more likely two humans will have coalesced by the time the Neanderthal migrates into the human population, decreasing the number of singletons. At the limit, $t_{admix} = t_{split}$, the spectrum should be flat but with a lower constant (see equation 4.2.1).

Bottleneck model

The demographic model is shown in Figure 4.3. We compare the case with a bottleneck with the case without. We looked separately at the case where the admixture happens before the bottleneck ($0.072 \leq t_{admix} \leq 0.4$, Figure 4.9), during the bottleneck ($0.07 \leq t_{admix} \leq 0.072$, Figure 4.10) or after the bottleneck ($t_{admix} \leq 0.07$, Figure 4.11). The legend is as follows: in orange, the overlap between the simple model and the bottleneck models, in red is the excess of SNPs of the simple model.

As expected, if there is no admixture, the spectrum is constant but that constant is different depending on the bottleneck. The overall effect is a loss of segregating sites. In particular, if the admixture happens before the bottleneck ($t_{admix} > t_{B_{start}}$) the excess of singletons, doubletons and tripletons, due to the admixture are lost for the most part. If the admixture happens after the bottleneck the effect is an overall decrease in segregating sites, but the sites unique to the admixture remain. To match the observed CEU data for example, the intuition is that either the admixture is fairly old with strong admixture or more recent

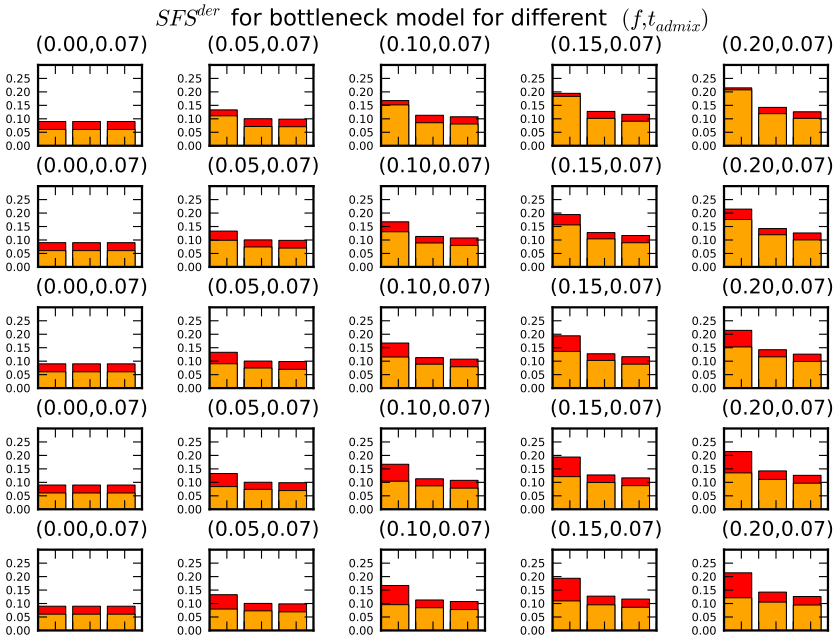

Figure 4.10: Theoretical $SFSs$ when the admixture happens during the bottleneck ($0.070 < t_{admix} < 0.072$). In red is the excess of SNPs of the simple model, in orange is the overlap between the simple model and the bottleneck model.

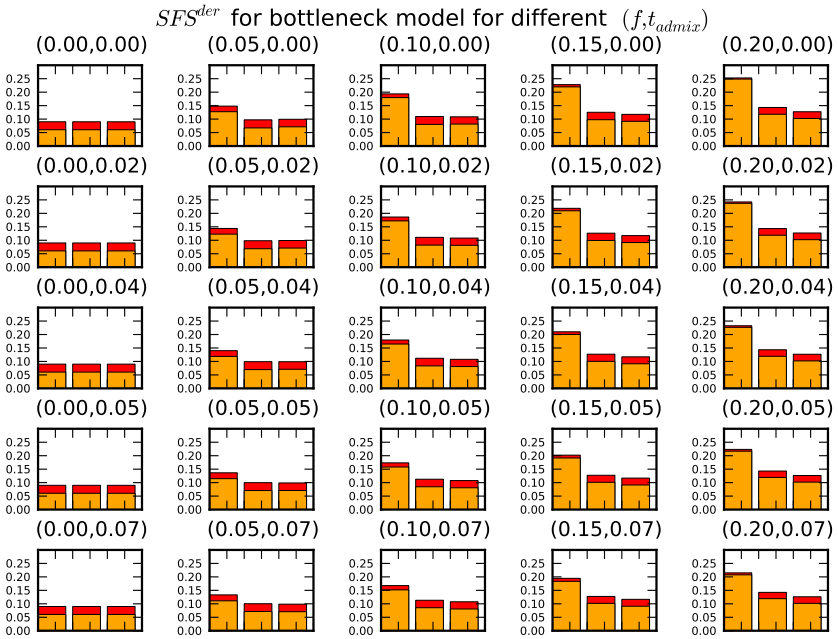

Figure 4.11: Theoretical $SFSs$ when the admixture happens after the bottleneck ($0 < t_{admix} < 0.070$). In red is the excess of SNPs of the simple model, in orange is the overlap between the simple model and the bottleneck model.

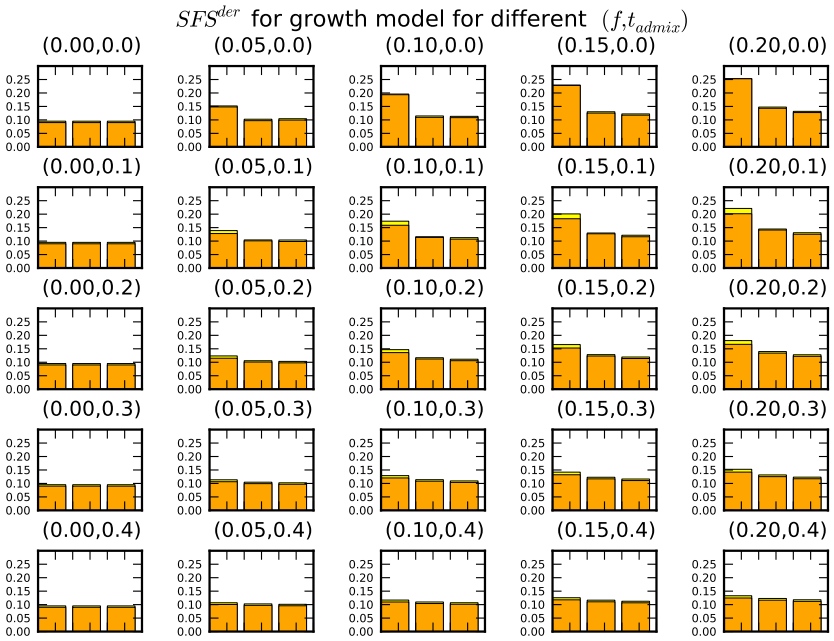

Figure 4.12: Theoretical $SFSs$ for growth model. In yellow is the excess of SNPs of the growth model, in orange is the overlap between the simple model and the growth model.

with weaker admixture. In other words the timing of the bottleneck has an impact on when the admixture occurred.

Growth model

The demographic model with growth is shown in Figure 4.4. The expected spectra are shown in Figure 4.12. We compare the case with growth with the case without growth. The legend is as follows: in orange the overlap between the simple and the growth model, in yellow is excess of SNPs of the growth model.

The overall effect is an excess of singletons. In particular the effect is stronger if the admixture happens just before the growth starts. This can be explained because the population size increases, slowing the rate of coalescence, leading to more singletons. But in general, the spectra are similar to those observed without growth. If there is no admixture, the spectrum remains flat, as expected by the theory, and there is very little effect of the growth event if the admixture event is very recent or very old.

4.3.4 Derived spectrum: parameter estimates

Likelihood function

We plot here the likelihood function for each population for the different demographic scenarios. The likelihood was computed over a 50×50 grid of t_{admix} and f values. These plots are helpful in determining that there is only one maximum, and also in determining that the optimization algorithms land on the maximum. We consider the different demographic

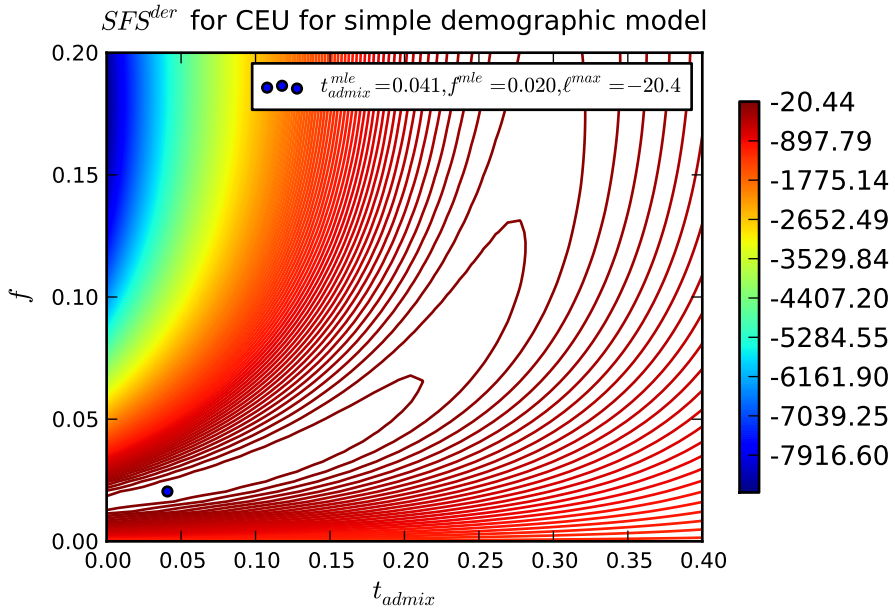

Figure 4.13: Likelihood surface for the simple model for $f \in (0, 0.2)$ and $t_{admix} \in (0, 0.4)$ for the CEU population.

scenarios separately and list the maxima found. We reserve the discussion of the actual values to the next section where we run the optimization algorithm, and find values closer to the actual maxima.

First we plot the results for the simple model. As seen in Figure 4.13 and 4.14 the maxima are found for CEU at $t_{admix} = 0.041$ and $f = 2.0\%$ and for YRI at $t_{admix} = 0.000$ and $f = 0.04\%$. Note that those times are not compatible with the archaelogical evidence of overlap between modern humans and Neanderthals.

Then we plot the results for the bottleneck model. The likelihood surfaces are shown in Figures 4.15, 4.16 and 4.18. We separate the cases where the admixture happens before, during, or after the bottleneck to be able to find the global maximum. Indeed, we expect a discontinuity. The maximum was found for the case where the admixture happens more recently than the bottleneck, that is $t_{admix} = 0.057$ and $f = 1.2\%$. Again, the time is not compatible with the archaelogical record.

Finally we looked at the growth model. The likelihood surface is plotted in Figure 4.18. The maxima for the sparse grid are found for $t_{admix} = 0.0001$ and $f = 0.4\%$. Again, the time is not compatible with the archaelogical record.

To conclude this section we note that for the simple model and the growth model, there is a single maximum. On the other hand, for the bottleneck model there are local maxima for the times of admixture before the bottleneck, during the bottleneck, and after the bottleneck. This discontinuity suggests that the timing of bottleneck has a strong impact on the parameter estimates, as mentioned above.

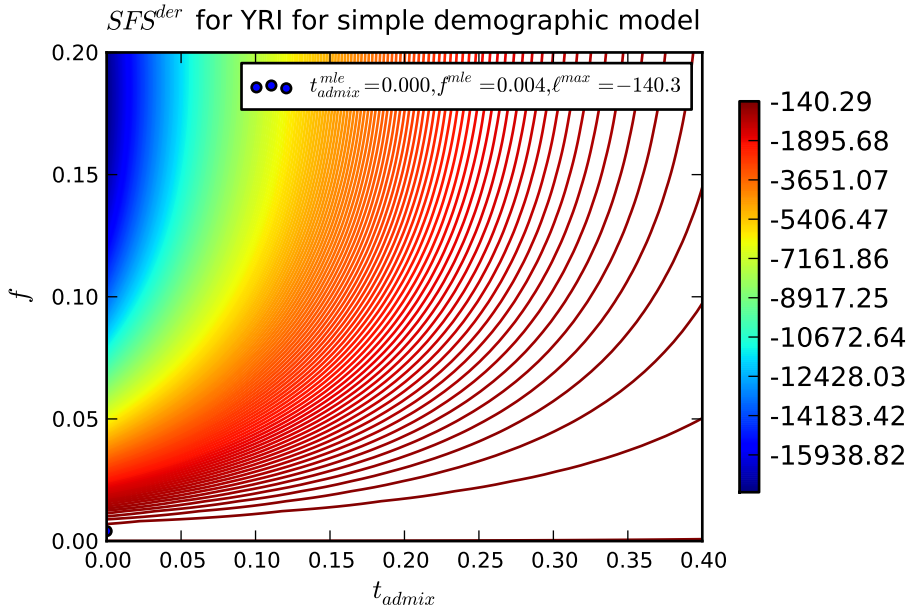

Figure 4.14: Likelihood surface for the simple model for $f \in (0, 0.2)$ and $t_{admix} \in (0, 0.4)$ for the YRI population.

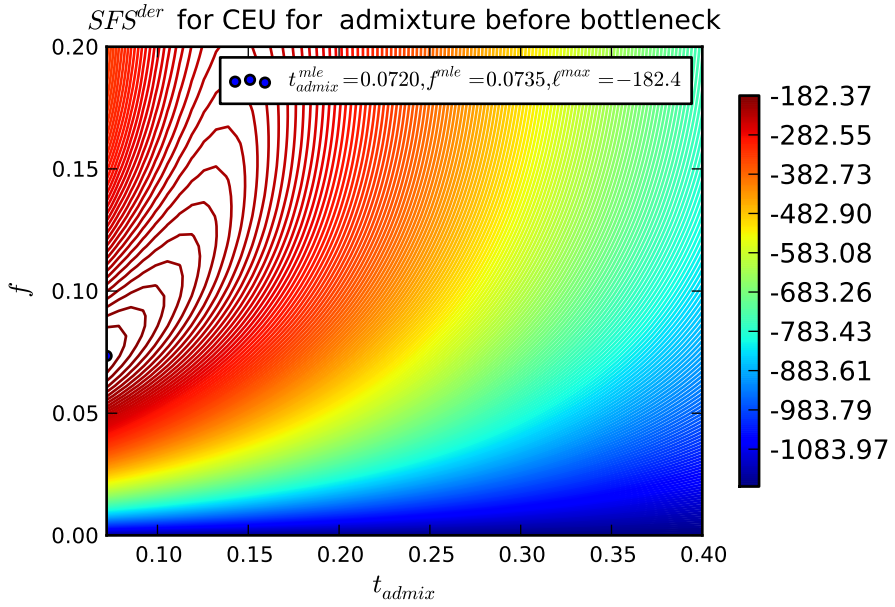

Figure 4.15: Likelihood surface for the bottleneck model if the admixture happens before the bottleneck ($0.072 \leq t_{admix} \leq 0.4$) for the CEU population.

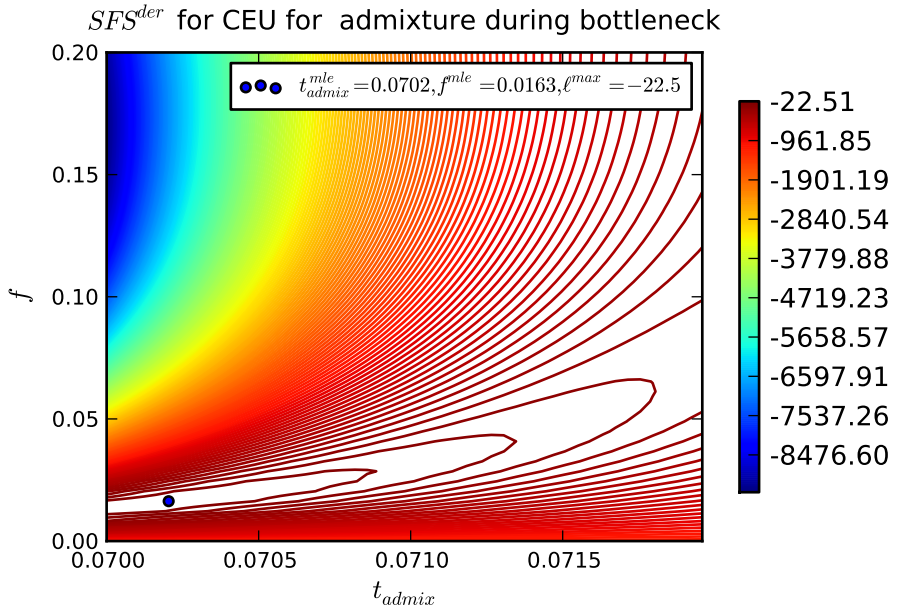

Figure 4.16: Likelihood surface for the bottleneck model when the admixture happens during the bottleneck ($0.070 \leq t_{admix} < 0.072$) for the CEU population.

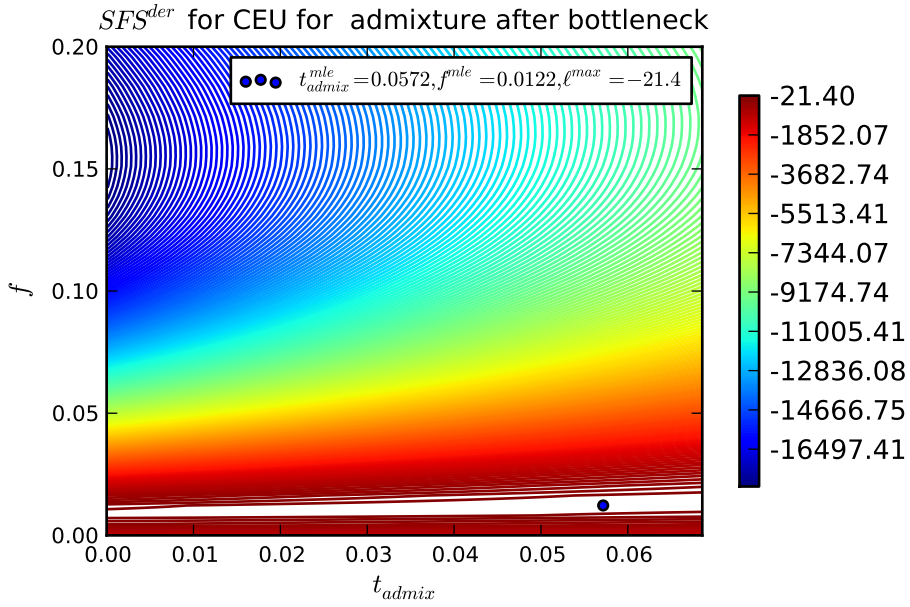

Figure 4.17: Likelihood for the bottleneck model when the admixture happens after the bottleneck ($0 \leq t_{admix} < 0.070$) for the CEU population.

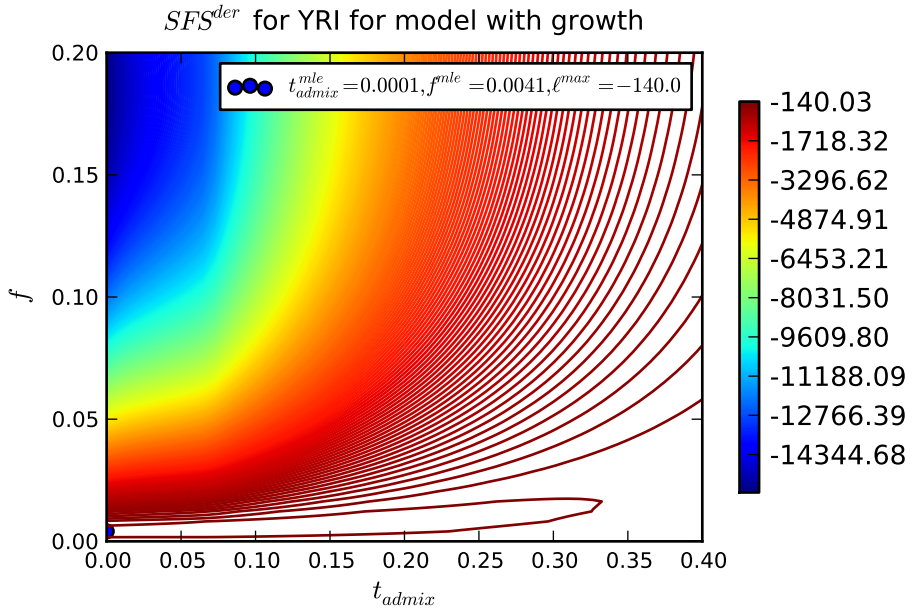

Figure 4.18: Likelihood surface for growth model for $f \in (0, 0.2)$ and $t_{admix} \in (0, 0.4)$ for the YRI population.

