Approaches For Inferring Past Population Size Changes From Genome-wide Genetic Data

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

im Fachgebiet
Informatik

vorgelegt

von Diplom Informatiker Christoph Theunert
geboren am 13.04.1983 in Lutherstadt Wittenberg.

Die Annahme der Dissertation wurde empfohlen von:

1. Prof. Dr. Peter F. Stadler, Universität Leipzig
2. Prof. Dr. Mark Stoneking, MPI EVA Leipzig
3. Prof. Dr. Dirk Metzler, LMU München

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 06.06.2014 mit dem Gesamtprädikat
Magna Cum Laude.

Abstract

The history of populations or species is of fundamental importance in a variety of areas. Gaining details about demographic, cultural, climatic or political aspects of the past may provide insights that improve the understanding of how populations have evolved over time and how they may evolve in future. Different types of resources can be informative about different periods of time.

One especially important resource is genetic data, either from a single individual or a group of organisms. Environmental conditions and circumstances can directly affect the existence and success of a group of individuals. Since genetic material gets passed on from generation to generation, traces of past events can still be detected in today's genetic data. For many decades scientists have tried to understand the principles of how external influences can directly affect the appearance and features of populations, leading to theoretical models that can interpret modern day genetic variation in the light of past events.

Among other influencing factors like migration, natural selection, etc., population size changes can have a great impact on the genetic diversity of a group of organisms. For example, in the field of conservation biology, gaining insights into how the size of a population evolves may assist in detecting past or ongoing temporal reductions of population size. This seems crucial since the reduction in size also correlates with a reduction in genetic diversity which in turn might negatively affect the evolutionary potential of a population. Using computational and population genetics methods, sequences from whole genomes can be scanned for traces of such events and, therefore, assist in new interpretations of historical details of populations or groups of interest.

This thesis focuses on the detection and interpretation of past population size changes. Two approaches to infer particular parameters from underlying demographic models are described. The first part of this thesis introduces two summary statistics which were designed to detect fluctuations in size from genome-wide Single Nucleotide Polymorphism (SNP) data. Demographic inferences from such data are inherently complicated due to recombination and ascertainment bias. Hence, two new statistics are introduced: allele frequency-identity by descent (AF-IBD) and allele frequency-identity by state (AF-IBS). Both make use of linkage disequilibrium information and exhibit defined relationships to the time of the underlying mathematical process. A fast and efficient Approximate Bayesian Computation framework based on AF-IBD and AF-IBS is constructed that can accurately estimate demographic parameters. These two statistics were tested for the biasing effects of hidden recombination events, ascertainment bias, and phasing errors. The statistics were found to be robust to a variety of these tested biases. The inference approach was then applied to genome-wide SNP data to infer the demographic histories of two human populations: (i) *Yoruba* from Africa and (ii) *French* from Europe. Results suggest that AF-IBD and AF-IBS are able to capture sufficient amounts of information from underlying data sets in order to accurately infer parameters of interest, such as the beginning, end, and strength of periods of varying size. Additionally, the results from empirical data suggest a rather stable ancestral population size with a mild recent expansion for Yoruba, whereas the French apparently experienced a rather long-lasting strong bottleneck followed by a drastic population growth.

The second part of this thesis introduces a new way of summarizing information from the site frequency spectrum. Commonly applied site frequency spectrum based inference methods make use of allele frequency information from individual segregating sites. Our newly developed method, the 2 point spectrum, summarizes allele frequency information from all possible pairs of segregating sites, thereby increasing the number of potentially informative values from the same underlying data set. These additional information are then incorporated into a Markov Chain Monte Carlo framework. This allows for a high degree of flexibility and implements an efficient method to infer population size trajectories over time. We tested the method on a variety of different simulated data sets from underlying demographic models. Furthermore, we compared the performance and accuracy of our method to already established methods like PSMC and diCal. Results indicate that this non-parametric 2 point spectrum method can accurately infer the extent and times of past population size changes and, therefore, correctly estimates the history of temporal size fluctuations. Furthermore, the initial results suggest that the amount of required data and the accuracy of the final results are comparable with other publicly available non-parametric methods. An easy to use command line program was implemented and will be made publicly available.

In summary, we introduced three highly sensitive summary statistics and proposed different approaches to infer parameters from demographic models of interest. Both methods provide powerful frameworks for accurate parameter inference from genome-wide genetic data. They were tested for a variety of demographic models and provide highly accurate results. They may be used in the settings as described above or incorporated into already existing inference frameworks. Nevertheless, the statistics should prove useful for new insights into populations, especially those with complex demographic histories.

Acknowledgements

I wish to thank all people who contributed to the success of this work. First of all I'd like to thank Mark Stoneking. He has always given me the freedom and support that I needed, but also helped and encouraged me whenever I lost track of the overall goal. His guidance and experience helped me a lot over all the years. Furthermore, Kun Tang, despite having been in Shanghai most of the time, introduced me to the topic of population genetics and raised my interest in the field of method development. Without their help and encouragement this work would not have been possible. Furthermore I wish to thank Peter Stadler. He has continuously given me the freedom and trust I needed to realize my own thoughts and ideas. It has always been a great pleasure to collaborate with these exceptional personalities. Kun Tang, Paul Jenkins, Daniel Falush and Daniel Živković have contributed significant parts to the current work. Another profound thanks goes to many colleagues from the MPI that have made my time in Leipzig such an amazing and wonderful period of my life. Thanks to all the present and past members of the Stoneking group, the Genetics Department, the institute and especially to David, Jesse, Alex, Frenchi, Torsten, Mark, Mimi and Lydi ... for uncountable hours of discussions, meetings, joy and pleasure. In addition, thanks to all the people who accompanied me throughout the whole decade in Leipzig. A special thanks to Matze for the exciting years in our first WG, the wonderful Cossi bike rides and our friendship. Arnoldo has been a big support and friend over all the years. The mutual encouraging, suffering and disputes have given me so much.

The last and most important thanks goes to my whole family (plus Sukrö) and my partner Anne. There's not much I can say, just because the support, trust and love you have given me all my life cannot be expressed in words. *Thank you.*

Danksagungen

Ich möchte mich bei allen Leuten, die zum Erfolg dieser Arbeit beigetragen haben, recht herzlich bedanken. Zu allererst möchte ich Mark Stoneking danken. Er hat mir immer die Freiheit und die Unterstützung gegeben die ich brauchte. Er hat mir geholfen und mich ermutigt wenn ich den Überblick verloren hatte. Seine Hilfe und Erfahrung haben mir über all die Jahre sehr geholfen. Weiterhin danke ich Kun Tang, der, trotzdem er die meiste Zeit in Shanghai gearbeitet hat, mein Interesse für dieses Gebiet der Populationsgenetik geweckt hat. Ohne ihre Hilfe und Ermutigung wäre diese Arbeit nicht möglich gewesen. Desweiteren möchte ich Peter Stadler danken. Er hat mir ununterbrochen die Freiheit und das Vertrauen gegeben, welche ich für die Realisierung meiner Arbeit benötigte. Es war mir eine große Ehre mit solch erfahrenen und außergewöhnlichen Persönlichkeiten zusammenzuarbeiten. Kun Tang, Paul Jenkins, Daniel Falush und Daniel Živković haben einen entscheidenden Teil zu dieser Arbeit beigetragen. Ein weiterer tiefer Dank gilt meinen

Kollegen vom MPI, die meine Zeit in Leipzig zu einem so wunderbaren Abschnitt in meinem Leben gemacht haben. Danke an alle aktuellen und damaligen Mitglieder der Stoneking Gruppe, der Genetik Abteilung und des Instituts. Besonderer Dank an David, Jesse, Alex, Frenchi, Torsten, Mark, Mimi und Lydi... für die unzähligen Diskussionen, Meetings, Spaß und Freude. Außerdem geht ein großer Dank an alle, die mich über die Jahre in Leipzig begleitet haben. Ein besonderer Dank gilt Matze für die aufregende Zeit in unserer ersten WG, die wunderbaren Fahrradtouren an den Cossi und unsere Freundschaft. Arnoldo war seit jeher eine große Unterstützung und ein Freund über all die Jahre. Die gegenseitigen Ermutigungen, das Leiden und die Streitigkeiten haben mir viel gegeben.

Der letzte aber wichtigste Dank gilt meiner gesamten Familie (inklusive Sukrö), sowie meiner Partnerin Anne. Es gibt nicht viel, was ich sagen kann. Die Unterstützung, das Vertrauen und die Liebe, die ihr mir mein ganzes Leben entgegengebracht habt, sind nicht in Worte zu fassen. *Danke.*

Contents

1 Introduction **1**
- 1.1 Motivation 1
- 1.2 Subject of this thesis 4
- 1.3 Organization of this thesis 5

2 Background **7**
- 2.1 The principles of population genetic inference 7
- 2.2 Simulating evolution 8
- 2.3 The features of population genetic data 10
  - 2.3.1 Single Nucleotide Polymorphisms 11
  - 2.3.2 Haplotype phasing 12
- 2.4 Bayesian inference, MCMC and ABC 12
  - 2.4.1 Bayesian inference 13
  - 2.4.2 Monte Carlo method 13
  - 2.4.3 Markov chains and Markov Chain Monte Carlo 14
  - 2.4.4 Approximate Bayesian Computation 15
- 2.5 Definitions 17