MLEs and confidence intervals for t_{admix} and f

In order to find the MLEs, we used the optimization algorithms that are implemented in $\partial a \partial i$ [‡]. For the CEU results we used the Nelder-Mead simplex algorithm (Nelder and Mead, 1965). For the YRI population, we saw in the previous section that the optima are close to the boundary of the parameter space. For most results, we used therefore the L-BFGS-B bounded algorithm (Byrd et al., 1995; Zhu et al., 1997). In some cases, the algorithm did not converge and we use the Nelder-Mead simplex algorithm instead.

We assume the parameters have values in the intervals $f \in (0, 1)$ and $t_{admix} \in (0.09, 0.8)$. The time of admixture is constrained by archaelogical evidence and the population split time. The confidence intervals were computed as described in section 4.2.3.

We plot the distribution of f_{-i} and $t_{admix_{-i}}$ for each jackknife replicate in Figure 4.20 for each demographic model and population. The standard deviation of those distributions allows to build a 95% confidence interval (also shown in Figure 4.20). All the MLEs estimates and the standard deviation are given on Table 4.1. We also plot the spectra of each jackknife replicates (see Figure 4.19). Each replicate consists of 99/100 of the original data.

The first thing to note is that the maxima are consistent with the surface of the previous section, suggesting that the optimization algorithms converge to the right point.

We also see that the spectra for the jackknife replicates are very similar to the original dataset. Consistently, we see that the distribution around the time of admixture is very narrow, at the lower bound of the parameter range, i.e. 0.09. This is the case for YRI

[‡]relying on the *fmin* and the *fmin_bfgs* methods of *scipy* (Jones et al., 2001b)

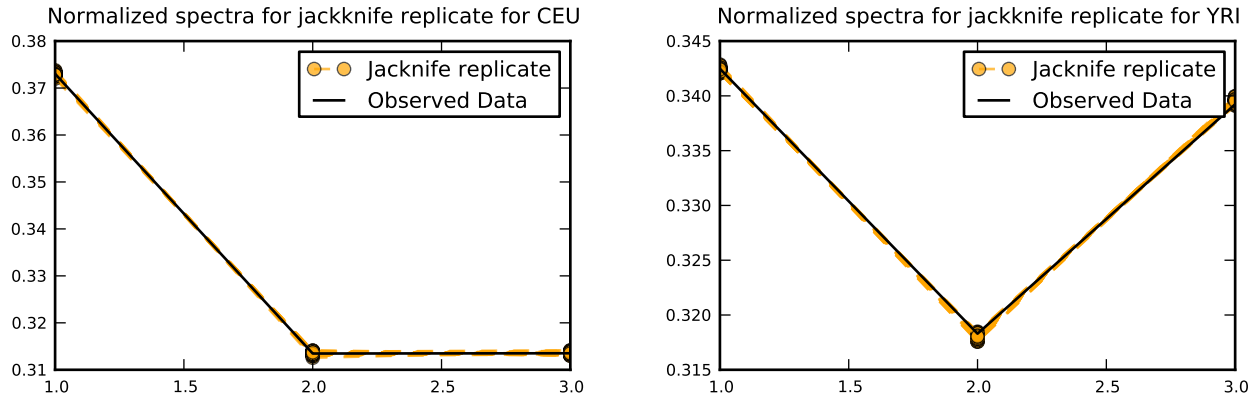

Figure 4.19: The SFS^{der} for each jackknife replicate is shown in orange and the SFS^{der} for the data is shown in black. Each spectrum is renormalized since the replicates have 99% of the original dataset. We see very little variation between replicates; there are actually 100 lines for the jackknife replicates.

and CEU for all demographic models. The value of 0.09 is consistent with the plots of the likelihood surface of the previous section. The maximum likelihood for the admixture time depends on the lower bound of the parameter range. Therefore, those results suggest that if those demographic models and the archaeological evidence are valid, the admixture happened shortly before $45,000$[years] for both YRI and CEU.

For the CEU population, the admixture proportion's 95% confidence interval is $(2.3\%, 3, 3\%)$ for the simple model and $(6.5\%, 10.7\%)$ for the model with bottleneck. Both intervals do not include the 0%. We can therefore conclude that, if the demographic model is realistic, there is evidence of admixture between CEU and Neanderthals. Note that, the two intervals do not overlap. The first is consistent with the conclusion of Green et al. (2010). The latter study found f to be between 1% and 4%. The second estimate we get with the bottleneck is closer to the estimate of Wall et al. (2009). These authors found evidence for admixture of about 14%. The bottleneck model is directly inspired from this second study. Therefore it is reasonable that we get closer estimates. In Wall et al. (2009), they also fix t_{admix} to a higher value of $50,000$[years]. This explains in part the fact that we get a lower value and we therefore conclude that our results for CEU are broadly consistent with previous studies.

For the YRI population we find the admixture proportion to be much lower, i.e. lower than 1%, but the confidence interval does not include 0%. Indeed, the admixture proportion's 95% confidence interval is $(0.3\%, 0.7\%)$ for the simple demographic model and $(0.3\%, 0.6\%)$ for the model with growth. As for the CEU, we can conclude that there is evidence for admixture between YRI and Neanderthals. This result was not directly discussed in Green et al. (2010). Moreover, this demographic model was not directly tested by Wall et al. (2009), and it is therefore hard to compare our results with theirs. In their paper, they assumed the admixture happens within the non-African population only. We could perhaps reconcile the two results by fixing a similar model with admixture with CEU only and migration between YRI and CEU. In fact such a model should lead to evidence of admixture for both CEU and YRI, depending on the migration rate, but a lower level with YRI (when the two

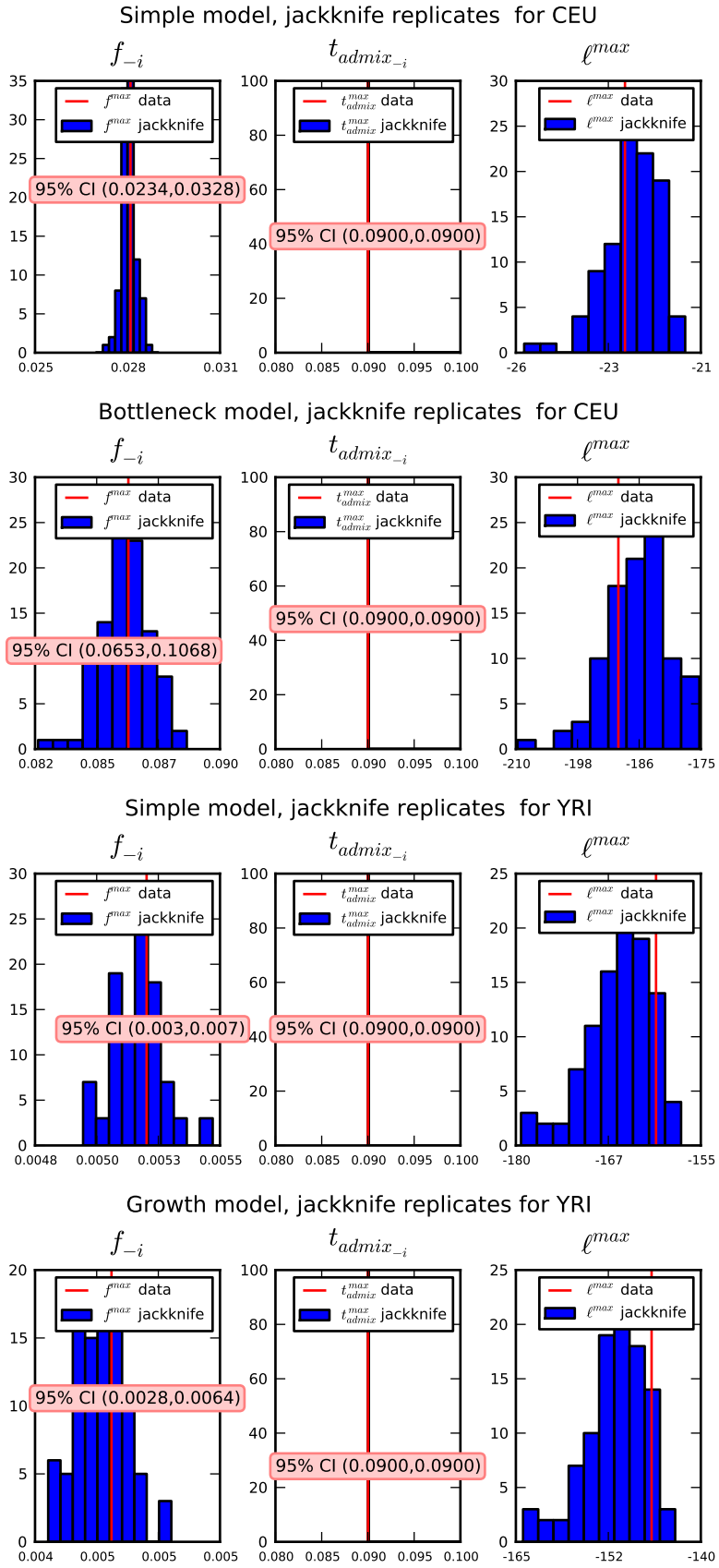

Figure 4.20: Distribution of MLEs for each replicate of the jackknife procedure. Each estimate is for a dataset for which a block was removed. In red is the maximum likelihood estimate for the whole dataset and the interval is the 95% confidence interval ($1.96\sigma_{-\Lambda}$).

YRI	f^{mle}	$\sigma_{f^{mle}}$	t^{mle}_{admix}	$\sigma_{t^{mle}_{admix}}$	ℓ^{max}
simple model	0.52%	0.09%	0.090 (45000[years])	0.000	-161.05
growth	0.46%	0.09%	0.090 (45000[years])	0.000	-146.65

CEU	f^{mle}	$\sigma_{f^{mle}}$	t^{mle}_{admix}	$\sigma_{t^{mle}_{admix}}$	ℓ^{max}
simple model	2.8%	0.2%	0.090 (45000[years])	0.000	-23.05
bottleneck	8.6%	1.1%	0.090 (45000[years])	0.000	-190.60

Table 4.1: MLEs and maximum likelihood for input values$f \in (0, 1)$ and $t_{admix} \in (0.09, 0.8)$.

populations are considered separately). To test a model like that, one would need to build the joint spectrum for all three populations, Neanderthals, YRI, and CEU. This would be an interesting future work.

We now plot for each model the resulting expected spectra from the best fit models (see Figure 4.21) and, the Anscombe residuals (see Figure 4.22) for each model. The latter are a normalized difference between the data and the model.

We can compare the spectra qualitatively. We see that the expected spectra, the maximum likelihood values, and the residuals suggest that, for the YRI, the model with growth is a better fit. It has a higher likelihood but also smaller residuals. Both overestimate the middle frequency SNPs and underestimate the high frequency SNPs.

For the CEU, the model without the bottleneck is actually a better fit. The simple demographic model for the CEU seems to provide a very good fit, while the bottleneck model overestimates the middle frequency SNPs and underestimates the low and high frequency SNPs. Both for the YRI and the CEU the two alternative models are not nested.

Validation: comparison with ms simulations

Another way to see qualitatively how the model fits the data is to simulate some new data with the same demographic model using the MLEs. We can use ms (Hudson, 2002) to simulate such data. We can then find the maximum likelihood for each simulation and compare this distribution to the values for the observed data. If the model is a good fit for the data, then the maximum likelihood of the observed data should not be smaller than most simulation replicates.

Ideally, one would want to simulate data with linkage between sites. The effect of linkage is to decrease the number of independent loci, i.e. to decrease the number of independent observations. Since we do not try to assign a significance value, we will assume that there is not linkage between sites.

We perform 400 simulations. We pick a θ value for each demographic scenario such that there is, on average, about one SNP per independent locus. We simulate sequence data so that the total number of polymorphic sites in CEU or YRI, where the Neanderthal is derived, is the same for the simulations and the observed data. For the YRI population there are about 400,000 such SNPs and for the CEU population about 330,000. For the simple demographic scenario we have to simulate about 4,300,000 loci for the YRI and about 3,200,000 for the CEU. For the demographic scenario with bottleneck, we simulate about 3,400,000 loci for the CEU. And finally for the demographic scenario with growth in

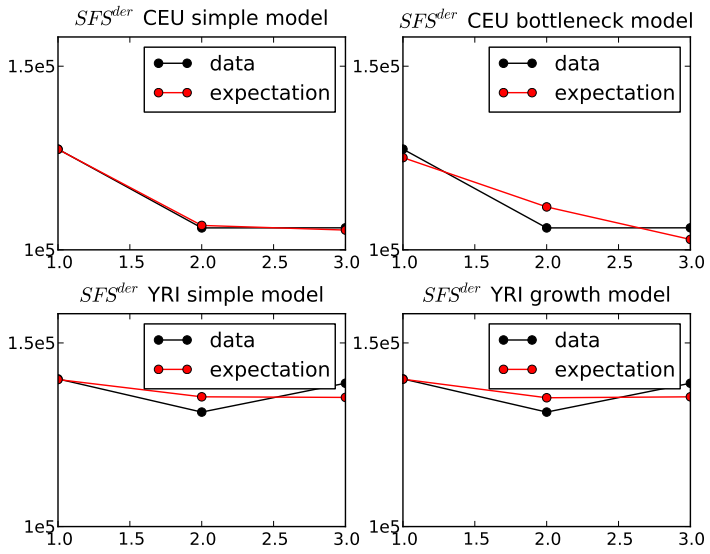

Figure 4.21: Comparison between the expected spectra from the models with the demographic models with the maximum likelihood estimates compared to the data.

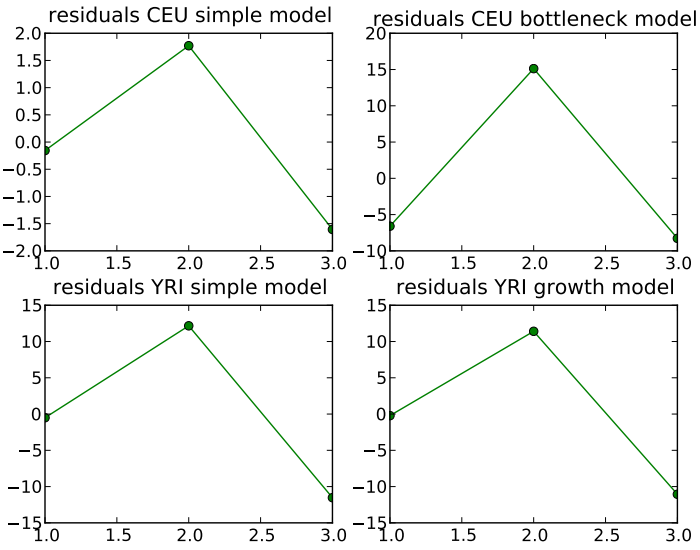

Figure 4.22: Ascombe residuals for the different demographic models for the CEU and the YRI data.

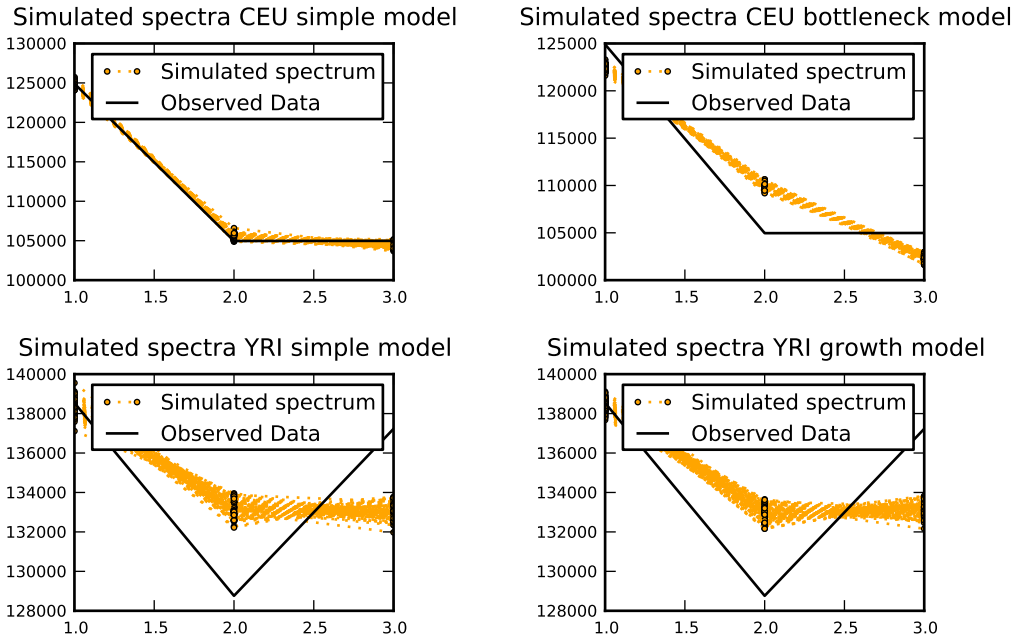

Figure 4.23: Spectra for the ms simulations. In orange the simulations and in black the observed spectrum. On top the CEU population and at the bottom the YRI with the two possible demographic scenarios for each.

YRI, we simulate about 4,200,000 loci. If we had modeled the right linkage, the number of independent loci would have been smaller, therefore the variance of the distribution would have been wider. Because of that, a confidence interval based on some simulations would be too narrow.

The results are shown in Figures 4.23, 4.24 and 4.25. They are consistent with the previous section. That is, the simple model for the CEU is a reasonable fit but the other models are not explaining the data very well. We see that first on the simulated spectra (Figure 4.23). Indeed, the observed spectra is within the range of simulated spectra for the simple model for CEU only. Then, we also see that, looking at the maximum likelihood distribution. Although we cannot assign a significance value, the observed maximum likelihood for the observed data is much smaller for all simulations. Finally, Figure 4.25 is a validation of the method itself (i.e. a validation of $\partial a \partial i$ for our particular demographic models). In fact, we recover the input parameters. On this Figure, we see that there is a strong correlation between the parameters. This result is intuitive and we already observed it in section 4.3.3.

On the other hand, the variance is high and somehow inconsistent with the interval of confidence we built with the non parametric jackknife method. This can already be expected from the spectra from the ms simulations of Figure 4.23 that have a higher variance than the spectra from the jackknife replicates in Figure 4.19. Indeed, the actual confidence interval based on these simulations should be larger if we had modeled the right linkage. It is not clear why this is the case and further work is necessary for both the jackknife replicates and the ms simulations to try to reconcile the results.

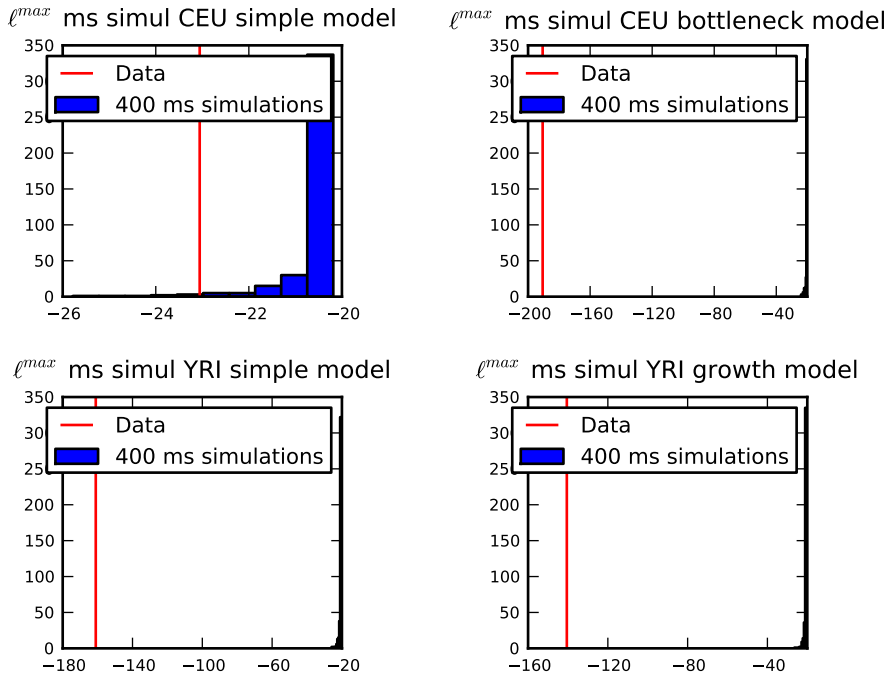

Figure 4.24: Distribution of the MLEs and ℓ^{max} for the ms simulations for each demographic model and population (CEU at the top, YRI at the bottom). For all but the simple model for CEU, the observed likelihood value is much smaller than the values of the ms simulations, suggesting that those three models do not explain the data well.

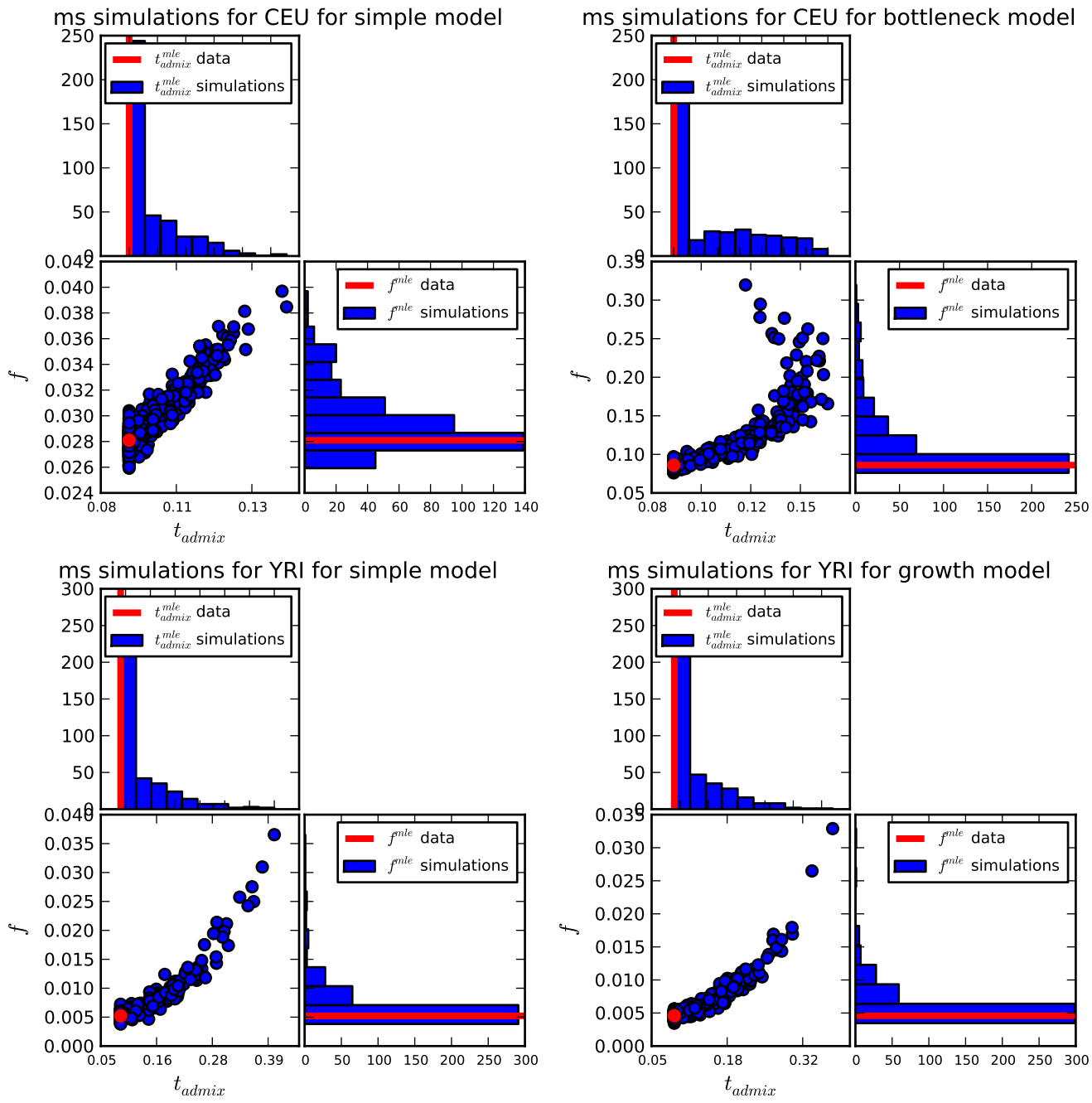

Figure 4.25: Distribution of the MLEs for the ms simulations for each demographic model and population. At the top the results for the CEU population and at the bottom the results for the YRI population. In all cases we observe a correlation between the two parameters. Indeed, intuitively and as observed in the theoretical spectra (see e.g. Figure 4.8), one can obtain similar results for high t_{admix} and f or low t_{admix} and f.

4.3.5 Effect of sequencing error on f

First we plot the expectation for the spectra with or without error, assuming as above $\theta = 1$, different values of admixture proportion and time of admixture for the simple model. We look at four error probabilities: $\epsilon = 0.01, 0.02, 0.1, 0.2$. The graphs are shown in Figure 4.26 for error probabilities of 0.01 and 0.02 and in Figure 4.27 for error probabilities of 0.1 and 0.2. The effect of mislabelling the Neanderthal becomes very large for error of the order of 10%. In all cases, we see an excess of all categories, with the excess being higher for smaller frequency. We can therefore expect an overestimate of the admixture proportion (or underestimate of admixture time).

We quantify the effect of sequencing error on f and t_{admix} separately. We look at the simple model only.

First we fix $t_{admix} = 0.09$ and quantify the effect on f. Then, for several true values of f, $f^{truth} \in \{0, 0.02, 0.04, ..., 0.2\}$, we build the extected spectrum and infer f back. We then compute the expected spectra SFS^{der}, SFS^{anc}. Assuming some amount of error ϵ, we compute SFS^{der}_{ϵ} and SFS^{anc}_{ϵ}. For each f^{truth} we infer f^{mle} using the resulting spectra as input data.

The corresponding graph is in Figure 4.28. The effect is an overestimate of the admixture proportion even for 1% or 2% error for all values of the truth, as could be expected from the spectrum. What is maybe surprising is that there is a "saturation" effect. For high error probabilities, the estimate is always about 0.2 for all values of the truth, that is, the effect of the error dominates for all admixture proportion.

For the effect on t_{admix} we proceed similarly. We fix $f = 0.04$ and then compute the MLE for t_{admix} assuming some amount of error ϵ.

The results are shown in Figure 4.29. As could be foreseen from the expected spectra, the t_{admix} is underestimated if ignoring error, even for an error of the order of 1%. Again there is a saturation effect, and the MLE for t_{admix} is 0 for all values of the truth when the error is above 10%.

In conclusion, to get an accurate estimate of f and t_{admix} based on the derived SFS one cannot ignore the effect of sequencing error. Moreover, if the error is above 10%, it is unlikely we can actually recover the truth at all.

4.4 Conclusion, future work and caveats

The admixture question is relevant to better understand the history of Neanderthals themselves, but is also of importance to define modern humans. The question has been addressed from several angles: archaeological, paleoanthropological and genetical. Interestingly, each field carries opponents and proponents of the admixture question.

The sequencing of the Neanderthal genome has made it possible to directly compare modern humans and Neanderthal on a large scale from a genetic perspective. But thus far, it is safe to say that the findings raise as many questions as they answer (Hodgson et al., 2010).

In this work we infer demographic parameters under a particular demographic scenario related to the *African Hybridization and Replacement Model* described in Stringer (2002).

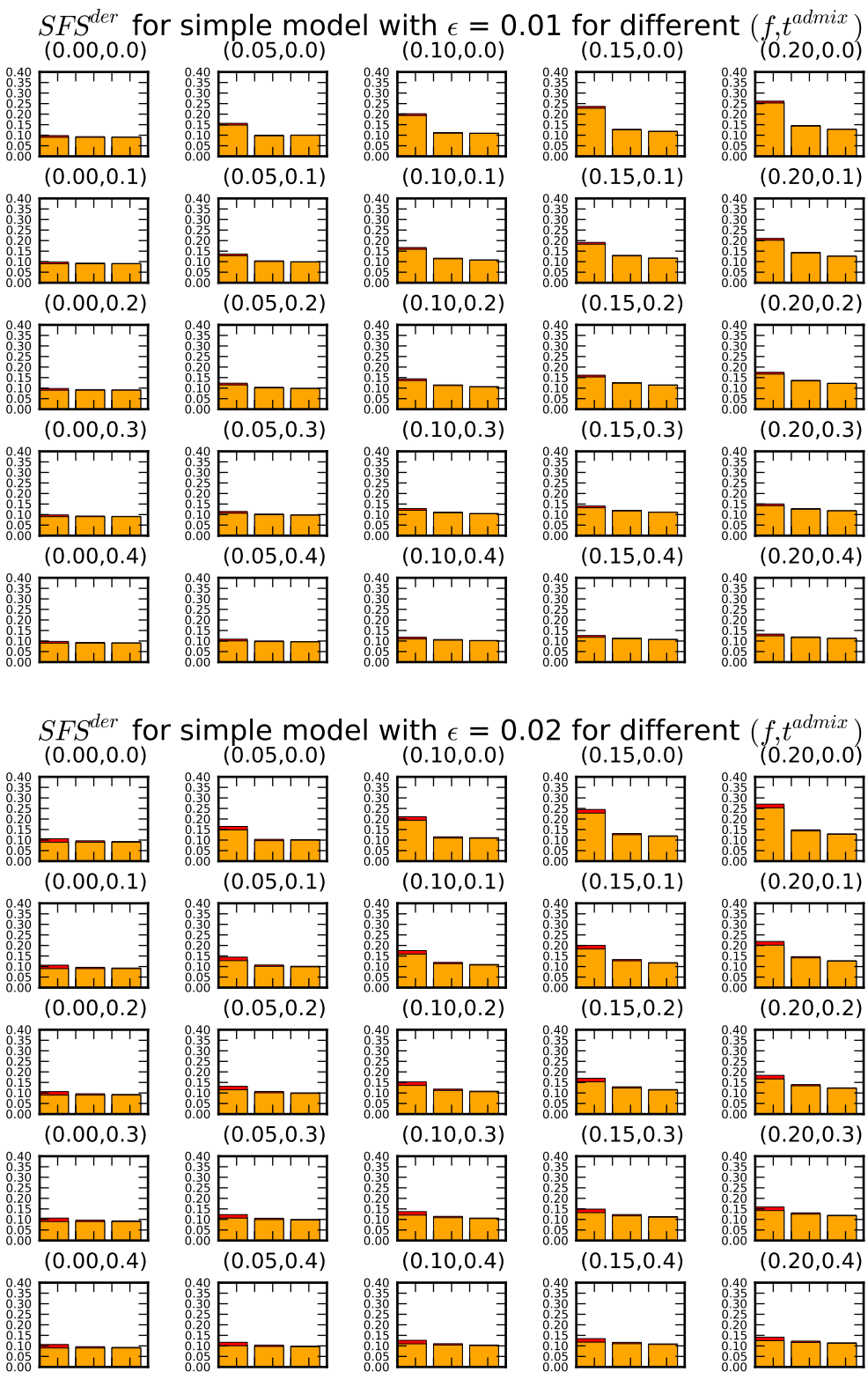

Figure 4.26: Theoretical SFS^{der}_{ϵ}. In orange is the overlap between spectra without error and spectra with error. In red is the excess due to an error of $\epsilon = 0.01$ above and $\epsilon = 0.02$ below.

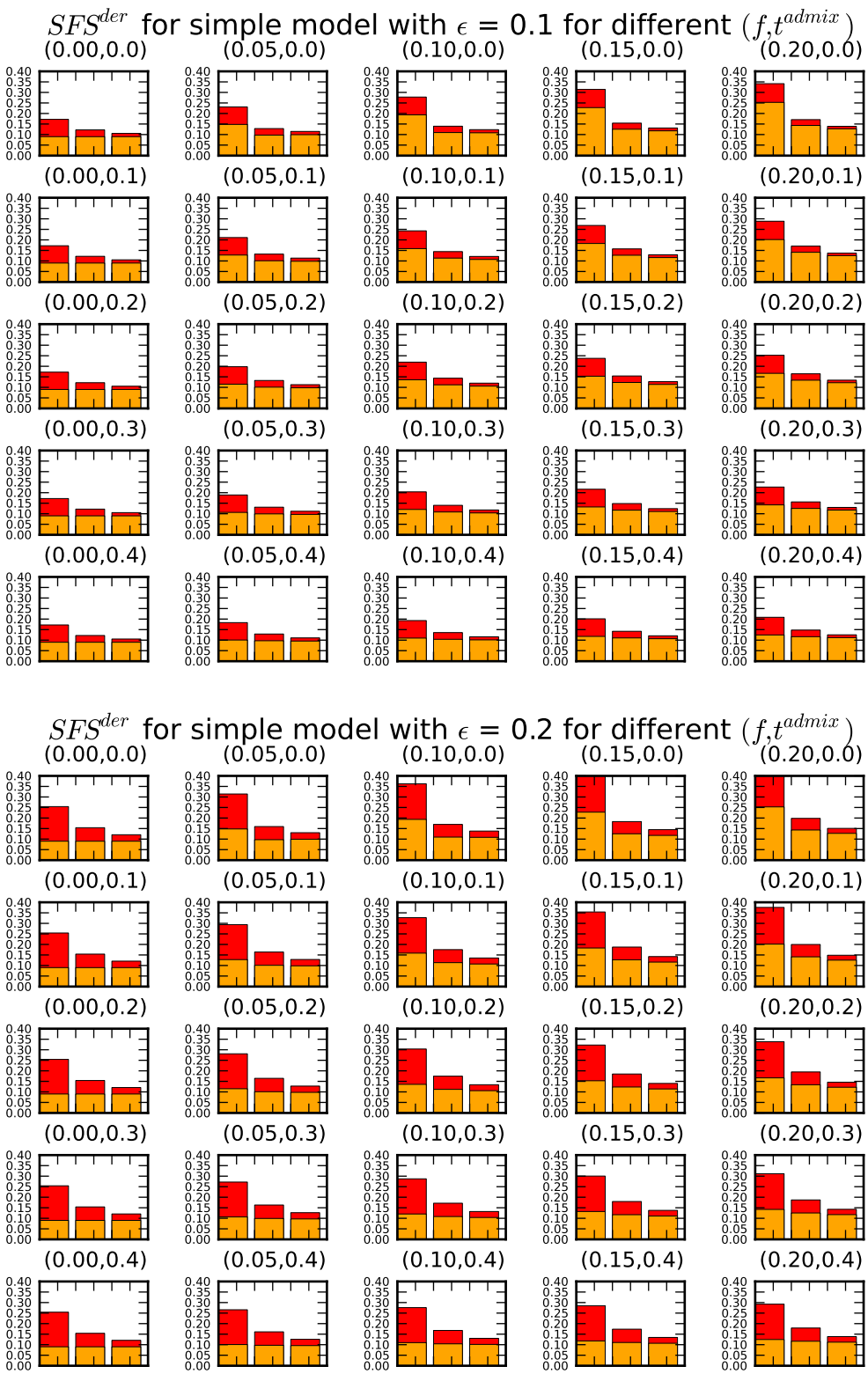

Figure 4.27: Theoretical SFS^{der}_{ϵ}. In orange is the overlap between spectra without error and spectra with error. In red is the excess due to an error of $\epsilon = 0.1$ above and $\epsilon = 0.2$ below.

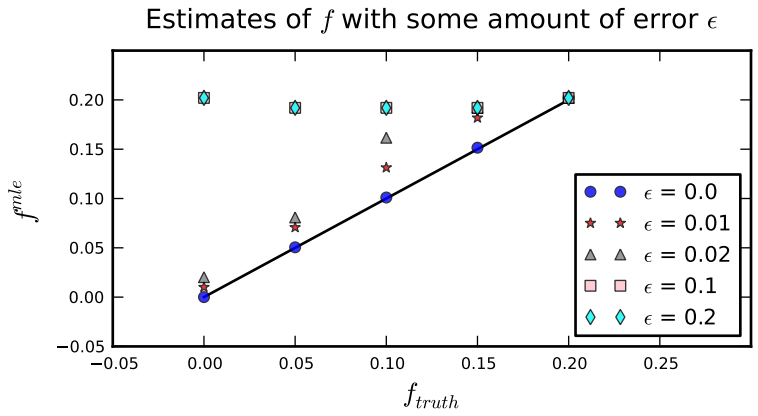

Figure 4.28: Effect of sequencing error on the admixture proportion estimate. On the x-axis the true f and on the y-axis the estimate using $\partial a \partial i$ and an exhaustive search for f values between 0 and 1.

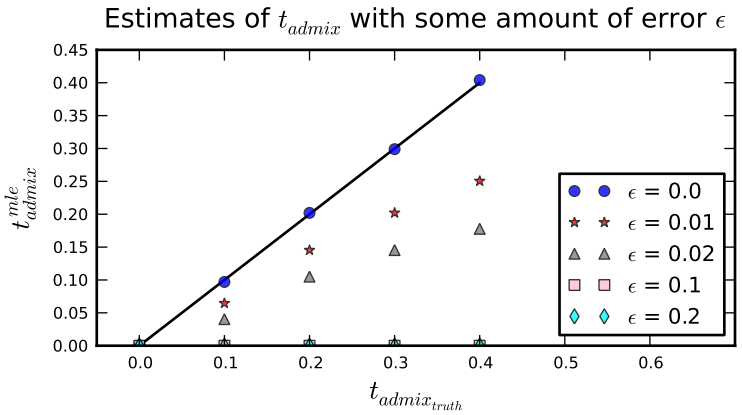

Figure 4.29: Effect of sequencing error on the admixture proportion estimate. On the x-axis the true t_{admix} and on the y-axis the estimate using $\partial a \partial i$ and an exhaustive search for t_{admix} values between 0 and 1.

We do not try to investigate other demographic scenarios that could explain the data (e.g. substructure within Africa (Green et al., 2010). It is therefore important to remain cautious when interpreting our results.

We show that the SFS is a powerful tool to estimate admixture time and admixture proportion for a recent event of admixture. Indeed, the method proposed provides estimates with narrow confidence intervals, even when considering the SFS for four chromosomes only. Note that the method is not new (Gutenkunst et al., 2009), only the application to that particular problem is new.

Assuming that the demographic models are close to reality, we also show evidence for admixture between Neanderthals and Europeans as shown before (e.g. Wall et al. (2009); Green et al. (2010)). We also estimate, to our knowledge for the first time, the age of admixture between Neanderthals and Europeans to about 45,000 years. The amount of admixture was found to be sensitive to the demographic model and ranges between 2% and 11% for the models considered.

We also show evidence for some admixture between Neanderthals and Africans. This result is not obviously consistent with the archaeological record and with past studies. It could be due however to an incomplete demographic model that does not consider Europeans and Africans simultaneously . In fact, it might be that the effect we observe is not due to admixture with Africans but rather to migration between Africans and Europeans after the Neanderthal-European admixture.

This study is a first step into using the joint SFS to estimate parameters of early modern human evolution. Several aspects of the data have been ignored in order to obtain this result. We discuss now several of those aspects and also provide some ideas on how to take them into account for future work.

Two aspects typical of concern when studying ancient DNA are contamination and damage (e.g. Hofreiter et al. (2001)). Both of them may have an impact on the demographic estimates. Damage should have a very similar effect to sequencing error, except that it would affect preferentially certain nucleotides. Contamination on the other hand should be investigated separately and would have an effect that would depend on the contaminants. One way to deal with all three aspects, i.e. sequencing error, damage and contamination, is to build a probabilistic model including all three.

Another aspect is the choice of demographic models and the parameters of those models - a particularly important point given the sensitivity of the estimated admixture proportion on the underlying model. Thus, those demographic models have of course an influence on the parameter estimates. Our estimates depend heavily on archaelogical evidence and on Wall et al.'s estimates. Although the last Neanderthal in Middle East was found around 45,000 years, maybe Neanderthals and modern humans overlapped until later. Therefore our choice to constrain the time of admixture to be older than 45,000 years is disputable. We also demonstrate that the timing of the bottleneck will have an effect on the admixture time in Europeans. On the other hand, the results for Africans seems to be robust to adding a growth event. Ideally one would want to estimate all parameters, including the time of the bottleneck, to get a demographic estimate. The problem is that the more parameters added, the higher the variance. For only four chromosomes, it is unlikely that we are able to estimate many parameters with confidence. There are several promising directions to increase power. The first is to build the joint spectrum for the data we use here. The second is to add

additional human genomes. Moreover, we decided to work with only one chromosome for the Neanderthal, though roughly half of the sites of interest are covered more than once. Potentially we could infer the genotype at each site and work with two chromosomes for the Neanderthal.