3 Software, algorithms and fundamental concepts **21**
- 3.1 Software 21
- 3.2 The basic coalescent 23
  - 3.2.1 The Wright Fisher model 24
  - 3.2.2 Generating genealogies 26
  - 3.2.3 Adding mutations to genealogies 27
- 3.3 Expanding the basic coalescent 29
  - 3.3.1 The coalescent with changes in population size 30
  - 3.3.2 Genealogical effects of variable population size 31
  - 3.3.3 The coalescent with recombination 33

4 Importance of population size changes **37**

4.1 Overview 37
    4.1.1 Classical population size inference approaches 38
    4.1.2 Recent genome-wide approaches 40
    4.1.3 Perspectives 45

5 The AF-IBD/ AF-IBS method 47
5.1 Introduction 47
5.2 AF-IBD & AF-IBS 48
    5.2.1 Examining AF-IBD under various demographic scenarios . . . 50
    5.2.2 Parameter estimation with AF-IBD using ABC 52
    5.2.3 Use of AF-IBS for sequence or polymorphism data 53
    5.2.4 Parameter estimation with AF-IBS using ABC 54
    5.2.5 Model misspecification 55
    5.2.6 Parameter estimation for empirical data 56
5.3 Results for AF-IBD & AF-IBS 57
    5.3.1 Properties of AF-IBD 57
    5.3.2 AF-IBD-ABC 58
    5.3.3 Properties of AF-IBS 59
    5.3.4 The IBS/IBD Ratio 59
    5.3.5 AF-IBS-ABC 59
    5.3.6 Application to genome-wide data 60
    5.3.7 Model misspecification 61
    5.3.8 ABC analysis for Yoruba 61
    5.3.9 ABC analysis for French 61
5.4 Discussion AF-IBD & AF-IBS 63

6 The 2 point spectrum method 71
6.1 Introduction 71
6.2 Idea of the 2 point spectrum 73
6.3 Implementation of the 2 point spectrum 75
6.4 Obtaining the joint moments 78
6.5 MCMC approach 83
    6.5.1 RJMCMC overview 83
6.6 Chain length 87
6.7 Performance optimization 88
6.8 Parameters and options 90
6.9 Likelihood calculation and data sets 91
6.10 Results 2 point spectrum 92
    6.10.1 RJMCMC results 94
6.11 Runtime observations and optimal parameter settings 102
6.12 Discussion 106

7 Discussion 111

8 Appendix 115
8.1 AF-IBD AF-IBS appendix . . . 115
8.1.1 Part I . . . 115
8.1.2 Part II . . . 116
8.1.3 Part III . . . 116
8.1.4 Part IV . . . 117
8.2 Exact calculation of E[$T_j T_k$] . . . 122

List of Abbreviations 123

List of Figures 125

List of Tables 127

Bibliography 129

Chapter 1

Introduction

> *"The scientist is not a person who gives the right answers, he's one who asks the right questions."*
>
> *Claude Lévi-Strauss (1908 - 2009)*

1.1 Motivation

Our world as it exists today is only a snapshot of a process that started several billion years ago. Some of the oldest known traces of life, which most likely represented bacterial organisms found in some banded iron formation rocks, date back to 3.5 billion years ago [35]. Ever since the conditions on our planet were sufficient for the origin of life, a process called evolution has constantly been changing the inheritable characteristics of organisms from generation to generation. Life on earth has been influenced by the prevalent environmental conditions, forces, and factors at respective periods over time. On the other hand, the appearance and features of our environment are as well influenced and shaped by all kinds of living forms. Mutual dependence and control have created an enormous diversity of which we as the human species only represent a small part.

The interest and curiosity about the origins of life in general, or about particular species, is widespread. This interest can be expressed on a personal level as historical research and genealogical investigation, or on a wider scale at the level of society (the general scientific investigation of our past). These goals can be achieved by means of a variety of different resources. Not only do we have access to the past by historical traditions (either oral or in writing), but by records of the past enclosed in ice, soil, rocks, etc. Each of the available sources provides information about different periods and ranges of time. The written word and oral histories were handed down through the centuries and their very first appearance can be dated back to around five thousand years ago, providing insights into the recent part of history [69]. Evolutionary linguistics is the scientific study of the origins and evolution of languages. By studying the history of spoken languages today and performing a detailed comparison among themselves, additional insights into the history of the populations that spoke them can be obtained. Linguistics can, therefore, assist in the identification of historical patterns of migrations and foundations of human populations. However, rough estimates suggest that languages do not retain evidence of their origins for more than six thousand years. Further down the timeline, overlapping with the origin of modern humans, physical objects that have been used and shaped by human contact can be used as archaeological record. Examples are ceramics, stone tools, ornaments, houses, etc. Depending on the location and age of these specimens, interesting facts about human and non-human history, as far back as two million years, can be acquired. Paleontology and Paleoclimatology are sciences that investigate and analyze fossil remains of living organisms or their traces, and physical remains (e.g. ice) that contain information about past climatic conditions. These sources can provide insights into the entire period of life existing on our planet [69].