Here, we present the results of the derived SFS. We therefore completely ignore part of the data, i.e. the ancestral SFS. This was by choice, given that the ancestral SFS depends on the population histories even without admixture, and is therefore less robust to misspecification of the demographic events. In the future, while we explore other demographic scenarios, it would be of interest to include the results of the ancestral SFS as it also contains information about the admixture event.

One point that remains quite weak in this work are the estimates of the uncertainty on the parameters. As shown, the parametric and non parametric methods produce different results. The two procedures are quite different. The ms simulations model the randomness of the coalescent process while the jackknife simulates the randomness of sampling from a given data distribution. A first direction for future work would be to assess the robustness of the standard deviation estimates depending on the block size of the jackknife. A second direction would be to correctly model the linkage between sites in order to run parametric simulations. This should allow for the calculation of LRTs for the various nested models.

Finally, although the SFS has proven to be quite powerful to test hypotheses in evolutionary biology, it is still a summary of the data and in particular does not contain any information about linkage. It would be beneficial to construct a method that would allow for the joint consideration of linkage among sites. We believe that such a method would afford considerably improved power.

Chapter 5

A likelihood method for jointly estimating the selection coefficient and the allele age for time serial data

Modern population genetics has progressed enormously from the simple one-locus, two-allele models of Fisher, Haldane, and Wright in the early 1930s. The Origins of Theoretical Population Genetics, William B. Provine, 2001.

5.1 Introduction

Time series analysis is widespread in several fields, such as economics (e.g. Hamilton (1994)). The related statistical models deal with a time ordered sequence of observations. Such observations are prevalent in several areas of biology as well. But until recently, time molecular series data was only available for time spanning a few generations in higher organisms. Therefore, in the context of population genetics, time serial data was mostly limited to viral evolution or experimental evolution for samples taken at time intervals of several coalescent units (e.g. Wichman et al. (2005); Bollback and Huelsenbeck (2007); Nelson and Holmes (2007); Gresham et al. (2008)).

With recent advances in DNA sequencing and DNA preparation techniques, the study of extinct and long dead organisms is now entering a new era. Previously limited to short segments of mitochondrial DNA, whole nuclear genomes are now available from several extinct species, thus providing new insights into deep evolutionary history (e.g. Rasmussen et al. (2010); Reich et al. (2010)). Moreover, it is now possible to target specific DNA regions in ancient organisms (e.g. Lalueza-Fox et al. (2007); Ludwig et al. (2009); Rusk (2009)). Therefore, time serial data will become increasingly available for a whole range of organisms allowing one to test evolutionary questions using not only present day samples, but also samples from extinct populations.

Theory to describe the temporal change in allele frequency has existed since the advent of population genetics (e.g. Fisher (1922); Wright (1931)). Although not very common, several statistical methods and estimators to deal with time serial data have been developed. For example, several methods to estimate the change in population size have been published

(e.g. Waples (1989); Williamson and Slatkin (1999); Anderson et al. (2000); Drummond and Rambaut (2007)). More recently, in 2008, Bollback et al. developed a method to co-estimate the effective population size, N_e, and the selection coefficient, s, from temporal allele frequency data. They model the evolution of the allele frequency of a di-allelic locus with a diffusion process that approximates a Wright-Fisher population genetic model (WF). They assume the locus is under constant natural selection that acts on diploid individuals.

Our work is a natural extension to Bollback et al.'s method to allow for the estimation of the allele age, t_0, as well. The age of an allele is the time since the mutation event. Allele age is an omnipresent parameter in population genetics and it is closely linked to the selection coefficient (see Slatkin and Rannala (2000) for a review). Bollback et al. assume that at the first time of sampling the population allele frequency is uniformly distributed. It follows from this assumption that even if the allele was not sampled at the oldest sampling time, it had to be present in the population. In this work, we would like to co-estimate s, N_e and t_0 by computing the likelihood of the data given the parameters.

In section 5.2.1 we explain how we approximate the WF model with a one step process. We then discuss the numerical details of the implementation in sections 5.2.2 and 5.3.1.We show how our method performs based on simulations in sections 5.2.3 and 5.3.2. To conclude, we analyze a dataset of horses for the ASIP locus for samples dating from the Pleistocene up to the present in sections5.2.4 and 5.3.3.

5.2 Materials and Methods

5.2.1 Theory

We assume that there is a single, panmictic population evolving according to a WF population genetic model. Under this model, the frequency of an allele A is a homogeneous discrete-time Markov chain. We denote the Markov chain describing the frequency of the allele A through time by X_t. We assume that selection is constant from the time the allele arose up to present. The allele under selection arises only once and there is no recurrent mutation. In other words, the only evolutionary forces acting on that allele are genetic drift and selection.

Selection is modeled as acting on diploid individuals. If we denote the two alleles by A and a, we can choose the genotypic fitness to be $w_{AA} = 1 + s$, $w_{Aa} = 1 + sh$ and $w_{aa} = 1$ where s is the selection coefficient and h is the dominance coefficient ($s > -1$ and $h \in [0, 1]$, see e.g. Ewens (2004)). If N_e is the effective population size, the states of X_t are the allelic frequencies that we can also write with respect to the population size $x_j = \frac{j}{2N_e}$ for $0 \leq j \leq 2N_e$. Therefore the state space is $\{0, \frac{1}{2N_e}, ..., \frac{2N_e-1}{2N_e}, 1\}$. We define the rescaled selection coefficient $\gamma = 2N_e s$.

We would like to compute the likelihood of the allele age t_0, the rescaled selection coefficient γ and the effective population size N_e. To simplify the notation, let us write $\theta \equiv (\gamma, N_e, t_0)$ for the parameters of interest. Assume we have samples from m distinct sampling time points. We suppose that $M = (n_1, n_2, \ldots, n_m)$ chromosomes were collected, among which $I = (i_1, i_2, \ldots, i_m)$ are of the A type and that the chromosomes where drawn at times $T = (t_1, t_2, \ldots, t_m)$, where time is measured in generations with $t_{k-1} < t_k$ (see

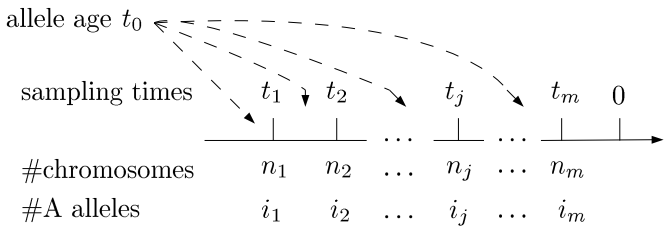

Figure 5.1: Notation used throughout the text. The chromosomes $M = (n_1, n_2, \ldots, n_m)$ are sampled at times $T = (t_1, t_2, \ldots, t_m)$ and there are $I = (i_1, i_2, \ldots, i_m)$ A alleles at each sampling time.

Figure 5.1). Then the likelihood function of the parameters, conditioning implicitly on M and h, is $\ell(\theta) = p(i_1, \ldots, i_m|\theta, T)$.

To compute the likelihood, we can condition and sum over all the population allelic frequencies, $x_{j_1}, \ldots, x_{j_m}$, at each sampling time $t_1, t_2, \ldots, t_m$. We can then rewrite the likelihood:

$$\ell(\theta) = \sum_{j_1} \ldots \sum_{j_m} p(i_1, \ldots, i_m|\theta, T, x_{j_1}, x_{j_2}, \ldots, x_{j_m}) p(x_{j_1}, x_{j_2} \ldots, x_{j_m}|\theta, T). \tag{5.2.1}$$

Conditional on the population allelic frequencies, the number of A alleles i_j at each sampling time are independent of one other. The right side of the summation of equation 5.2.1 becomes

$$p(i_1, \ldots, i_m|\theta, T, x_{j_1}, x_{j_2}, \ldots, x_{j_m}) = p(i_1|x_{j_1}) \ldots p(i_m|x_{j_m}). \tag{5.2.2}$$

In the WF model the population is large and panmictic, therefore we can assume that we sample the chromosomes with replacement, for $k \in \{0, .., m\}$ can be written as:

$$p(i_k|x_{j_k}) = \begin{pmatrix} n_k \\ i_k \end{pmatrix} x_{j_k}^{i_k} (1 - x_{j_k})^{n_k - i_k}. \tag{5.2.3}$$

Since X_t is a Markov chain we have that the left side of equation 5.2.1 as:

$$p(x_{j_1}, x_{j_2}, \ldots, x_{j_m}|\theta, T) = p(x_{j_m}|x_{j_{m-1}}, \theta, T) p(x_{j_{m-1}}|x_{j_{m-2}}, \theta, T) \ldots p(x_{j_1}|x_{j_0}, \theta, T), \tag{5.2.4}$$

where x_{j_0} is the frequency of the allele when it first arose in the population, i.e. $x_{j_0} = \frac{1}{2N_e}$. We can rewrite the transition probabilities of X_t $p(x_{j_k}|x_{j_{k-1}}, \theta, T) = p_{t_k - t_{k-1}}(x_{j_{k-1}}, x_{j_k})$, conditioning implicitely on θ and T. By substituting equation 5.2.4 and 5.2.3 into 5.2.1 we get:

$$\ell(\theta) = p(i_1, \ldots, i_m|\theta, T) = \sum_{j_m=0}^{2N_e} p(i_m|\frac{j_m}{2N_e}) \sum_{j_{m-1}=0}^{2N_e} p_{t_m - t_{m-1}}(\frac{j_{m-1}}{2N_e}, \frac{j_m}{2N_e}) \cdots p(i_2|\frac{j_2}{2N_e}) \sum_{j_1=0}^{2N_e} p_{t_2-t_1}(\frac{j_1}{2N_e}, \frac{j_2}{2N_e}) \cdot p(i_1|\frac{j_1}{2N_e}) p_{t_1-t_0}(\frac{1}{2N_e}, \frac{j_1}{2N_e}). \quad (5.2.5)$$

The solution for the transition probabilities for the non-neutral case of the WF model is elaborate (Ewens (2004) and citations therein). But if we rescale the time by $2N_e$, the Markov chain, X_t, can be approximated by a diffusion process ("WF diffusion process"), Y_τ (see e.g. (Durrett, 2008)). Time is now in units of $2N_e$ generations and is continuous and we replace T by $\mathcal{T} = (\tau_1, \ldots, \tau_m)$ where $\tau_i = \frac{t_i}{2N_e}$. The state space is also continuous with states denoted by $y \in [0, 1]$. This holds in the limit of large N_e, where $X_{[\tau 2N_e]} \simeq Y_\tau$. The transition probabilities of the diffusion process are denoted $p(y_k|y_{k-1}, \theta, \mathcal{T}) = p_{\tau_k - \tau_{k-1}}(y_{k-1}, y_k)$. In this work we approximate the diffusion process itself by a one step process that we denote by Z_τ (see e.g. Van Kampen (1992)). A one step process is a continuous-time Markov chain (i.e. discrete in space and continuous in time) where jumps are only allowed between two states that are adjacent to each other. As before, the states of the process Z_τ are the population allelic frequencies that we denote by $\{z_0, z_1, \ldots, z_{H-1}\}$, where H is an integer. The states are chosen such that z_0 and z_{H-1} are respectively the 0 and 1 allelic frequencies, and they are absorbing states since there is no recurrent mutation. The other states are chosen such that $0 < z_k < 1$ and $z_{k-1} < z_k$ for $0 < k < H - 1$. The infinitesimal generator Q of such a process is a tridiagonal $H \times H$ matrix. By denoting β_i (respectively δ_i) the rate of jumping to the right (respectively the left) of state i, we have that:

$$Q = \begin{pmatrix} 0 & & & \ldots & & & 0 \\ \delta_1 & \eta_1 & \beta_1 & 0 & & & \\ 0 & \ddots & \ddots & \ddots & 0 & \vdots & \vdots \\ & 0 & \delta_k & \eta_k & \beta_k & 0 & \\ \vdots & & 0 & \ddots & \ddots & \ddots & 0 \\ & & & 0 & \delta_{H-2} & \eta_{H-2} & \beta_{H-2} \\ 0 & & & \ldots & & 0 & 0 \end{pmatrix} \quad (5.2.6)$$

where $\eta_k = -(\beta_k + \delta_k)$. The transition probability between two states $z_{j_{k-1}}$ and z_{j_k} of the process is $p_{\tau_k - \tau_{k-1}}(z_{j_{k-1}}, z_{j_k}) = (\exp(Q(\tau_{k+1} - \tau_k)))_{j_{k-1}, j_k}$. With the appropriate choice of β_i and δ_i (see Appendix A), one can show that for large H, $Z_\tau \simeq Y_\tau$. In particular β_i and δ_i will be functions of z_j, z_{j-1}, z_{j+1}, γ and h. Note that Y_τ is a continuous variable whereas Z_τ is discrete. Therefore, choosing $y_{k-1} = z_{j_{k-1}}$ and $y_k = z_{j_k} \notin \{0, 1\}$ we have that:

$$p_{\tau_k-\tau_{k-1}}(y_{k-1},y_k) \backsimeq \frac{p_{\tau_k-\tau_{k-1}}(z_{j_{k-1}},z_{j_k})}{\left(\frac{z_{j_k+1}-z_{j_k-1}}{2}\right)} = \frac{\left(\exp(Q(\tau_k-\tau_{k-1}))\right)_{j_{k-1},j_k}}{\left(\frac{z_{j_k+1}-z_{j_k-1}}{2}\right)} \tag{5.2.7}$$

where the denominator is necessary since Y_τ has a continuous state space and Z_τ has a discrete state space. We can rewrite the likelihood described in equation 5.2.5 by replacing the original process X_t by the one step process Z_τ. We then have:

$$\begin{aligned}\ell(\theta) &= p(i_1,\ldots,i_m|\theta,\mathcal{T}) = \\ &\sum_{j_m=0}^{H-1} p(i_m|z_{j_m}) \sum_{j_{m-1}=0}^{H-1} p_{\tau_m-\tau_{m-1}}(z_{j_{m-1}},z_{j_m})\cdots \\ &p(i_2|z_{j_2}) \sum_{j_1=0}^{H-1} p_{\tau_2-\tau_1}(z_{j_1},z_{j_2})\cdot \\ &p(i_1|z_{j_1}) p_{\tau_1-\tau_0}(\frac{1}{2N_e},z_{j_1}).\end{aligned} \tag{5.2.8}$$

where $p(i_k|z_{j_k}) = \left(\begin{array}{c} n_k \\ i_k \end{array}\right) z_{j_k}^{i_k}(1-z_{j_k})^{n_k-i_k}$ from equation 5.2.3.

In the case of experimental evolution this unconditional process should be realistic since in principle one might want to estimate the selection coefficient for any locus. We will now consider one special case of what is presented above, motivated by ancient DNA data. We will assume that the allele is segregating at the last sampling time (i.e., the process is never reaching the states 0 or 1). This case corresponds to what we think is a realistic scenario for how ancient DNA data would be collected, where presumably the locus of interest is polymorphic at present. Indeed, only such loci would be selected for inference.

We can rewrite the likelihood as follows:

$$\ell^C(\theta) = p(i_1,\ldots,i_m|\theta,\mathcal{T},z_{j_m}\notin\{0,1\}) \;=\; \frac{p(i_1,\ldots,i_m,z_{j_m}\notin\{0,1\}|\theta,\mathcal{T})}{\sum_{j_m=1}^{H-2} p_{\tau_m-\tau_0}(\frac{1}{2N_e},z_{j_m})} \tag{5.2.9}$$

where

$$\begin{aligned}&p(i_1,\ldots,i_m,z_{j_m}\notin\{0,1\}|\theta,\mathcal{T}) = \\ &\sum_{j_m=1}^{H-2} p(i_m|z_{j_m}) \sum_{j_{m-1}=1}^{H-2} p_{\tau_m-\tau_{m-1}}(z_{j_{m-1}},z_{j_m})\cdots \\ &p(i_2|z_{j_2}) \sum_{j_1=0}^{H-2} p_{\tau_2-\tau_1}(z_{j_1},z_{j_2})\cdot \\ &p(i_1|z_{j_1}) p_{\tau_1-\tau_0}(\frac{1}{2N_e},z_{j_1}).\end{aligned} \tag{5.2.10}$$

We can consider the subprocess Z^C_τ defined on the reduced state space $\{z_1, \ldots, z_{H-2}\} \subset \{z_0, z_1 \ldots z_{H-2}, z_{H-1}\}$. The infinitesimal generator q^C of such a process is the matrix Q without the first and last rows and columns, i.e.,:

$$q^C = \begin{pmatrix} \eta_1 & \beta_1 & 0 & & \ldots & & 0 \\ \delta_2 & \eta_2 & \beta_2 & 0 & & & \\ 0 & \ddots & \ddots & \ddots & 0 & & \vdots \\ & 0 & \delta_k & \eta_k & \beta_k & 0 & \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\ & & & 0 & \delta_{H-3} & \eta_{H-3} & \beta_{H-3} \\ 0 & & \ldots & & 0 & \delta_{H-2} & \eta_{H-2} \end{pmatrix}. \tag{5.2.11}$$

Denoting $p^C_{\tau_k - \tau_{k-1}}(z_{j_{k-1}}, z_{j_k})$ the transition probabilities of this subprocess we have that $p_{\tau_k - \tau_{k-1}}(z_{j_{k-1}}, z_{j_k}) = p^C_{\tau_k - \tau_{k-1}}(z_{j_{k-1}}, z_{j_k})$ for $\forall\, j_{k-1}, j_k \notin \{0, H-1\}$ (see appendix B for more details).

Finally in order to compute the likelihood of equations 5.2.8 and 5.2.9, the only difficulty is to compute the matrix exponentiation $e^{Q\tau}$ and $e^{q^C\tau}$, respectively.

5.2.2 Numerics

We evaluate numerically the matrix exponentiation. The advantage of the current approach compared to Bollback et al.'s is that we do not need to do a numerical integration step since the state space is already finite. The description of the matrix exponentiation is given in appendix B.

Although asymptotically the one step process is equivalent to the WF model, since the state space of Z_τ has a finite number of states, the accuracy of the approximation will depend on the choice of the states, or what we call from now on "the grid". We investigate three grids strongly inspired by Gutenkunst et al. (2009). The first one is a uniform grid with a point added at $\frac{1}{2N_e}$. The second and third grid are a "quadratic grid" and an "exponential grid". The last two grids were chosen to be refined around the boundaries such that the distance between adjacent points changes smoothly. The details for the grids are given in Appendix B. All three grids have a point at $\frac{1}{2N_e}$.

The likelihood function is complex. We were not able to compute the maximum of the function analytically. Therefore, in order to find the maximum we first computed the likelihood over a large range of parameters. We verified that there is a single maximum for each time interval defined by adjacent sampling times, i.e., if $t_0 < t_1$, the time intervals are $(-\infty, t_0)$, (t_1, t_2),...,(t_{m-1}, t_m), and that the likelihood surface is smooth. We used the *SciPy* (Jones et al., 2001a) implementation of the Nelder-Mead simplex algorithm (Nelder and Mead 1965) to find the maximum for each time interval.

Our implementation is written in *Python* and *C++* making use of *Numpy* (Oliphant, 2006), *SciPy* and *mpack* (Nakata, 2010) libraries for computations and of the *Matplotlib* library (Hunter, 2007) for plotting.

5.2.3 Simulations

In order to test our model, we simulate several datasets with the WF model forward in time. Simulating with the WF model can be time consuming if the population size is large, so we picked a small population size ($N_e = 500$). But in principle the conclusions hold for higher population size. We then infer the maximum likelihood estimates (MLEs) using our one step method. We use two different sampling schemes. The first one is similar to the real dataset we analyze below, i.e., 6 sampling times each with 50 chromosomes. And a second one corresponding to having twice as many sampling times with half the number of chromosomes, i.e., 12 sampling times and 25 chromosomes. We searched for the MLEs across a finite domain, i.e., $N_e \in [100, 1000]$, $t_0 \in [-3000, 0]$, and $\gamma \in [-200, 200]$. We can finally assess the accuracy of our estimator and compare the sampling schemes by looking at the bias of the estimates and the root mean square error (RMSE).