All mentioned data and methods are stand-alone sciences, but their interaction can be an even more powerful way of studying the past focusing on different periods of times from a variety of different materials and resources. Nevertheless, one important field of research has not been mentioned so far. Every time living organisms reproduce, parts of their genomes are mixed and new genomic variations are generated. The genetic material we can analyze today contains a record of the past. Since environmental conditions (i.e. changes) directly affect the behavior and existence of individuals, these factors leave traces in the inheritable material of each individual over time. Information can either be obtained from genomes of present day organisms, or ancient DNA from especially well preserved samples. Of course, all the available data we have today, no matter from which source, are biased representations of the true past, since only a tiny part is still existent today. Social rules, climatic conditions, environmental changes, demographic events, natural selection, etc. are limiting the amount, quality, and geographical locations of present day records that provide information about the foretime. It should, therefore, be mentioned, that even analyzing all possible information together is still just an approximation to the truth. However, with an increasing standard of technology and computational power this approximation is continuously getting more and more reliable.

Depending on the addressed question, suitable data and information can be obtained, or vice versa, depending on the available data, only certain types of questions can be answered. This is due to the fact that different records only provide insights into particular periods of time (and into either social, demographic, genetic, or climatic, etc. aspects). Nevertheless, although the combined analysis of interdisciplinary records might assist in answering particular questions, they are to some extent independent pictures of the past and might give different conflicting signals that are even harder to interpret and to disentangle than single results. Therefore, a standardized and simple approach to analyze all kinds of different data and to answer questions as diverse as possible does not exist at all. To account for this, caution should be applied at any time and specific methods and analyses need to be designed for every new question.

The current work focuses on one particular aspect of records of the past, namely genetic data. As already mentioned, what we see in today's genomes provides insights into how populations of a species of interest evolved tens, or hundreds, or even thousands of generations ago. Therefore, one of the most overwhelming biological discoveries of the last centuries is the fact that all features, characteristics, and genetic properties of organisms alive today are derived from ancestors that can be traced back over millions of years. In 1859, Charles Darwin, a British biologist, published his highly significant work "*On the Origin of Species*" [25], stating that all species are related by common descent and that the vast diversity observed is just a product of the accumulation of small, but favorable, modifications over large periods of time. "... say, that after a certain unknown number of generations, some bird had given birth to a woodpecker, and some plant to the mistletoe, and that these had been produced perfect as we now see them; but this assumption seems to me to be no explanation, for it leaves the case of the coadaptation of organic beings to each other and to their physical conditions of life, untouched and unexplained." (Charles Darwin, On the origin of species, 1959, p. 3-4, [25]). This quote expresses a deep sense of curiosity and skepticism that is required to unravel the mysteries and hidden secrets of our past. Because Humans share a common ancestor with every other species on this planet - with the never-ending process of evolution adjusting the genomes of individuals to face new environmental challenges, allowing to conquer new habitats and geographical locations - genetic data represent a book full of historical information. This common ancestry allows us to draw conclusions by looking at more or less related and diverged species - conclusions that could not be drawn by just looking at the human genome.

As I hope will be clear at the end of this thesis, the role of genetics in the past decades has largely been used for testing hypotheses that were derived based on data from other disciplines. This is mainly due to the fact that different evolutionary processes (i.e. different histories) can create similar genetic patterns whose interpretation is not easily feasible by means of only genetic methods. However, with the advent of ancient DNA analyses (e.g. [46]) and improved and sophisticated inference methods, more reliable conclusions from genetic data have been and possibly will be obtained in the future.

The question *why* one should study the evolution and history of living organisms is a complex combination of smaller problems and interests. For example, unlike many other non-human species, human populations have never in the past been as large as they are now[1]. How long have species been in growth or decline phases? How and when did they expand to new habitats and geographical regions? When and how did patterns of gene flow across multiple populations start and evolve? But also medical aspects are becoming more and more important, focusing on the analysis of variable drug response in different individuals, association studies, and inferences about gene function from patterns of genetic variation [43]. For example, reconstructing the demographic history of species is essential for the purpose of finding evidence for adaptation at the molecular level (e.g. [29]) and results of these kinds of investigations can assist in the fine determination of disease gene positions (e.g. [70]).

Evolutionary anthropology and population genetics, both independent sciences, try to answer questions concerning the origin, history, structure, migration patterns, and relationships of groups of individuals, be it from humans (anthropology) or any other species on our planet (e.g. [69]). They also try to develop methods by which one can infer how populations are shaped by evolutionary or demographic factors like mutation, selection, recombination, population size changes, or changes in the availability of resources, etc. The goal is to understand the forces that produce and maintain genetic variation within and between species.

1.2 Subject of this thesis

Over the past few years I have been involved in a variety of population genetic studies, mostly as part of a group of colleagues from Leipzig, but also as part of studies that were lead by scientists from all over the world. Because of the diverse topics of the particular projects I was able to deepen my knowledge in a large number of studies from theoretical and practical population genetics. Although I could talk about a couple of exciting topics, the current work only focuses on my two most comprehensive projects. Since for this work I took a major role in the design and implementation of all the procedures, this thesis is written in first-person form throughout most of the parts. However, I want to emphatically stress that *none of the achieved results would have been possible without the ideas, support, guidance, and brilliance of a lot of people. They encouraged me and suggested valuable ideas throughout all the states of my time at the Max Planck Institute in Leipzig.* So, whenever I use the first-person form it should be clear that this work was deeply dependent on many people's ingenious supports. I explicitly mention and thank all the involved people in the acknowledgements and here again would like to express my deepest appreciation. In the text I explicitly cite the work that other people contributed.

---

[1] http://www.census.gov/popclock/, last visited on 14/12/2013

The idea for this thesis arose from the need for new methods to infer demographic details from the history of organisms, in particular the past population size changes. Modern population genetics methods can make use of a large amount of genetic data. Often these data are first summarized by some summary statistics, each of them capturing specific aspects of the underlying diversity distribution (e.g. [113]). Depending on the question of interest, great efforts have been payed to develop new summary statistics or to modify the usage of conventional ones in order to gain better insights. On the other hand, data are often biased toward certain directions, therefore, statistics should be constructed to be robust toward bias. The more robust a summary statistic is to a bias, the more reliable the results of parameter inference are.

Part of the main goal of the current work was not to rely only on known and common statistics, but to develop new ways of summarizing data in order to contribute new details to the overall question of population history conclusions. The first project in this thesis (AF-IBS/IBD) is a consequent derivative of the results of my Diploma thesis from 2009 [131]. The original goal of that previous study was to find summary statistics that were sensitive enough to detect past population size changes and, at the same time, are robust to ascertainment bias (see chapter 5). Summarizing this prior work, we were able to show that the statistics we chose did exhibit distinct patterns when calculated from data based on different demographic scenarios. However, as discussed in its outline, further steps were needed and did include the use of these statistics in a comprehensive framework of parameter inference. This then inspired the work I have been doing for the last couple of years in this field, leading to the results discussed in the current dissertation (including the second project presented here, the 2 point spectrum method, see chapter 6).

In summary, I will present two methods to infer past population size changes from genetic data. The methods I propose here are mostly tested for human demographic models, but I emphasize that *the methods are generally applicable to genetic data from other species as well, as long as sufficient amounts of data are available.*

1.3 Organization of this thesis

This thesis is structured in a way that allows the reader to gradually understand the initial idea and all aspects that rely on it. The remainder of this work is organized as follows: After the introduction, chapter 2 briefly explains all general backgrounds that are needed to introduce the reader to this field of population genetic analyses. Chapter 3 gives a short summary on the software that was used, the details of the underlying mathematical models, and fundamental concepts on which all my analyses are built on. In chapter 4 similar work and projects that aim to infer past population size changes, mostly based on single statistics that try to detect deviations from a standard model of constant population size, are introduced. Furthermore, that chapter will give a brief introduction into more recent genome-wide approaches. The core part of this thesis contains two main projects that will be discussed separately in chapter 5 and 6. Chapter 5 describes the first project, the AF-IBD/IBS statistics and the use of Approximate Bayesian Computation. Chapter 6 presents a work that implements a MCMC approach using the 2 point spectrum statistics. Both chapters 5 and 6 were written in the style of research articles, containing their own introduction, methods, results, and discussion section. After chapter 6, a final discussion section will be given, trying to compare both approaches in a more general view. In addition it provides an outlook on the potentials and possible further developments of the newly proposed methods. After all, the appendix and the references of this thesis are listed. Since the methods I have been developing during my time as a PhD student at MPI EVA are advancements of my Diploma thesis, chapters 2 and 3 are inspired by this work [131]. Brief definitions and explanations of unknown technical terms can be found at the end of chapter 2. A list of all nomenclatures and abbreviations is given at the end of the appendix chapter 8.

Chapter 2

Background

"Millions saw the apple fall, Newton was the only one who asked why?"

Bernard M. Baruch (1870 - 1965)

The purpose of this chapter is to give a basic introduction to topics necessary for an understanding of this work.