5.2.4 Real data

In 2009, Ludwig et al. (2009) sequenced several loci encoding coat color in horses. Each locus had been shown to be linked with a color phenotype in present day horses. In other words, the phenotype associated with each locus is segregating in present populations. We re-analyze in this work one of the loci encoding for the agouti-signaling-protein (ASIP), that controls the distribution of the black pigment (Rieder et al. (2001)). The hypothesis is that at the beginning of domestication, some coat colors in horse were positively selected for.

The samples sequenced were obtained from Siberia, Middle and Eastern Europe, China and the Iberian Peninsula. As in Ludwig et al. (2009) we grouped the samples into six sampling times, $t_1 \simeq -20000$, $t_2 \simeq -13100$, $t_3 \simeq -37000$, $t_4 \simeq -2800$, $t_5 \simeq -1100$ and $t_6 \simeq -500$ where the unit is years BC. We assumed that the generation time of horses is 5 years, following Ludwig et al. (2009). The wild type horses are presumed to have been of bay color. The mutation of interest is recessive, since only horses homozygous for the *ASIP* locus will be black. So, in this case $h = 0$.

To compute a possible range for the population sizes we use data from Cieslak et al. (2010). They sequenced part of the control region of the mtDNA for 78 samples that are part of Ludwig et al. (2009)'s dataset. The control region of the mtDNA is a non coding region. One way to compute the population size N_e is to compute the diversity π of the samples. Then, assuming the region is neutral and ignoring hitchhiking effects due to nearby selected sites, we use the relationship that relates the diversity of a sample to the population size, $\pi = 2N_e\mu \Rightarrow N_e = \frac{\pi}{2\mu}$, where μ is the mutation rate per base pair per generation. To get an estimate of the mean and standard error of π of the mtDNA sample, we use the maximum likelihood method implemented in *MEGA* (Tamura K et al., 2011) with default parameters. The standard error for the diversity was computed performing 1000 bootstraps. We use Jazin et al. (1998)'s estimate for the mutation rate (i.e., $\mu \in (3.0 \cdot 10^{-6}, 4.4 \cdot 10^{-5})$). Those authors used human families to get direct estimates of the mutation rate for mtDNA control region for a single generation. Although the mutation rate is an important parameter, we do not have direct estimate in horses and we have to rely on results for other species . To get conservative upper lower bounds for N_e we use the 95% confidence interval (CI) bounds of the mutation rate and the diversity. If the CIs for μ and π are denoted (μ_{low}, μ_{up}) and

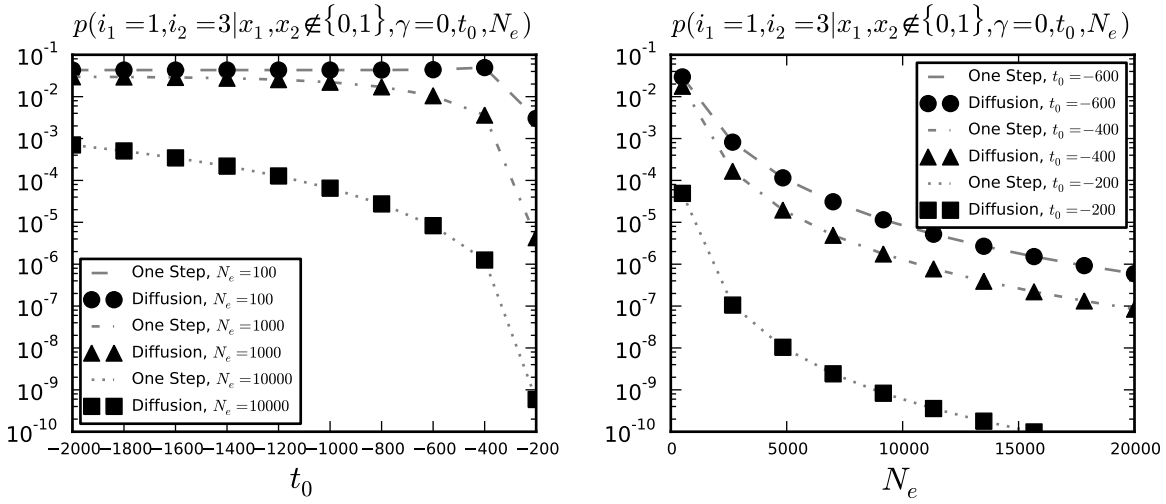

Figure 5.2: Likelihood for the neutral case for several values of N_e and t_0. The likelihood is for two samples taken at times -200 and 0 generations of size $M = (4, 4)$ and with $I = (1, 3)$ derived alleles. On the left (right), we fix N_e (respectively t_0) to several values and plot the likelihood versus t_0 (respectively N_e).

(π_{low}, π_{up}) respectively, we defined $N_{e_{low}} = \frac{\pi_{low}}{2\mu_{up}}$ and $N_{e_{up}} = \frac{\pi_{up}}{2\mu_{low}}$.

In order to find the MLEs we use a domain defined by $N_e \in [200, 5000]$, $t_0 \in [-10000, 0]$, and $\gamma \in [-200, 200]$ for the parameters. We fix $H = 400$ for this computation.

For the CIs, there exist several asymptotic results that apply for maximum likelihood, especially for a time serial Markov chain. But here, the sample sizes are generally small, therefore we chose to compute the CIs with a parametric bootstrap approach.

Note that several assumptions of our model are violated with this dataset, such as constant population size, potentially random mating (since the samples are taken from all around the world), but also, the *MC1R* locus, encoding a melanocortin receptor and related to the black pigment production, is known to have an epistatic interaction with *ASIP* (Rieder et al.). Nevertheless we decided to analyze these data to be able to compare those with the results obtained with Bollback et al.'s method on the same dataset.

5.3 Results and Discussion

5.3.1 Numerics

In order to validate the method we compared several known analytical results for the WF model with the one step process. For the neutral case, it is possible to compute the likelihood since the transition probabilities are known for the diffusion process (see e.g. Ewens (2004)). We plot the results in Figure 5.2 for a quadratic grid of size 100 for two samples of size $M = (4, 4)$ and number of A alleles $I = (1, 3)$, sampled at times $T = (-200, 0)$ for several values of N_e and t_0. The plots suggest that even for a grid of size 100 the one step process is a very good approximation of the diffusion process.

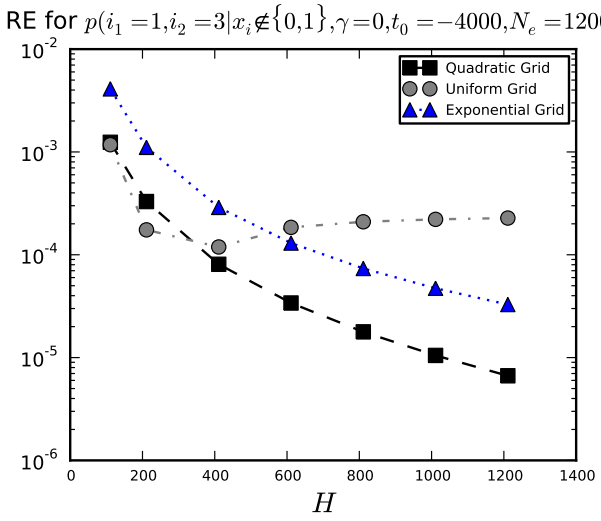

Figure 5.3: Relative error (RE) for the three grids discussed in 5.3.1 for the likelihood of 2 samples taken at times -3000 and 0, with $M = (4, 4)$. The parameter H describes the size of the grid. The y-axis is in logarithmic scale. In this example, the one step process converges towards the diffusion process faster when using the quadratic grid rather than the other two grids.

We then compare the relative error between the diffusion and the one step process and demonstrate that, when we increase the grid size the one step process converges towards the diffusion process. The results for a particular choice of parameters is shown in Figure 5.3 for the three grids discussed in 5.3.1. First we note that the one step process does converge as expected with increasing grid size. In this example, the convergence is faster for the quadratic grid. We looked at several combinations of parameters, and we observe that the quadratic grid and the exponential grid perform better than the uniform grid in general but that the ordering between the other two grids depends on the parameters. Indeed, if the allele age is close to the first sampling time a grid more refined around the frequency $\frac{1}{2N_e}$ performs better. In the applications below we will use a quadratic grid of size between 100 and 400.

5.3.2 Simulations

We picked a population size of $N_e = 500$ and set the allele age to $t_0 = -1400$. We fix the selection coefficient to seven potential values: $\gamma \in \{-10, -5, 0, 5, 10, 15, 20\}$.

First, we fix the sampling times to $T = (-1000, -800, -600, -400, -200, 0)$ generations and sample 50 chromosomes at each time point. Then we look at a scheme where the samples are taken every 100 generations from -1100 up to 0 (i.e. 12 samples). At each sampling time we sample 25 chromosomes. The intent is to quantify whether it is better to sample more chromosomes at fewer time points, or the opposite.

The boxplot results for these simulations are shown on Figure 5.4. They are standard

boxplots showing the five point summary (the minimum, the first quartile, the median, the third quartile, and the maximum). Then we plot the bias and the RMSE on Figure 5.5 for both schemes.

For the population size, the MLEs span all the potential range of N_e values, but the bulk of the results exclude very low population sizes. This suggests nevertheless that it is hard to estimate N_e with our method, at least with a precision higher than one order of magnitude. Our estimator is biased upwards for both schemes but this might be explained by the presence of outliers since the median is largely accurate. Moreover, the second scheme, with less chromosomes and more sampling schemes leads to a smaller bias and a smaller RMSE for most cases.

In contrast, the results for the selection coefficient are essentially unbiased, with a symmetric distribution, and the median matching the mean of the distribution. The variance remains large and only when γ is quite high can one reject neutrality. In particular, the higher the selection coefficient, the higher the variance. The RMSE this time is worse for the second sampling scheme.

The results for the allele age also exhibit a large variance. The tail of the distribution is large. This can be explained by the use of the conditional process. Indeed for weak selection, if the number of derived alleles is high at the first sampling time the likelihood becomes uninformative for the allele age (i.e., the likelihood is flat for older allele ages; Figure 5.2). This leads to difficulty for the optimization algorithm to converge to the global maximum. The results seem to be systematically biased upwards, although the median is accurate. For strong selection the likelihood is more informative and the estimator is unbiased. Also, for strong selection the scheme with more samples through time performs considerably better.

In conclusion, especially for strong selection, sampling fewer chromosomes over more sampling times will lead to better results.

5.3.3 Real data

The change in allelic frequency of this locus is shown in Figure 5.6. Although the frequency is increasing in around 3,000 generations from 0 to ∼0.8 between the first and the third sampling time, suggesting positive selection, it then drops down to 0.4 in around 500 generations. It is interesting to note that the archaeological evidence for domestication suggests a date of 3500 years BC (Outram et al., 2009), which would correspond to the third sampling time (i.e. when the sample frequencies start decreasing).

The first step is to choose a potential range for the population size. We found $\pi = 0.024$ with a 95% CI of $(0.018, 0.030)$. Together with the 95% CI of the mutation rate, this leads to a range for N_e of $(200, 5000)$. This is a small population size. It might be explained by the fact the horses are a domesticated species and most samples are taken after the beginning of domestication, resulting in a small N_e. On the other hand it might be that the mutation rate calculated for the human population for the control region is not appropriate for horses.

We first plot the likelihood surface for 4 values of N_e on Figure 5.7. This will help us confirm that we have found a global maximum. We note that the higher the population size the higher the γ and the older the allele age that maximizes the likelihood. For example if the population is fixed at $N_e = 200$ then $\gamma^{max} = -1.5$ and $t_0^{max} = -2567$. In contrast, if we fix $N_e = 5000$, then $\gamma^{max} = 9.1$ and $t_0^{max} = -3550$. In other words, if the mutation rate is

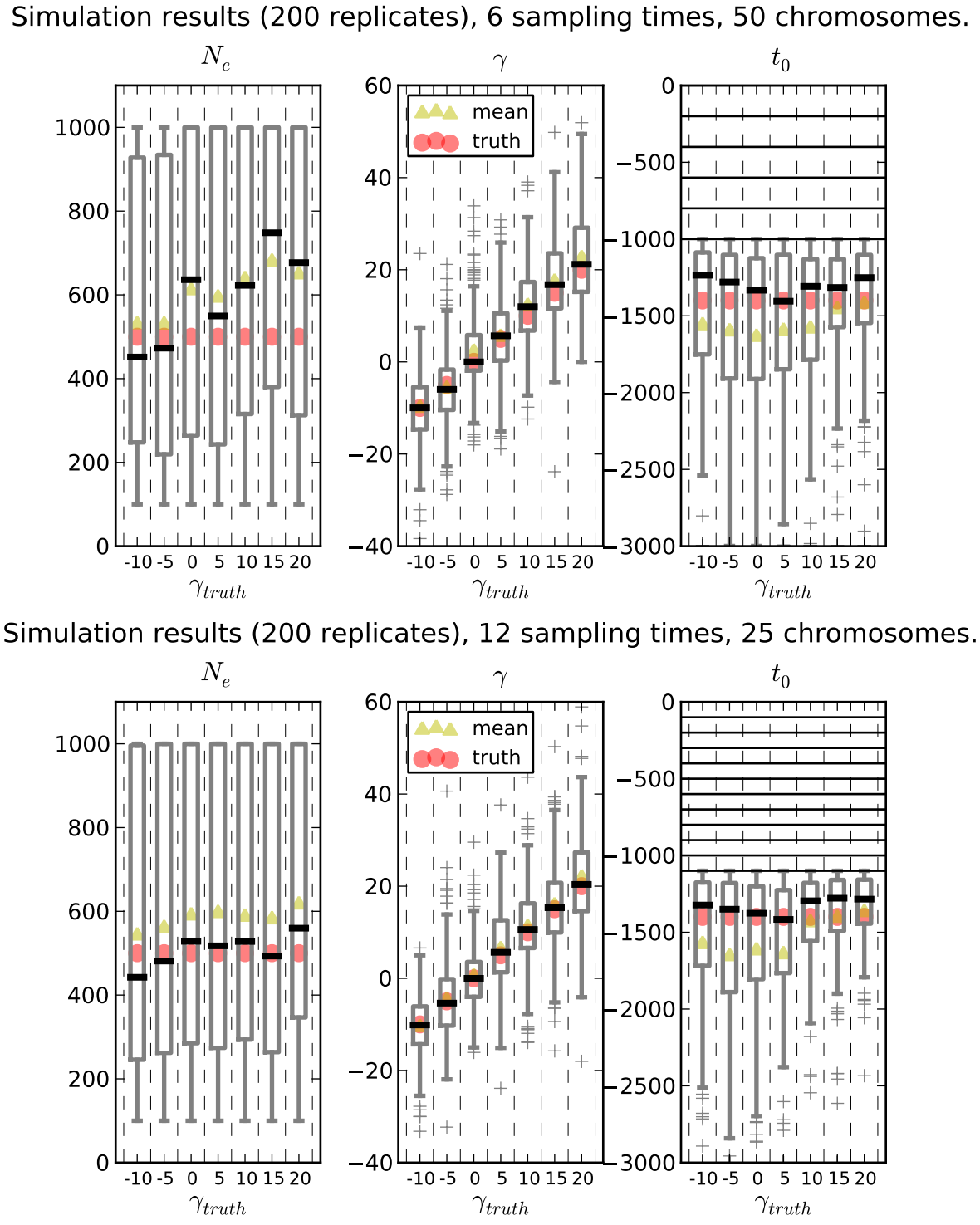

Figure 5.4: Boxplots for the MLEs of each simulation replicate, for seven different parameter combinations. At the top is the scheme with 6 sampling times and 50 chromosomes sampled. At the bottom, the scheme with 12 sampling times and 25 chromosomes sampled. On each plot, left are the estimates for the population size, N_e, in the middle for the rescaled selection coefficient, γ, and right for the allele age, t_0. For all subplots the triangle represents the mean of the estimates, and the circle the true value. The rectangles of the boxplots are for the first and third quartile and the black line represents the median. The outliers are also indicated by crosses.

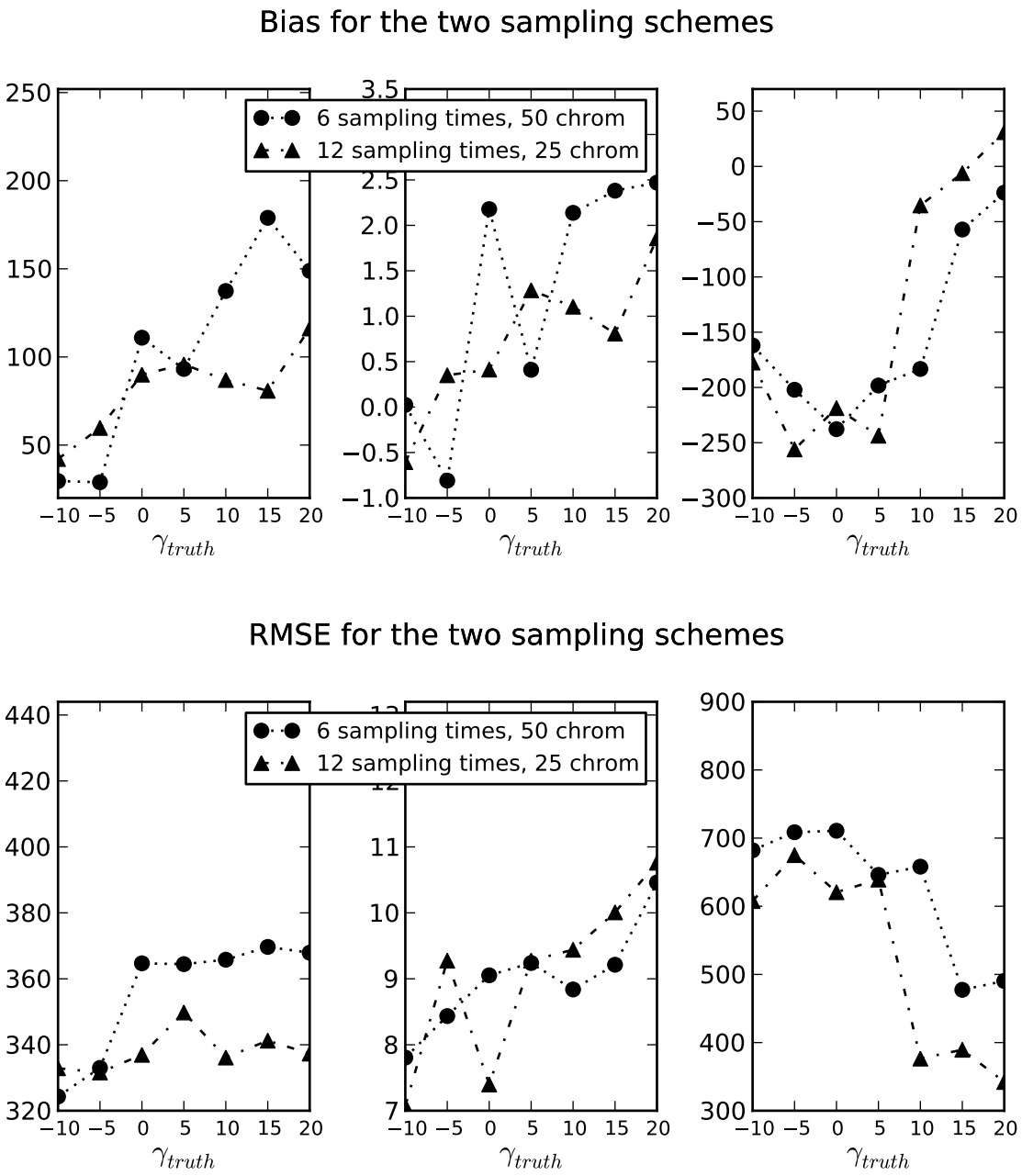

Figure 5.5: Bias (top plot) and RMSE (bottom plot) results for the MLEs for seven different sets of simulations also presented in Figure 5.4

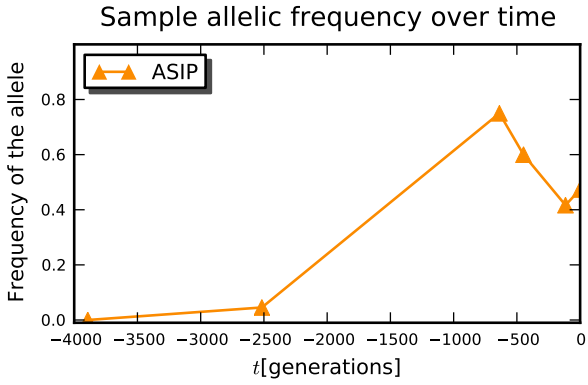

Figure 5.6: Change in allelic frequency over time for the ASIP locus. The sample sizes are $M = (10, 22, 20, 20, 36, 38)$ and the number of derived alleles $I = (0, 1, 15, 12, 15, 18)$. The times have been offset so that the last sampling time is 0. Domestication is thought to have happened around -3500 years BC which would correspond to around -600 generations on this plot, i.e. the 3rd sampling time.

overestimated by say an order of magnitude, our potential range for the population size will also be much higher, affecting the results.