2.1 The principles of population genetic inference

One of the fundamental questions in population genetics is how to make use of data. In this field *data* means genetic information sampled from any kind of living creatures, be it animals (including humans) or plants. There are two basic ways of looking at data, in particular non-parametric and parametric methods. In the non-parametric case one tries to reason things by not relying on models, whereas parametric methods provide explicit models for the data. For example, in phylogenetics, the study of evolutionary relatedness among several groups of organisms, methods that use a parsimony (non-parametric) approach are applied. They assess a topology by using a relatively simple metric. This metric can be the minimum number of character state changes necessary to generate the data on a given tree, not presuming any specific model or distribution. A different class of methods uses a likelihood (parametric) approach, which assesses a topology using the assumption, that there exists an evolutionary model from which the data are identically distributed, to infer phylogenetic trees.

Population genetic models are used to combine the information about observed sequence variation (resulting in the part of an organism's genome that makes it unique compared to other organisms of the same species, see section 2.3.1) with assumptions about the history and demographic influences that shape the picture of a population. Most of the methods used to infer and to understand the history of such are model based. The overall goal of every model is to depict a reality under examination (in this case genetic data) by means of certain declarations within the scope of a scientific theory (population genetic models). One of the first and most basic models was the standard Wright-Fisher model, a stochastic process used to describe how genes from one generation get transferred to the next one. This model has been widely used and still continues to be. Compared to realistic populations, all models developed so far contain many unrealistic assumptions. For example, the Wright-Fisher model assumes a population of constant size, with uniform rates of mutation and recombination across the genome, with all individuals being equally fit, i.e. random mating [141]. Migration models like the Island model [141] assume no mutation, no selection, and that populations are divided into 'islands' of an equal size which can exchange genes per generation at the same rate and persist indefinitely. Another migration model, the stepping stone model, tries to bypass the assumption of no population substructure by introducing the possibility of exchanging genes between adjacent discrete subpopulations [76]. Despite all the advantages and simplifications these models yield, the main problem is to decide which one to use and regarding the tremendous variety of model parameters, demographic scenarios, and biological causes, it is still not possible to get the "correct" model from data. If assumptions are not met, inferred results can be incorrect. A natural approach for modeling would be to start with the simplest possible model that appropriately characterizes the data. Dealing with populations (e.g. humans) one, therefore, assumes no population structure, no recombination, and no selection which also defines the null hypothesis: "Nothing interesting ever happens in biology". The first step is to estimate some parameters of the so called null model, assuming it is correct, e.g. mutation rate and population size. After that, questions of interest, for example whether this simple model is a good description of the data, or, if not, what can be said about the forces that cause the deviation from the null model, can be asked. As a consequence, each step includes the comparison of observed (or empirical) data and artificial data.

2.2 Simulating evolution

To make inference about the past and answer questions of interest (as indicated in the previous section), simulated data are often required for each step. The challenge of this is to emulate a process of creating genetic data representing the true evolution as well as possible. Unfortunately, it is yet not feasible to include all existing natural forces, circumstances, and parameters that shape the genetic diversity of organisms, since this would result in a search with an underlying high dimensional space not computable in a fair amount of time. Due to this fact, the complexity of the simulation process needs to be simplified in a way that still ensures to get reasonable results.

In principle there exist two ways of simulation that can be distinguished. The most obvious and natural way of simulating evolution would be to start in the past and end in the present, thus time is running forwards. But there exist problems for which considering time running backwards is easier and faster. In fact, in the last 20 years, the most important tool has been backward or coalescent simulation [77–79]. This development was a natural consequence of the availability of genetic data sampled at the present but shaped by past processes. To get an idea of how this approach can be applied, a family tree can be imagined. One way of drawing this kind of graph is to start with an individual in the present and go backwards in time following the lineages through parents, grandparents, etc. This example can be used as a good illustration for why, at least in some cases, time should be considered running backwards. To apply the opposite method, drawing the tree forward in time, also called forward simulation (e.g. [108]), each individual alive from a specific population at a certain point in the past needs to be considered. With the information about each parent to child relationship the tree could be drawn for each generation until the present day, implying that every family lineage can be traced back to a day forward in time. The crucial point is that just a very small part of all the information is sufficient to create the family tree. Since just the lineages leading to a specific individual need to be kept, forward simulation represents a memory inefficient way of using data. But it should be noted that the family tree example only serves as an example for a better understanding. These so called *genealogies*, the ancestry of sequences back to their most recent common ancestor (MRCA), refer to the genetic ancestry of a sample at a locus in a genome and not to the described usual definition of a genealogy being the family relationships of a set of individuals. Figure 2.1 shows the basic principle of the coalescent and one possible graphical representation of genealogies. A detailed description of the coalescent process itself will be given in section 3.2.

Referring to figure 2.1, "genealogy" and "tree" are used interchangeably throughout this thesis. Both cases refer to the use of edges, nodes, and node times to represent a rooted history. Coalescent approaches are fast and have been the best choice for several years, for they are easy to handle and were computable with the available computer power at that time. But one important drawback needs to be mentioned. The unusual way of the coalescent process to emulate the evolution results in limitations in terms of the complexity of scenarios that can be simulated, such as environmental effects, selection, population structures, recombination events, or the complexity of natural genetic data. The continuous progress in computational power and memory capacities exploit new capabilities to benefit from the more realistic approach of forward simulations.