Since there is no mutant allele at the first time of sampling, the allele might have arisen after the first sampling time. We denote *"dom1"* the range between $(-\infty, -3893]$ generations, and *"dom2"* the range $(-3893, -2516]$. As discussed before, the likelihood is therefore discontinuous as a function of the allele age with discontinuities at sampling times. It is important to look for the global maximum in *dom1* and *dom2* separately. Moreover, we compute the 95% CI in *dom1* and in *dom2* separately. We build the confidence interval as a union of (potentially) disconnected domains.

The values for the MLEs and 95% CI are shown in Table 5.1. The first thing to note is that they are compatible with the results of Figure 5.7. The MLEs were found in dom2: $t_0^{mle} \cong -2577$ $\gamma^{mle} \cong -1.3$ and $N_e^{mle} = 652$.

In Figure 5.8 we plot the distribution for the bootstrap replicates for each parameter and for the maximum likelihood values. The confidence interval was constructed as the 2.5th and 97.5th percentile. We ran a total of 1400 replicates. For about 30 of those simulations, the optimizer did not converge. Among successful runs, ∼500 did not have an MLE in *dom1* or *dom2* and were discarded. From the remaining, about 823 were found in *dom2* and 34 in *dom1*.

The MLEs and the bootstrap results have several implications. First, we do not find evidence for positive selection as could be anticipated by the archaeological evidence for domestication. The discrepancy between this study and Ludwig et al. (2009) is first the method used and second the parameter range assumed. Indeed, the results in Ludwig et al. (2009) were obtained using Bollback et al. (2008)'s method. Since our t_0^{mle} is in *dom2*, and Bollback et al. 2008 assume that the allele was already present in the first time of sampling, it is to be expected that our results will be very different. Moreover, the potential range for the population size in Ludwig et al. 2009 is from 10,000 to 100,000, i.e., it does not overlap with the range for N_e that we assume here. As noted above, if we had assume a larger

	dom1	*dom2*
	optimum	optimum
ℓ	14.9	**13.1**
t_0	-3893	**-2577**
γ	-0.61	**-1.3**
N_e	1617	**652**

	95% CI
t_0	$(-4759, -3893] \bigcup (-3892, -2516]$
γ	(-27.7,60.7)
N_e	(200,5000)

Table 5.1: Maximas and CI for the ASIP locus sequenced in (Ludwig et al., 2009). The MLEs are on the right most column of the first table. The CIs were obtained through parametric bootstrap, see Figure 5.8

population size, the γ^{mle} would be larger.

The distribution of each parameter from the bootstrap replicates are almost unbiased relative to the true value (as could be expected from the results in the simulation section). The distribution for γ is close to a normal distribution while the distribution for N_e and t_0 are not as simple. For N_e, the distribution is bimodal with a second mode at the upper bound. This mode is a reflection of the finite domain we impose on the search for the MLE rather than an actual mode. Similarly, for t_0 there is a mode at the lower bound for *dom2*, an artifact of the bounds from the sampling times.

As could be expected from the simulations above, the 95% CI for N_e suggests that with these data we have no power to estimate N_e. Similarly, we have no power to distinguish between negative and positive selection as γ's CI is between -27.7 and 60.7. On the other hand, the bootstrap replicates suggest that the allele arose in *dom2*. We can indeed test the hypothesis that the allele age is not in *dom2*, $H_0 : t_0 \notin$ *dom2* versus the hypothesis that the allele age is in *dom2*, $H_1 : t_0 \in$ *dom2*. We can reject the nul hypothesis H_0 with pvalue $1 - \frac{823}{823+34} = 0.04$. The domain *dom2* corresponds to -20,000 to -13,100 years BC. In other words, from the data, one could have already deduced that the allele had to be present before -13,100 years (i.e., before the presumed start of domestication). Indeed, domestication in horses is thought to have started about 3,500 years BC (Outram et al., 2009). Our analysis shows that it is likely to have arisen within the last 20,000 years.

5.4 Conclusion

The allele age, the strength of selection and the population size are all crucial parameters in population genetics. Although molecular data is growing exponentially in recent years, it often remains a challenge to estimate those key parameters.

We develop a maximum likelihood approach to estimate these parameters that deals with a particular type of data - temporal data. Our method is based on an approximation to the WF diffusion process, and has the advantage of being quite flexible and appropriate for hypothesis testing. Moreover, it is fast for small γ, as one evaluation of the likelihood

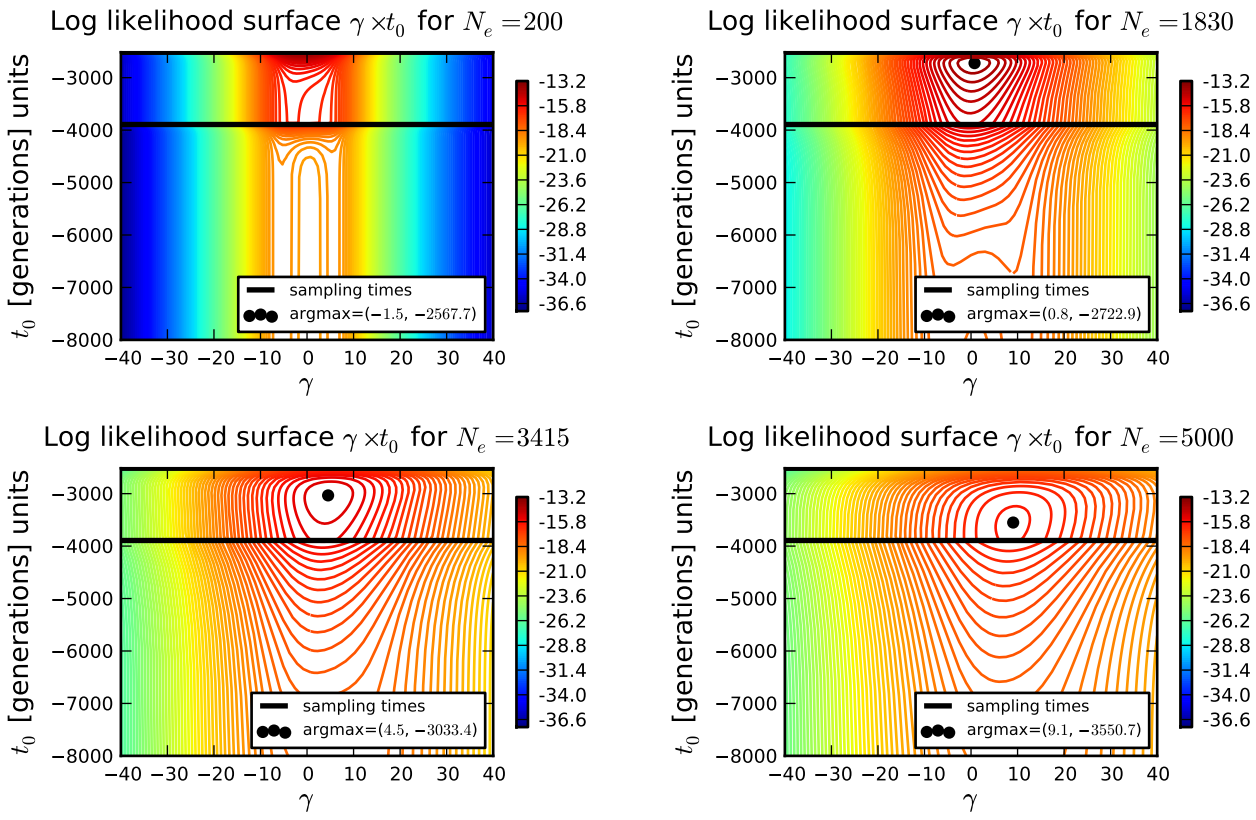

Figure 5.7: Likelihood surfaces for various values of N_e when analyzing the *ASIP* locus. In each case the local maximum is indicated.

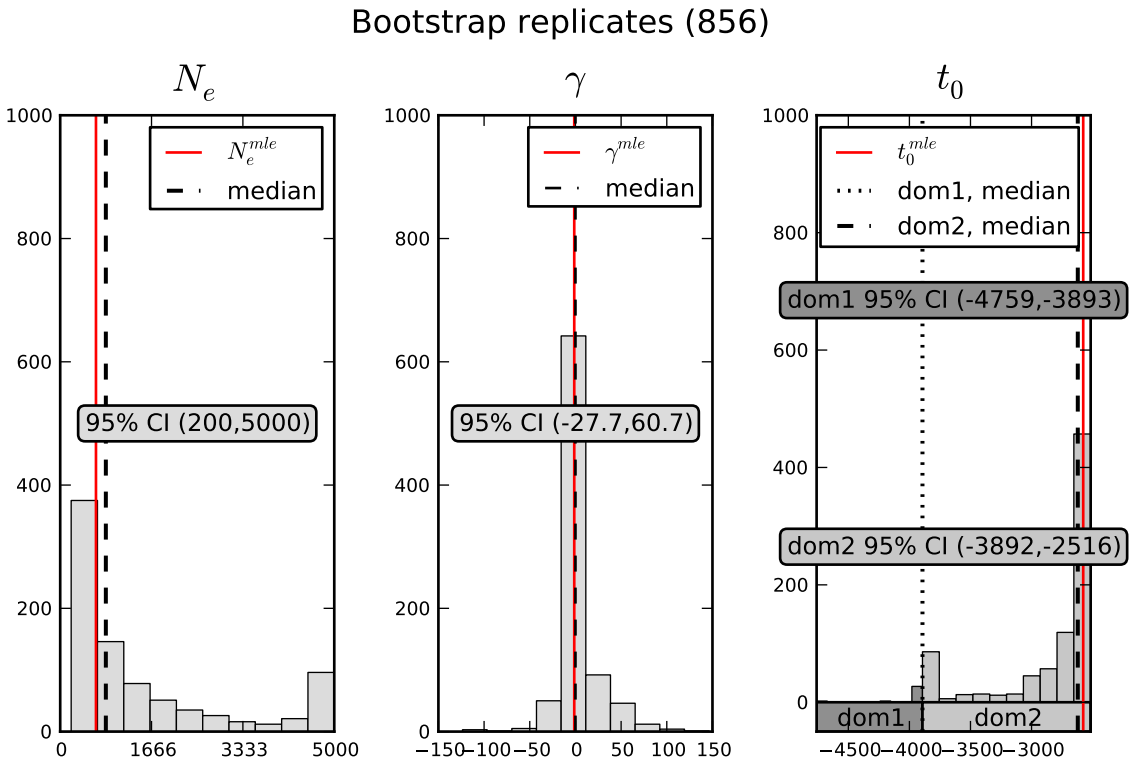

Figure 5.8: Bootstrap estimate of the sampling distribution of ML estimators of the three variables N_e, γ and t_0 for the parametric bootstrap. Of 1400 simulations, 856 were compatible with the data (i.e. the maximum for t_0 was in *dom1* or *dom2*. In each case the local maximum is indicated.

function takes ∼0.1 seconds for $\gamma \lesssim 40$ on a laptop with a i5 2.53 GHz CPU, for a dataset like the one we analyze here.

We show through simulations that for a realistic sample of realistic size, although the variance of our estimator is quite large, our MLE is unbiased for estimating selection and is nearly unbiased for the age of the allele and the effective population size. On the other hand, our method is not appropriate for estimating the population size, even for simulations where the model used to simulate the data match the method used to infer the parameters. Indeed, for a realistic sampling scenario, the MLEs for N_e, although unbiased, can span several orders of magnitude. This is not surprising. The effective population size is a parameter notoriously difficult to estimate, and our method considers only a single locus.

The sampling scheme has of course an impact on the accuracy of the estimator. We investigate two different sampling strategies and conclude that, in the cases considered, it is better to increase the number of sampling times rather than the number of samples per time point. It is indeed intuitive that in order to be able to estimate the allele age, for the conditional process, it is necessary to have a sample close to the allele age. Indeed, in the conditional process, an allele will never get fixed or lost. Thus, after several coalescent units, the likelihood is flat.

We re-analyze a locus that was previously found to be under positive selection, *ASIP,* by evaluating samples ranging from the Pleistocene to present. In this study, we do not have enough power to distinguish positive from negative selection for this locus. This could also be due to an underestimate of the effective population size, or a violation of one or more assumptions of our null model, as discussed earlier. Although we are not able to estimate the selection coefficient precisely, we find the age of the ASIP mutation to range between -20000 to -13100 with an MLE at -13400 years BC, which well predates domestication.

Even though we analyze a mammalian dataset, our method can in principle be applied to datasets obtained in experimental evolution or viral data. But, it is important to note that our approximation to the WF model will only be valid provided that $s \sim \mathcal{O}\left(\frac{1}{N_e}\right)$ as for the WF diffusion process.

It is clear that several violations to the assumptions of the model can lead to increased bias of our estimator. In particular, we assume a constant, panmictic population size. If the population is fluctuating in size, this can lead to biased estimates of selection coefficients and allele age.

Finally, our framework could be extended to include multiple loci at a time, using several loci simultaneously to estimate N_e while inferring selection at the target locus.

Appendix A

One step process, Q matrix

We denote by L the generator of the diffusion process Y_τ. We have that

$$L = \frac{1}{2}a(y)\frac{d^2}{dy^2} + b(y)\frac{d}{dy} \tag{A.0.1}$$

where $a(y)$ and $b(y)$ are the infinitesimal variance and mean of our diffusion process. For the WF model with additive selection (see main text) those functions are:

$$a(y) = y(1-y) \tag{A.0.2}$$
$$b(y) = \gamma y(1-y)(y + h(1-2y)) \tag{A.0.3}$$

By definition the generator can also be written as

$$\lim_{\tau \downarrow 0} \frac{\mathbb{E}^y[f(Y_\tau)] - f(y)}{\tau} = Lf(y) \tag{A.0.4}$$

Ignoring the $\Delta\tau^2$ terms, we have for the infinitesimal mean:

$$\mathbb{E}^y[Y_{s+\Delta\tau} - Y_s \mid Y_s] \cong \gamma Y_s(1-Y_s)(Y_s + h \cdot (1-2Y_s))\Delta\tau = b(Y_s) \cdot \Delta\tau \tag{A.0.5}$$

Similarly the infinitesimal variance is:

$$\mathbb{E}^y\left[\{Y_{s+\Delta t} - Y_s - \gamma Y_s(1-Y_s)(Y_s + h \cdot (1-2Y_s))\}^2 \mid Y_s\right] \cong Y_s(1-Y_s)\Delta\tau = a(Y_s) \cdot \Delta\tau \tag{A.0.6}$$

We want to choose the Markov chain Z such that $Z \simeq Y$, in the sense that the probability distribution governing the samples of Z is close to the probability distribution governing the samples of Y. To achieve that, we can match the infinitesimal mean and variance of Z and Y (see Durrett (2008)). By definition of the generator of Z_τ (see equation 5.2.6), we know the probabilities of transition in time $\Delta\tau$. Assuming the process starts at $Z_s = z_i$:

$$Z_{s+\Delta\tau} = \begin{cases} z_i & \text{with probability } 1 - (\beta_i + \delta_i)\Delta\tau + \mathcal{O}(\Delta\tau^2) \\ z_{i+1} & \text{with probability } \beta_i\Delta\tau + \mathcal{O}(\Delta\tau^2) \\ z_{i-1} & \text{with probability } \delta_i\Delta\tau + \mathcal{O}(\Delta\tau^2) \end{cases} \tag{A.0.7}$$

We can rewrite equations A.0.5 and A.0.6 replacing Y_τ by Z_τ. We have for the infinitesimal mean

$$\begin{aligned}\mathbb{E}^{z_i}\left[\{Z_{s+\Delta t}-z_i\}\right] &\cong z_i\cdot(1-(\beta_i+\delta_i))+z_{i+1}(\beta_i\Delta\tau)+z_{i-1}(\delta_i\Delta\tau)-z_i\\ &= (\beta_i(z_{i+1}-z_i)+\delta_i(z_i-z_{i-1}))\Delta\tau\\ &= b(z_i)\cdot\Delta\tau. \end{aligned} \tag{A.0.8}$$

And for the infinitesimal variance:

$$\begin{aligned}\mathrm{Var}(Z_{s+\Delta t}-z_i) &= \mathbb{E}^{z_i}\left[\{Z_{\Delta t}-z_i\}^2\right]-\mathbb{E}^{z_i}\left[\{Z_{\Delta t}-z_i\}\right]^2\\ &\cong (z_{i+1}-z_i)^2\cdot\beta_i\Delta\tau+(z_{i-1}-z_i)^2\cdot(\delta\Delta\tau-(z_i-z_i)^2(1-\beta_i-\delta_i)\Delta\tau\\ &= (\beta_i(z_{i+1}-z_i)^2)+(\delta_i(z_i-z_{i-1})^2)\Delta\tau\\ &= a(z_i)\cdot\Delta\tau. \end{aligned} \tag{A.0.9}$$

We have therefore two equations A.0.8 and A.0.9 with two unknowns δ_i and β_i. Solving the system we have:

$$\beta_i=\frac{(-1+z_i)\cdot z_i\cdot(-1-z_i^2\cdot\gamma+h\cdot(-1+2\cdot z_i)\cdot(z_i-z_{i-1})\cdot\gamma+z_i\cdot z_{i-1}\cdot\gamma)}{(z_i-z_{i+1})\cdot(z_{i-1}-z_{i+1})} \tag{A.0.10}$$

$$\delta_i=\frac{-((-1+z_i)\cdot z_i\cdot(-1-z_i^2\cdot\gamma+h\cdot(-1+2\cdot z_i)\cdot(z_i-z_{i+1})\cdot\gamma+z_i\cdot z_{i+1}\cdot\gamma))}{(z_i-z_{i-1})\cdot(z_{i-1}-z_{i+1})} \tag{A.0.11}$$

Note that since we require that $\delta_i,\beta_i>0\ \forall i$, the range of the possible parameters γ depends on the choice of the states z_{i-1}, z_i, z_{i+1}, or on the grid. In particular if we use a uniform grid we get: $\{z_0,z_1,...,z_{H-1}\}=\{0,\frac{1}{H-1},...,\frac{H-2}{H-1},1\}$ and $\beta_i=\frac{(-1+H-k)k(1+H^2+k\gamma+H(-2+h\gamma)-h(\gamma+2k\gamma))}{2(-1+H)^2}$ and $\delta_i=\frac{(-1+H-k)k(1+H^2-k\gamma-H(2+h\gamma)+h(\gamma+2k\gamma))}{2(-1+H)^2}$. Most likely the locus of interest is either dominant, co-dominant of recessive, i.e. $h\in\{0,\frac{1}{2},1\}$. In those three cases for a uniform grid the range of γ is easy to compute. If $h=\frac{1}{2}$ then $-2(H-1)<\gamma<2(H-1)$, if $h=0$, $-(H-1)<\gamma<(H-1)$, and if $h=1$, $-\frac{(H-1)^2}{H-2}<\gamma<\frac{(H-1)^2}{H-2}$. In other words, we will need a large grid for high values of γ, slowing our computation.

Appendix B

Numerics

B.1 Matrix exponentiation

We would like to compute the matrix exponential of the matrix Q and the matrix q^C for the conditional process. We will focus on the non conditional process as the conditional process follows easily. We use the convention of numbering the elements of a matrix starting from 0 to $H-1$ for the unconditional process, and from 1 to $H-2$ for the conditional process. We seek to compute

$$\exp(Qt),$$

where the $H \times H$ matrix Q is a tridiagonal matrix with all entries above and below the diagonal strictly positive. We implement two different approaches to compute the matrix exponentiation.
The first approach is a scaling and squaring algorithm with a Padé approximation. This approach is described in detail in Moler and Van Loan (2003) and is implemented in *SciPy*. This method works for a general matrix and takes advantage of the properties of the matrix Q.
The matrix Q is in general not symmetric ($\delta_i \neq \beta_i$ when $s \neq 0$). Nevertheless all eigenvalues are real. In particular two eigenvalues are 0 and the others are negative. Thus, when we remove the first and last column and row, the resulting matrix is the tridiagonal matrix q^C. We can transform the matrix q^C into a symmetric matrix with a similarity transformation. More precisely, there exists a diagonal matrix

$$d = \begin{pmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ ... & ... & ... & ... \\ 0 & 0 & d_{H-3} & 0 \\ 0 & 0 & 0 & d_{H-2} \end{pmatrix} \tag{B.1.1}$$

such that $s = d^{-1}q^C d$ is a symmetric matrix. The d_i can be defined recursively as follows $d_1 = 1$,$d_2 = \sqrt{\delta_2/\beta_1} \cdot d_1$, $d_3 = \sqrt{\delta_3/\beta_2} \cdot d_2$, ... Note that the square root exists since $\beta_i, \delta_i > 0$. The matrices q^C and s have the same eigenvalues, and the eigenvalues of a symmetric matrix are all real. In particular they are also eigenvalues of the original matrix Q. The two remaining eigenvalues of Q are the two zero eigenvalues (this can be seen writing

the characteristic polynomials). Therefore all eigenvalues are real. We can build a matrix D adding a first and last row and column to the matrix d:

$$D = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & d_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & d_2 & 0 & 0 & 0 \\ 0 & ... & ... & ... & ... & 0 \\ 0 & 0 & 0 & d_{H-3} & 0 & 0 \\ 0 & 0 & 0 & 0 & d_{H-2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{B.1.2}$$

It follows that $R = D^{-1}QD$ symmetries the interior part of Q (the matrix q^C) and is a tridiagonal matrix as well. Since $s = d^{-1}qd$ is symmetric there exists an orthogonal matrix, o, such that $\ell = o^T so$ is diagonal. This matrix ℓ has the following form:

$$\ell = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ ... & ... & ... & ... \\ 0 & 0 & \lambda_{H-3} & 0 \\ 0 & 0 & 0 & \lambda_{H-2} \end{pmatrix} \tag{B.1.3}$$

We can construct the matrix O as the matrix D before, with o in its center and adding first and last rows and columns with zeros everywhere but the diagonal entries $(0,0)$ and $(H-1,H-1)$. Then we see that $T = O^T RO$ has an inner part equal to ℓ the coefficients of the first and last lines remain equal to 0, and the coefficients on the first and last columns are non-zero. We denote $T(0,j) = v_{0,j}$ with $j = 1,\ldots,H-2$ and $T(H-1,j) = v_{H-1,j}$ with $j = 1,\ldots,H-2$. That is

$$T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ v_{0,1} & \lambda_1 & 0 & 0 & 0 & v_{H-1,1} \\ v_{0,2} & 0 & \lambda_2 & 0 & 0 & v_{H-1,2} \\ v_{0,..} & ... & ... & ... & ... & v_{H-1,..} \\ v_{0,H-3} & 0 & 0 & \lambda_{H-3} & 0 & v_{H-1,H-3} \\ v_{0,H-2} & 0 & 0 & 0 & \lambda_{H-2} & v_{H-1,H-2} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{B.1.4}$$

where the $v_{i,j} \neq 0$. We can rewrite $T = \Lambda + V$ where

$$\Lambda = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_2 & 0 & 0 & 0 \\ 0 & ... & ... & ... & ... & 0 \\ 0 & 0 & 0 & \lambda_{H-3} & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_{H-2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{B.1.5}$$

and

$$V = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ v_{0,1} & 0 & 0 & 0 & 0 & v_{H-1,1} \\ v_{0,2} & 0 & 0 & 0 & 0 & v_{H-1,2} \\ v_{0,..} & ... & ... & ... & ... & v_{H-1,..} \\ v_{0,H-3} & 0 & 0 & 0 & 0 & v_{H-1,H-3} \\ v_{0,H-2} & 0 & 0 & 0 & 0 & v_{H-1,H-2} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{B.1.6}$$

We can note that V is nilpotent and that $V\Lambda = 0$. It follows that for $k \geq 1$, $(\Lambda + V)^k = \Lambda^k + \Lambda^{k-1}V$, which we can see by induction. There is another identity that will be useful. If we define:

$$\Lambda' = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1/\lambda_1 & 0 & 0 & 0 \\ 0 & 0 & ... & 0 & 0 \\ 0 & 0 & 0 & 1/\lambda_{H-2} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{B.1.7}$$

where λ_1, λ_2... are the diagonal entries of l. We see that for $k \geq 2$, $\Lambda^{k-1} = \Lambda^k\Lambda'$. Since

$$\begin{aligned} T &= (DO)^{-1}Q(DO) \\ Q &= (DO)T(DO)^{-1} \\ Qt &= (DO)Tt(DO)^{-1}, \end{aligned} \tag{B.1.8}$$

we have:

$$\exp(Qt) = \sum_{k=0}^{\infty} \frac{1}{k!}(Qt)^k = \sum_{k=0}^{\infty} \frac{1}{k!}((DO)Tt(DO)^{-1})^k = DO\left(\sum_{k=0}^{\infty} \frac{1}{k!}(Tt)^k\right)(DO)^{-1} \tag{B.1.9}$$

And then we have that

$$\begin{aligned} \sum_{k=0}^{\infty} \frac{1}{k!}(Tt)^k &= \mathtt{I} + \sum_{k=1}^{\infty} \frac{1}{k!}((\Lambda + V)t)^k = \\ &= \mathtt{I} + \sum_{k=1}^{\infty} \frac{1}{k!}(\Lambda t)^k + \sum_{k=1}^{\infty} \frac{t^k}{k!}(\Lambda^{k-1}V) = \\ &= \exp(\Lambda t) + tV + \left(\sum_{k=2}^{\infty} \frac{t^k}{k!}\Lambda^{k-1}\right)V = \\ &= \exp(\Lambda t) + tV + \left(\sum_{k=0}^{\infty} \frac{t^k}{k!}\Lambda^k - \mathtt{I} - \Lambda t\right)\Lambda' V = \\ &= \exp(\Lambda t) + tV + (\exp(\Lambda t) - \mathtt{I} - \Lambda t)\,\Lambda' V. \end{aligned} \tag{B.1.10}$$

And finally:

$$\exp(Qt) = DO\left(\exp(\Lambda t) + tV + \left(\exp(\Lambda t) - \mathtt{I} - \Lambda t\right)\Lambda' V\right)(DO)^{-1} \tag{B.1.11}$$

So that, in terms of computing time, this requires us to compute o using an algorithm for hermitian matrices,then to compute d by recursion and the rest should follow from matrix multiplications. The advantage compared to the Padé approach described above is that most of the work is done once D and O are computed only once and reused for all time intervals.

In practice, the condition number of the matrix o can be very high leading to instabilities in the matrix exponentiation. Indeed the higher the condition number, the more sensitive the matrix will be to numerical operation. The condition number of our matrix can be of the order of 10^6 for large γ and is therefore ill-conditioned. Note that for the approximation of the diffusion process to the WF model, γ has to be on the order of 1. Thus, the matrix exponentiation becomes harder when the conditions for approximating the WF model with the diffusion are not necessarily met.

In order to overcome this problem we implemented the matrix exponentiation in *C++* using a library, *mpack* (Nakata, 2010), for multiple precision arithmetic. The library *mpack* is a multiple precision arithmetic version of *LAPACK* and *BLAS*. Although this allows to exponentiate the matrix for any γ in principle, it makes the matrix exponentiation step much slower. We therefore empirically test for which parameter range we require more precision than the double precision of *numpy* or *SciPy* that rely on *LAPACK*.

To do so, for a particular matrix $Q = Q(H, h, \gamma)$ we compute

$$\text{test}(Q) = \text{norm}\left((D \cdot O) \cdot (O^T D^{-1})\right) - \text{trace}\left((D \cdot O) \cdot (O^T D^{-1})\right) \tag{B.1.12}$$

where $\text{norm}(A) = \text{norm}((a_{ij})) = \sum_{i,j} |a_{ij}|$. The value of $\text{test}(Q)$ should be equal to 0. We choose a threshold value ϵ such that if $\text{test}(Q) > \epsilon$, we do not trust the default *SciPy* implementation and we invoke the higher precision computation. For this work we used $\epsilon = 10^{-5}$.

We plot on Figure B.1 the Boolean $\text{test}(Q) > \epsilon$ for different values of N_e and γ for $h = 0$. We can see on those plots that the matrix instability does depend on γ but not on the population size. For all the population sizes, the default implementation becomes unstable for $\gamma \gtrsim 40$.

To conclude, we use one existing method to exponentiate the matrix (Pad) and implemented one more method, with the possibility of increasing the double precision. Which method to use depends on the type of dataset and the parameter range one needs to explore. For high values of γ, if there are many time intervals, a method based on the spectral decomposition would be faster, otherwise the Pad approximation works well.

B.2 Choice of grids

As said in the main text, we investigated several grids inspired by Gutenkunst et al.. No matter the parameters, to compute the likelihood we need to approximate the transition probabilities between the original frequency of the A allele, $\frac{1}{2N_e}$, and another frequency between 0 and 1. Although we could extrapolate, we decided to use grids that all include the point $\frac{1}{2N_e}$.

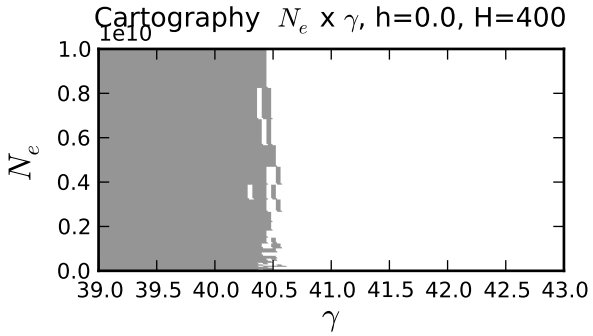

Figure B.1: One example cartography of the parameter combination that require higher precision for $\epsilon = 10^{-5}$. We plot the result of the Boolean operation test$(Q(H, h, \gamma, N_e) \leq \epsilon$ The legend is True for gray and white for False. We fix $H = 400$ and we plot N_e versus γ.

The first is a uniform grid with a point added at $\frac{1}{2N_e}$. We call this grid the "uniform grid". Then we investigate a quadratic grid and an exponential grid. The last two grids were chosen so that, as opposed to the uniform grid, the distance between adjacent points changes smoothly.

As before, let's denote $\{z_0, z_1, ..., z_H\}$ the state space of the one step process or the grid. The quadratic grid is described by a cubic equation, i.e., the difference between adjacent points is quadratic. We will assume for simplicity of notation that H is a multiple of 20 (it is straightforward to generalize), and that $G = \frac{H}{10}$. We set the first $G + 1$ points to form a uniform grid between 0 and $\frac{2}{2N_e}$, so that the median of this grid is $\frac{1}{2N_e}$. In other words, $z_j = \frac{j}{N_e G}$ for $0 \leq j \leq G$. Now we assume first that $\{q_0, \ldots, q_{H-G-1}\}$ is a uniform grid between 0 and 1. In other words, $q_0 = 0$, $q_{H-G-1} = 1$ and $q_j = \frac{j}{H-G-1}$. The remaining points are described by

$$z_{G+j} = aq_j^3 + bq_j^2 + cq_j + d \tag{B.2.13}$$

where $d = \frac{2}{2N_e}$, $c = \frac{1}{2N_e G}$, $b = -3(\frac{1}{H-G-1} + c + \frac{d}{H-G-1})\frac{1}{H-G-1}$, $a = -\frac{2}{3}b$.

The exponential grid will be defined as follows. If $\{u_0, ..., u_{H-1}\}$ is a uniform grid between -1 and 1 (i.e., $u_0 = -1$, $u_{H-1} = 1$ and $u_j = -1 + j\frac{2}{H-1}$), then the grid is

$$z_j = \frac{\frac{1}{1+\exp(-\beta u_j)} - \frac{1}{1+\exp(\beta)}}{\frac{1}{1+\exp(-\beta)} - \frac{1}{1+\exp(\beta)}}, \tag{B.2.14}$$

where β is a parameter that defines the density of the grid around the boundaries. We pick β such as $z_{\left[\frac{H}{10}\right]} = \frac{1}{2N_e}$, with [] denoting the integer part. To do so, we solve numerically the equation B.2.14 for $j = \left[\frac{H}{10}\right]$.

We plot the grids of interest versus uniform grids and the spacing between each point in Figure B.2.

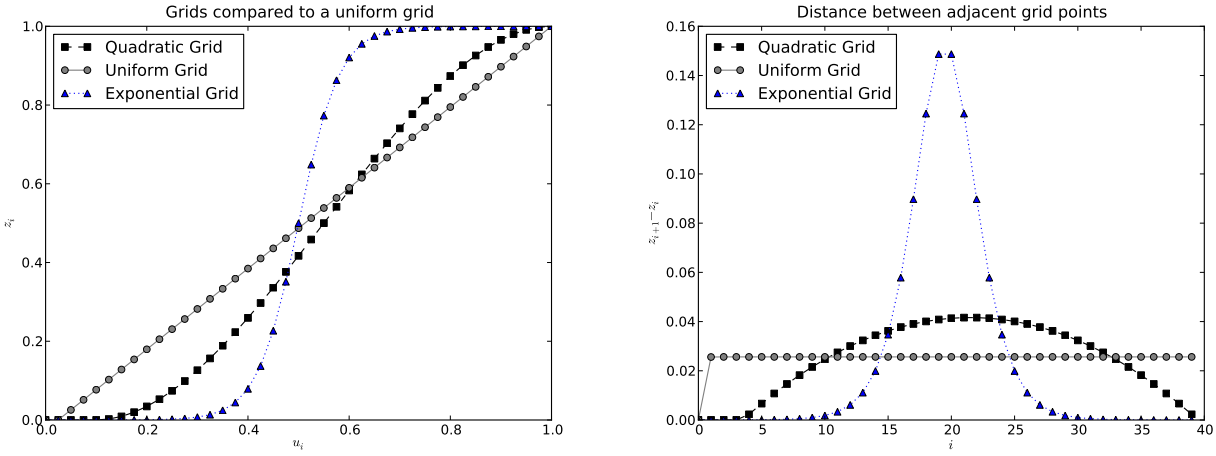

Figure B.2: Description of three different grids tested of size H=41 and $N_e = 10^4$. Left: the grids are plotted against a uniform grid of points between 0 and 1. Right: the spacing of adjacent points.

Bibliography

Adams, A. M. and Hudson, R. R. (2004). Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics*, 168(3):1699–712.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Contr ACM*, 19:716–723.

Albert, R., Berna, F., and Goldberg, P. (2010). Insights on Neanderthal fire use at Kebara Cave (Israel) through high resolution study of prehistoric combustion features: Evidence from phytoliths and thin sections. *Quaternary Int.*

Albrechtsen, A., Castella, S., Andersen, G., Hansen, T., Pedersen, O., and Nielsen, R. (2007). A Bayesian multilocus association method: allowing for higher-order interaction in association studies. *Genetics*, 176(2):1197–1208.

Anderson, E. C., Williamson, E. G., and Thompson, E. A. (2000). Monte Carlo evaluation of the likelihood for N(e) from temporally spaced samples. *Genetics*, 156(4):2109–18.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29.

Baker, M., Litvan, I., Houlden, H., Adamson, J., Dickson, D., Perez-Tur, J., Hardy, J., Lynch, T., Bigio, E., and Hutton, M. (1999). Association of an extended haplotype in the tau gene with progressive supranuclear palsy. *Hum Mol Genet*, 8(4):711–5.

Balding, D. J. (2005). *Weight-of-evidence for Forensic DNA Profiles*. John Wiley and Sons Ltd, Chichester, England.

Balding, D. J. and Nichols, R. A. (1994). DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Sci Int*, 64(2-3):125–40.

Balding, D. J. and Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96(1-2):3–12.

Bhaskar, A. and Song, Y. (2009). Multi-locus match probability in a finite population: A fundamental difference between the moran and wright-fisher models. *Proceedings of the 17th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB), Bioinformatics*, 25:i187–i195.

Bishop, Y., Fienberg, S., and Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge.

Blum, M. G. and Rosenberg, N. A. (2007). Estimating the number of ancestral lineages using a maximum-likelihood method based on rejection sampling. *Genetics*, 176(3):1741–57.

Bollback, J. P. and Huelsenbeck, J. P. (2007). Clonal interference is alleviated by high mutation rates in large populations. *Mol Biol Evol*, 24(6):1397–406.

Bollback, J. P., York, T. L., and Nielsen, R. (2008). Estimation of 2Nes from temporal allele frequency data. *Genetics*, 179(1):497–502.

Briggs, A. W., Good, J. M., Green, R. E., Krause, J., Maricic, T., Stenzel, U., Lalueza-Fox, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Schmitz, R., Doronichev, V. B., Golovanova, L. V., de la Rasilla, M., Fortea, J., Rosas, A., and Paabo, S. (2009). Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science*, 325(5938):318–21.

Briggs, A. W., Stenzel, U., Johnson, P. L., Green, R. E., Kelso, J., Prufer, K., Meyer, M., Krause, J., Ronan, M. T., Lachmann, M., and Paabo, S. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci U S A*, 104(37):14616–21.

Budowle, B., Shea, B., Niezgoda, S., and Chakraborty, R. (2001). CODIS STR loci data from 41 sample populations. *J Forensic Sci*, 46(3):453–89.

Bustamante, C. D., Wakeley, J., Sawyer, S., and Hartl, D. L. (2001). Directional selection and the site-frequency spectrum. *Genetics*, 159(4):1779–88.

Byrd, R. H., Lu, P., and Nocedal, J. (1995). A Limited Memory Algorithm for BoundConstrained Optimization. *SIAM Journal on Scientific and Statistical Computing*, 16(5):1190–1208.

Cadieu, E., Neff, M. W., Quignon, P., Walsh, K., Chase, K., Parker, H. G., Vonholdt, B. M., Rhue, A., Boyko, A., Byers, A., Wong, A., Mosher, D. S., Elkahloun, A. G., Spady, T. C., Andre, C., Lark, K. G., Cargill, M., Bustamante, C. D., Wayne, R. K., and Ostrander, E. A. (2009). Coat variation in the domestic dog is governed by variants in three genes. *Science*, 326(5949):150–153.

Caramelli, D., Lalueza-Fox, C., Condemi, S., Longo, L., Milani, L., Manfredini, A., de Saint Pierre, M., Adoni, F., Lari, M., Giunti, P., Ricci, S., Casoli, A., Calafell, F., Mallegni, F., Bertranpetit, J., Stanyon, R., Bertorelle, G., and Barbujani, G. (2006). A highly divergent mtDNA sequence in a Neandertal individual from Italy. *Curr Biol*, 16(16):R630–2.

Chen, H., Green, R. E., Paabo, S., and Slatkin, M. (2007). The joint allele-frequency spectrum in closely related species. *Genetics*, 177(1):387–98.

Cieslak, M., Pruvost, M., Benecke, N., Hofreiter, M., Morales, A., Reissmann, M., and Ludwig, A. (2010). Origin and history of mitochondrial DNA lineages in domestic horses. *PLoS One*, 5(12):e15311.

Committee on DNA Forensic Science: An Update, N. R. C. (1996). *The Evaluation of Forensic DNA Evidence*. National Academy Press.

Conroy, G. (2005). *Reconstructing human origins*. W. W. Norton, second edition edition.

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11:2463–2468.

Currat, M. and Excoffier, L. (2004). Modern humans did not admix with Neanderthals during their range expansion into Europe. *PLoS Biol*, 2(12):e421.

Currat, M., Excoffier, L., Maddison, W., Otto, S. P., Ray, N., Whitlock, M. C., and Yeaman, S. (2006). Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens" and "Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans". *Science*, 313(5784):172; author reply 172.

Curry, A. (2010). Decoding DNA: Evidence Suggests Early Humans Mated with Neanderthals. *Spiegel Online*.

Dessibourg, O. (2010). Quelque chose en nous de Neandertal. *Le Temps*.

Di Rienzo, A., Peterson, A., Garza, J., Valdes, A., Slatkin, M., and Freimer, N. (1994). Mutational processes of simple-sequence repeat loci in human populations. *Proc Natl Acad Sci USA*, 91(8):3166–3170.

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, 26:363–397.

Dikic, I. and Blaukat, A. (1999). Protein tyrosine kinase-mediated pathways in G protein-coupled receptor signaling. *Cell Biochemistry and Biophysics*, 30(3):369–387.

Drton, M., Sturmfels, B., and Sullivant, S. (2009). *Lectures on algebraic statistics*. Basel: Birkhüser, Oberwolfach Seminars, Vol. 40.

Drummond, A. J. and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*, 7:214.