Since the complexity of demographic models used in this work is relatively simple, e.g. no selection is assumed and very time consuming ABC and MCMC approaches

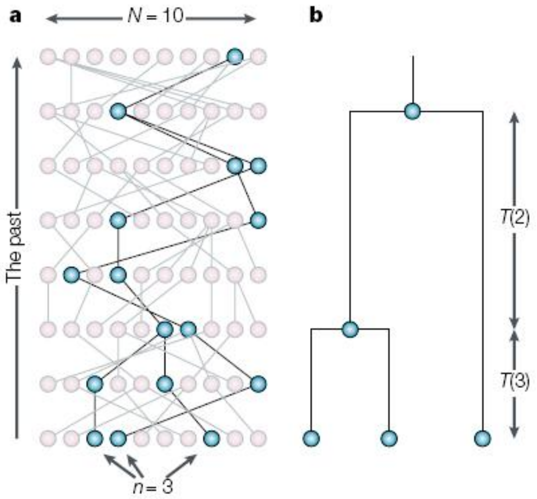


Figure 2.1: The basic principle behind the coalescent. A) shows a so called genealogy for a population of size N=10. The black lines represent the backtracking of n=3 sampled lineages to their common ancestor. The present day population is at the bottom and the most ancient population at top of the picture, indicating that time is running backwards. B) shows the sub genealogy for the 3 sampled lineages. Tj, j= 2, 3 is the time while there are j ancestors to the sampled 3 lineages. Image was adopted from [116]

are applied, coalescent software was used to simulate huge amounts of genetic data in the current work. For this purpose the advantages of the coalescent (speed and simplicity) prevailed.

2.3 The features of population genetic data

As in the field of conventional statistical inference the amount and quality of the available data is usually sufficient to get satisfactory results. Not only are there independent data points and a sample space of low dimensions, but also analytical formulations for inference using all possible information are feasible. In the case of population genetics data usually represent a single draw from the evolutionary process. Due to the tremendous amount of natural forces that shape the genetic structure of a population (e.g. recombination, migration, natural selection, genetic drift - see section 2.5), the sample space that needs to be considered consists of many dimensions. All these facts implicate that analytical formulations for inference using all potential information are generally impossible or at least hard to derive.

2.3.1 Single Nucleotide Polymorphisms

Since the empirical data used for the AF-IBD/S method in this thesis are SNP data, this section gives a short introduction to SNPs (pronunciation: "snips"). Most of the DNA between different members of a species or between paired chromosomes of an individual is completely the same [1]. In the case of evolutionary genetics, the study of how one individual genome differs from another, implications about the past and the present of organisms are derived from the varying part of a genome. Often there is little or no additional information when considering the positions between the SNPs in a particular population [104]. Imagine the DNA sequence ACGT**A**. A mutation can cause a change to a single nucleotide within this sequence, creating ACGT**G**. These are positions that vary within or between populations. If this mutation gets transmitted and consequently occurs in a certain number of individuals it can be referred to as a SNP, a single base pair mutation at a specific locus, resulting in two alleles, namely A and G.

Beside a large amount of various genetic markers, for example RFLP (Restriction fragment length polymorphism), AFLP (Amplified fragment length polymorphism), STR (Short tandem repeat), etc., single nucleotide polymorphisms are frequently chosen for studies of linkage and for studies of historical demography. Compared to other markers, SNPs are comprehensively available in the (human) genome. Furthermore they can be efficiently assayed and analyzed [137]. Approximately 90% of the variation of the human genome is represented by SNPs and they can be found every 100 to 300 bases along the human genome which consists of 3 billion bases [41]. If the SNP occurs in a protein coding gene and the amino acid is changed by the mutation, the SNP is called non-synonymous, otherwise it is called synonymous.

There exist various ways of identifying SNPs, e.g. genotyping, sequencing, screening databases for expressed sequence tags (ESTs), etc. [104, 128]. Due to economic reasons SNPs are often first discovered in a small sample of individuals ('discovery panel') and later genotyping of these SNPs in a larger sample follows. This approach results in the fact that only SNPs detected in the small sample can later be typed in the larger sample and the probability that a SNP can be identified in the small sample is crucially dependent on the allele frequency [19]. In general there are at least three possibilities how a candidate site is determined to be a SNP. Those sites may be classified as a SNP because it is a polymorphic site in the actual sample ("sample SNPs"), or because it is a polymorphic site in a panel drawn from the same population ("panel SNP"), or it is a polymorphic site in a panel from a different, but related, population ("different population panel SNP") [83].