Duarte, C., Mauricio, J., Pettitt, P. B., Souto, P., Trinkaus, E., van der Plicht, H., and Zilhao, J. (1999). The early Upper Paleolithic human skeleton from the Abrigo do Lagar Velho (Portugal) and modern human emergence in Iberia. *Proc Natl Acad Sci U S A*, 96(13):7604–9.

Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Mol Biol Evol*.

Durbin, R. M., Abecasis, G. R., Altshuler, D. L., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73.

Durret (2010). *Probability Models for DNA Sequence Evolution*. Springer, 2 edition.

Durrett, R. (2008). *Probability Models for DNA Sequence Evolution*. Springer, 2nd ed. edition.

Ellegren, H. (2000). Microsatellite mutations in the germline::: implications for evolutionary inference. *Trends in Genetics*, 16(12):551–558.

Emily, M., Mailund, T., Hein, J., Schauser, L., and Schierup, M. H. (2009). Using biological networks to search for interacting loci in genome-wide association studies. *European Journal of Human Genetics*, 17(10):1231–1240.

Eren, M. I., Greenspan, A., and Sampson, C. G. (2008). Are Upper Paleolithic blade cores more productive than Middle Paleolithic discoidal cores? A replication experiment. *J Hum Evol*, 55(6):952–61.

Eswaran, V., Harpending, H., and Rogers, A. R. (2005). Genomics refutes an exclusively African origin of humans. *J Hum Evol*, 49(1):1–18.

Evans, P. D., Gilbert, S. L., Mekel-Bobrov, N., Vallender, E. J., Anderson, J. R., Vaez-Azizi, L. M., Tishkoff, S. A., Hudson, R. R., and Lahn, B. T. (2005). Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. *Science*, 309(5741):1717–20.

Evett, I. and Weir, B. (1998). *Interpreting DNA Evidence*. Sunderland.

Ewens, W. J. (2004). *Mathematical Population Genetics*. Springer, second edition edition.

Fagundes, N. J., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F. M., Bonatto, S. L., and Excoffier, L. (2007). Statistical evaluation of alternative models of human evolution. *Proc Natl Acad Sci U S A*, 104(45):17614–9.

Fienberg, S. E. (1970). An iterative procedure for estimation in contingency tables. *Annals of Mathematical Statistics*, 41:907–917.

Finlayson, C. (2010). To mate, or not to mate: The Neandderthal question. *BBC news*.

Fisher, R. (1922). On the dominance ratio. *Proc. Roy. Soc. Edin.*, 42:321–341.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1995). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Goddard, M. E. and Hayes, B. J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat Rev Genet*, 10(6):381–91.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H., Hansen, N. F., Durand, E. Y., Malaspinas, A. S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prufer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Hober, B., Hoffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., and Paabo, S. (2010). A draft sequence of the Neandertal genome. *Science*, 328(5979):710–22.

Green, R. E., Krause, J., Ptak, S. E., Briggs, A. W., Ronan, M. T., Simons, J. F., Du, L., Egholm, M., Rothberg, J. M., Paunovic, M., and Paabo, S. (2006). Analysis of one million base pairs of Neanderthal DNA. *Nature*, 444(7117):330–6.

Green, R. E., Malaspinas, A. S., Krause, J., Briggs, A. W., Johnson, P. L., Uhler, C., Meyer, M., Good, J. M., Maricic, T., Stenzel, U., Prufer, K., Siebauer, M., Burbano, H. A., Ronan, M., Rothberg, J. M., Egholm, M., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Wikstrom, M., Laakkonen, L., Kelso, J., Slatkin, M., and Paabo, S. (2008). A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell*, 134(3):416–26.

Greig, A. V., Linge, C., and Burnstock, G. (2008). Purinergic receptors are part of a signalling system for proliferation and differentiation in distinct cell lineages in human anagen hair follicles. *Purinergic Signalling*, 4(4):331–338.

Gresham, D., Desai, M. M., Tucker, C. M., Jenq, H. T., Pai, D. A., Ward, A., DeSevo, C. G., Botstein, D., and Dunham, M. J. (2008). The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genet*, 4(12):e1000303.

Griffiths, R. C. (2003). The frequency spectrum of a mutation, and its age, in a general diffusion model. *Theor Popul Biol*, 64(2):241–51.

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 Genet*, 5(10):e1000695.

Hallgrimsdottir, I. B. and Yuster, D. S. (2008). A complete classification of epistatic two-locus models. *BMC Genetics*, 9:17.

Hamilton, D. (1994). *Time Series Analysis*. Princeton University Press.

Harding, R. M., Fullerton, S. M., Griffiths, R. C., Bond, J., Cox, M. J., Schneider, J. A., Moulin, D. S., and Clegg, J. B. (1997). Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.*, 60:772–789.

Harpending, H. C., Batze, M. A., Gurven, M., Jorde, L. B., Rogers, A. R., and Sherry, S. T. (1998). Genetic traces of ancient demography. *Proc. Nat. Acad. Sci.*, 95:1961–1967.

Harvati, K., Frost, S. R., and McNulty, K. P. (2004). Neanderthal taxonomy reconsidered: implications of 3D primate models of intra- and interspecific differences. *Proc Natl Acad Sci U S A*, 101(5):1147–52.

Herrera, K. J., Somarelli, J. A., Lowery, R. K., and Herrera, R. J. (2009). To what extent did Neanderthals and modern humans interact? *Biol Rev Camb Philos Soc*, 84(2):245–57.

Hodgson, J. A., Bergey, C. M., and Disotell, T. R. (2010). Neandertal genome: the ins and outs of African genetic diversity. *Curr Biol*, 20(12):R517–9.

Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M., and Paabo, S. (2001). Ancient DNA. *Nat Rev Genet*, 2(5):353–9.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–8.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95.

Jazin, E., Soodyall, H., Jalonen, P., Lindholm, E., Stoneking, M., and Gyllensten, U. (1998). Mitochondrial mutation rate revisited: hot spots and polymorphism. *Nat Genet*, 18(2):109–10.

Jones, E., Oliphant, T., Peterson, P., and & others (2001a). {SciPy}: Open source scientific tools for {Python}.

Jones, E., Oliphant, T., Peterson P., et al. (2001b). SciPy: Open source scientific tools for Python. "http://www.scipy.org/".

Jones, P., Chase, K., Martin, A., Davern, P., Ostrander, E. A., and Lark, K. G. (2008). Single-nucleotide-polymorphism-based association mapping of dog stereotypes. *Genetics*, 179(2):1033–44.

Joris, O. and Street, M. (2008). At the end of the 14C time scale–the Middle to Upper Paleolithic record of western Eurasia. *J Hum Evol*, 55(5):782–802.

Karlsson, E. K., Baranowska, I., Wade, C. M., Salmon Hillbertz, N. H., Zody, M. C., Anderson, N., Biagi, T. M., Patterson, N., Pielberg, G. R., Kulbokas, 3rd, E. J., Comstock, K. E., Keller, E. T., Mesirov, J. P., von Euler, H., Kampe, O., Hedhammar, A., Lander, E. S., Andersson, G., Andersson, L., and Lindblad-Toh, K. (2007). Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat Genet*, 39(11):1321–8.

Krause, J., Orlando, L., Serre, D., Viola, B., Prufer, K., Richards, M. P., Hublin, J. J., Hanni, C., Derevianko, A. P., and Paabo, S. (2007). Neanderthals in central Asia and Siberia. *Nature*, 449(7164):902–4.

Krings, M., Capelli, C., Tschentscher, F., Geisert, H., Meyer, S., von Haeseler, A., Grossschmidt, K., Possnert, G., Paunovic, M., and Paabo, S. (2000). A view of Neandertal genetic diversity. *Nat Genet*, 26(2):144–6.

Krings, M., Geisert, H., Schmitz, R. W., Krainitzki, H., and Paabo, S. (1999). DNA sequence of the mitochondrial hypervariable region II from the neandertal type specimen. *Proc Natl Acad Sci U S A*, 96(10):5581–5.

Krings, M., Stone, A., Schmitz, R. W., Krainitzki, H., Stoneking, M., and Paabo, S. (1997). Neandertal DNA sequences and the origin of modern humans. *Cell*, 90(1):19–30.

Kuensch, H. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241.

Lalueza-Fox, C., Krause, J., Caramelli, D., Catalano, G., Milani, L., Sampietro, M. L., Calafell, F., Martinez-Maza, C., Bastir, M., Garcia-Tabernero, A., de la Rasilla, M., Fortea, J., Paabo, S., Bertranpetit, J., and Rosas, A. (2006). Mitochondrial DNA of an Iberian Neandertal suggests a population affinity with other European Neandertals. *Curr Biol*, 16(16):R629–30.

Lalueza-Fox, C., Rompler, H., Caramelli, D., Staubert, C., Catalano, G., Hughes, D., Rohland, N., Pilli, E., Longo, L., Condemi, S., de la Rasilla, M., Fortea, J., Rosas, A., Stoneking, M., Schoneberg, T., Bertranpetit, J., and Hofreiter, M. (2007). A melanocortin 1 receptor allele suggests varying pigmentation among Neanderthals. *Science*, 318(5855):1453–5.

Lalueza-Fox, C., Sampietro, M. L., Caramelli, D., Puder, Y., Lari, M., Calafell, F., Martinez-Maza, C., Bastir, M., Fortea, J., de la Rasilla, M., Bertranpetit, J., and Rosas, A. (2005). Neandertal evolutionary genetics: mitochondrial DNA data from the iberian peninsula. *Mol Biol Evol*, 22(4):1077–81.

Laurie, C. and Weir, B. S. (2003). Dependency effects in multi-locus match probabilities. *Theor Popul Biol*, 63(3):207–19.

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., and Karlsson, E. K. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069):803–19.

Ludwig, A., Pruvost, M., Reissmann, M., Benecke, N., Brockmann, G. A., Castanos, P., Cieslak, M., Lippold, S., Llorente, L., Malaspinas, A. S., Slatkin, M., and Hofreiter, M. (2009). Coat color variation at the beginning of horse domestication. *Science*, 324(5926):485.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461:747–753.

Marchini, J., Donnelly, P., and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, 37:413–417.

Mather, K. A., Caicedo, A. L., Polato, N. R., Olsen, K. M., McCouch, S., and Purugganan, M. D. (2007). The extent of linkage disequilibrium in rice (Oryza sativa L.). *Genetics*, 177(4):2223–2232.

McCarroll, S. A., Huett, A., Kuballa, P., Chilewski, S. D., Landry, A., Goyette, P., Zody, M. C., Hall, J. L., Brant, S. R., Cho, J. H., Duerr, R. H., Silverberg, M. S., Taylor, K. D., Rioux, J. D., Altshuler, D., Daly, M. J., and Xavier, R. J. (2008). Deletion polymorphism upstream of irgm associated with altered irgm expression and crohn's disease. *Nature Genetics*, 40:1107–1112.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9:356–369.

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):1720–2.

Moler, C. and Van Loan, C. (2003). Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later*. *SIAM Review*, 45(1):3–000.

Nagylaki, T. (1980). The strong-migration limit in geographically structured populations. *Journal of Mathematical Biology*, 9(2):101–114.

Nakata, M. (2010). The MPACK (MBLAS/MLAPACK); a multiple precision arithmetic version of BLAS and LAPACK. URL: http://mplapack.sourceforge.net/. Enter text here.

Nei M. and Graur D. (1984). Extent of protein polymorphism and the neutral mutation theory. *Evol Biol.*, 17:73–118.

Nelder, J. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, 7:308–313.

Nelson, M. I. and Holmes, E. C. (2007). The evolution of epidemic influenza. *Nat Rev Genet*, 8(3):196–205.

Noonan, J. P., Coop, G., Kudaravalli, S., Smith, D., Krause, J., Alessi, J., Chen, F., Platt, D., Paabo, S., Pritchard, J. K., and Rubin, E. M. (2006). Sequencing and analysis of Neanderthal genomic DNA. *Science*, 314(5802):1113–8.

Nordborg, M. (1998). On the probability of Neanderthal ancestry. *Am J Hum Genet*, 63(4):1237–40.

Oliphant, T. (2006). *Guide to NumPy*. Trelgol Publishing.

Orlando, L., Darlu, P., Toussaint, M., Bonjean, D., Otte, M., and Hanni, C. (2006). Revisiting Neandertal diversity with a 100,000 year old mtDNA sequence. *Curr Biol*, 16(11):R400–2.

Outram, A. K., Stear, N. A., Bendrey, R., Olsen, S., Kasparov, A., Zaibert, V., Thorpe, N., and Evershed, R. P. (2009). The earliest horse harnessing and milking. *Science*, 323(5919):1332–5.

Ovchinnikov, I. V., Gotherstrom, A., Romanova, G. P., Kharitonov, V. M., Liden, K., and Goodwin, W. (2000). Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature*, 404(6777):490–3.

Patou-Mathis, M. (2006). *Neanderthal - Une autre humanité*. Perrin.

Pluzhnikov, A., Di Rienzo, A., and Hudson, R. R. (2002). Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics*, 161(3):1209–18.

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics*, 69:124–137.

Provine, W. B. (2001). *The Origins of Theoretical Population Genetics*. The University of Chicago Press.

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J. S., Albrechtsen, A., Moltke, I., Metspalu, M., Metspalu, E., Kivisild, T., Gupta, R., Bertalan, M., Nielsen, K., Gilbert, M. T., Wang, Y., Raghavan, M., Campos, P. F., Kamp, H. M., Wilson, A. S., Gledhill, A., Tridico, S., Bunce, M., Lorenzen, E. D., Binladen, J., Guo, X., Zhao, J., Zhang, X., Zhang, H., Li, Z., Chen, M., Orlando, L., Kristiansen, K., Bak, M., Tommerup, N., Bendixen, C., Pierre, T. L., Gronnow, B., Meldgaard, M., Andreasen, C., Fedorova, S. A., Osipova, L. P., Higham, T. F., Ramsey, C. B., Hansen, T. V., Nielsen, F. C., Crawford, M. H., Brunak, S., Sicheritz-Ponten, T., Villems, R., Nielsen, R., Krogh, A., Wang, J., and Willerslev, E. (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, 463(7282):757–62.

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L., Maricic, T., Good, J. M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E. E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M. V., Derevianko, A. P., Hublin, J. J., Kelso, J., Slatkin, M., and Paabo, S. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468(7327):1053–60.

Rieder, S., Taourit, S., Mariat, D., Langlois, B., and Guerin, G. (2001). Mutations in the agouti (ASIP), the extension (MC1R), and the brown (TYRP1) loci and their association to coat color phenotypes in horses (Equus caballus). *Mamm Genome*, 12(6):450–5.

Rincon, P. (2010). So we're part Neanderthal. What now? *BBC news*.

Roebroeks, W. and Villa, P. (2011). On the earliest evidence for habitual use of fire in Europe. *Proc Natl Acad Sci U S A*, 108(13):5209–14.

Rusk, N. (2009). Targeting ancient DNA. *Nature Methods*, 6:629.

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Res*, 15(11):1576–83.

Schmitz, R. W., Serre, D., Bonani, G., Feine, S., Hillgruber, F., Krainitzki, H., Paabo, S., and Smith, F. H. (2002). The Neandertal type site revisited: interdisciplinary investigations of skeletal remains from the Neander Valley, Germany. *Proc Natl Acad Sci U S A*, 99(20):13342–7.

Serre, D., Langaney, A., Chech, M., Teschler-Nicola, M., Paunovic, M., Mennecier, P., Hofreiter, M., Possnert, G., and Paabo, S. (2004). No evidence of Neandertal mtDNA contribution to early modern humans. *PLoS Biol*, 2(3):E57.

Slatkin, M. and Rannala, B. (2000). Estimating allele age. *Annu Rev Genomics Hum Genet*, 1:225–49.

Smith, F. H., Trinkaus, E., Pettitt, P. B., Karavanic, I., and Paunovic, M. (1999). Direct radiocarbon dates for Vindija G(1) and Velika Pecina late Pleistocene hominid remains. *Proc Natl Acad Sci U S A*, 96(22):12281–6.

Smith, J. M. and Haigh, J. (1974). The hitchhiking effect of a favourable gene. *Genetical Research*, 23:23–35.

Solecki, R. (1975). Shanidar IV, a Neanderthal Flower Burial Northern Iraq. *Science*, 190(4217):880–881.

Song, Y. S., Patil, A., Murphy, E. E., and Slatkin, M. (2009). Average probability that a "cold hit" in a DNA database search results in an erroneous attribution. *J Forensic Sci*, 54(1):22–7.

Song, Y. S. and Slatkin, M. (2007). A graphical approach to multi-locus match probability computation: revisiting the product rule. *Theor Popul Biol*, 72(1):96–110.

Stringer, C. (2002). Modern human origins: progress and prospects. *Philos Trans R Soc Lond B Biol Sci*, 357(1420):563–79.

Sutter, N. B., Eberle, M. A., Parker, H. G., Pullar, B. J., Kirkness, E. F., Kruglyak, L., and Ostrander, E. A. (2004). Extensive and breed-specific linkage disequilibrium in Canis familiaris. *Genome Res*, 14(12):2388–96.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S (2011). MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution*, In press. in press.

Tattersall, I. and Schwartz, J. H. (1999). Hominids and hybrids: the place of Neanderthals in human evolution. *Proc Natl Acad Sci U S A*, 96(13):7117–9.

Van Kampen, N. (1992). *Stochastic processes in physics and chemistry*. Elsevier.

Voight, B. F., Adams, A. M., Frisse, L. A., Qian, Y., Hudson, R. R., and Di Rienzo, A. (2005). Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc Natl Acad Sci U S A*, 102(51):18508–13.

Wade, N. (2010). Signs of Neanderthals Mating with Humans. *The New York Times*.

Wakeley, J. (2008). *Coalescent Theory: An Introduction*. Roberts & Company Publishers, Greenwood Village, Colorado, 1 edition.

Wall, J. D. and Kim, S. K. (2007). Inconsistencies in Neanderthal genomic DNA sequences. *PLoS Genet*, 3(10):1862–6.

Wall, J. D., Lohmueller, K. E., and Plagnol, V. (2009). Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol Biol Evol*, 26(8):1823–7.

Waples, R. S. (1989). A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics*, 121(2):379–91.

Wayne, R. K. and Ostrander, E. A. (2007). Lessons learned from the dog genome. *Trends Genet*, 23(11):557–67.

Weger, N. and Schlake, T. (2005). Igf-I signalling controls the hair growth cycle and the differentiation of hair shafts. *Journal of Investigative Dermatology*, 125(5):873–882.

Weir, B. S. (2004). Matching and partially-matching DNA profiles. *J. Forensic. Sci.*, 49:1009–1014.

Wichman, H. A., Millstein, J., and Bull, J. J. (2005). Adaptive molecular evolution for 13,000 phage generations: a possible arms race. *Genetics*, 170(1):19–31.

Williamson, E. G. and Slatkin, M. (1999). Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics*, 152(2):755–61.

Wiuf, C. (2006). Consistency of estimators of population scaled parameters using composite likelihood. *J Math Biol*, 53(5):821–41.

Wolff, G. L., Stanley, J. S., Ferguson, M. E., Simpson, P. M., Ronis, M. J., and Badger, T. M. (2007). Agouti signaling protein stimulates cell division in "viable yellow" (A^{vy}/a) mouse liver. *Experimental Biology and Medicine (Maywood)*, 232(10):1326–1329.

Wright, F. A., Huang, H., Guan, X., Gamiel, K., Jeffries, C., Barry, W. T., Pardo-Manuel de Villena, F., Sullivan, P. F., Wilhelmsen, K. C., and Zou, F. (2007). Simulating association studies: a data-based resampling method for candidate regions or whole genome scans. *Bioinformatics*, 23:2581–2588.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, 16:97–159.

Zhang, Y. and Liu, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, 39:1167–1173.

Zhu, C., Byrd, R. H., and Nocedal, J. (1997). L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization. *ACM Transactions on Mathematical Software,*, 23(4):550–560.

Zilhao, J., Angelucci, D. E., Badal-Garcia, E., d'Errico, F., Daniel, F., Dayet, L., Douka, K., Higham, T. F., Martinez-Sanchez, M. J., Montes-Bernardez, R., Murcia-Mascaros, S., Perez-Sirvent, C., Roldan-Garcia, C., Vanhaeren, M., Villaverde, V., Wood, R., and Zapata, J. (2010). Symbolic use of marine shells and mineral pigments by Iberian Neandertals. *Proc Natl Acad Sci U S A*, 107(3):1023–8.