As a consequence all statistical properties of genotype frequencies of the larger sample are more or less different from what would be expected by full resequencing of this sample [137]. For example, SNPs at a high frequency in the larger panel are more likely to go undiscovered when these SNPs only have a low frequency in the smaller sample. This fact, called *ascertainment bias*, has a massive impact on how inference of any underlying history or parameter has to be carried out. Not considering the effect of ascertainment bias can lead to erroneous results when inference is solely based on the site frequency spectrum of the larger sample. Several methods of correcting for ascertainment bias are published (e.g. [86, 105]) and consist of predictions of properties of those SNPs that are missing in the larger sample, resulting in a set of SNPs that would have been observed with full resequencing.

---

[1] see http://hapmap.ncbi.nlm.nih.gov/whatishapmap.html.en, last visited on 04/12/13

2.3.2 Haplotype phasing

With the new advances in technology an increasing amount of data is produced from comprehensive and low-cost genome-wide SNP microarrays and from ever more affordable whole-genome and whole-exome sequencing tools. Empirical data are usually obtained as unphased genotype data, which is subject to an additional statistical calculation of phase reconstruction to infer the haplotype composition. The problem arises from the fact that sequence and SNP array data generally take the form of unphased genotypes, i.e. it is not directly observed which of the two parental chromosomes, or haplotypes, a particular allele falls on [12]. However, there exist a large number of computational methods to infer the phase of haplotypes from underlying data. In general, the number of unrelated individuals present in a sample is an important factor in determining how well the phase can be estimated: the more individuals, the better and easier the estimation. In order to illustrate the mentioned problem imagine a diploid organism and two bi-allelic loci (e.g. SNPs) on the same chromosome with the first locus having alleles G or C and the second locus having alleles A or T. Hence, there exist three possible genotypes at both loci: (GG, GC, and CC) and (AA, AT, and TT). More information can be found in a relatively recent study that reviews publicly available programs and methods and discusses general issues [12].

2.4 Bayesian inference, MCMC and ABC

Since Markov Chain Monte Carlo (*MCMC*) and Approximate Bayesian Computation (ABC) are applied in this work, this section will give a short introduction explaining the background, advantages and disadvantages of both. By doing so, the reasons for choosing ABC for the first and MCMC for the second method should become clear. If not explicitly stated, all theoretical facts and information are based on [5, 20, 42, 54].

Markov Chain Monte Carlo means Monte Carlo integration using Markov chains. To make inferences about model parameters it is often necessary to integrate over high-dimensional probability distributions. In Bayesian inference, a kind of statistical inference, Bayesians need to make predictions about a posterior distribution of model parameters given the data.

2.4.1 Bayesian inference

In Bayesian statistics observations can be used to update a probability of any given thesis being true. Probability acts as a direct measure of uncertainty that might or might not represent a long term frequency as it would in the frequentistic way of inference.

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{\int P(\theta)P(D|\theta)d\theta} \tag{2.1}$$

Equation 2.1 is the conditional distribution of θ given the data D, where θ denotes model parameters and missing data and D are the observed data. This is the object of all Bayesian inference and is called *posterior distribution* of θ. It consists of two parts, the prior $P(\theta)$ and the likelihood $P(D|\theta)$. The prior can be seen as any available information inferred before the data were observed and the likelihood is the conditional probability of the data D given a particular parameter θ. Any desired properties of this distribution are valid for Bayesian inference, e.g. quantiles, high density regions, moments, mode, mean etc. In high dimensional spaces it is however often not possible either to solve the integrations in a feasible time, or to solve them at all. In most cases an analytical evaluation of equation 2.1 is impossible. That's why a common problem in practical Bayesian statistics is that the distribution is only known up to a specific normalizing constant. $P(\theta)P(D|\theta)$ is known, but the normalizing constant $\int P(\theta)P(D|\theta)d\theta$, integrating over all possible models respectively parameters, is unknown and a convenient way of defining the posterior distribution is $posterior \propto likelihood \times prior$. Equation 2.1 encloses the beliefs about θ after having observed the data and in consideration of the chosen prior information [20].

2.4.2 Monte Carlo method

In many fields of applications one needs to make inference about a distribution of interest. In population genetics, e.g. the distribution of the population mutation rate given the data needs to be calculated. In this case one object would be to calculate a function of interest f() for any vector Y, denoting model parameters and or missing data, with ψ being the posterior distribution. As described in section 2.4.1, the expectation E[f(Y)] of this function cannot be easily calculated, since the posterior distribution is known only up to a normalizing constant. But one way of doing this is to draw samples {Y$_t$,t=1,...,n} from ψ and making use of the law of large numbers and approximate

$$E[f(Y)] \approx \frac{1}{n}\sum_{t=1}^{n} f(Y_t) \tag{2.2}$$

In theory this approximation can be made as precise as desired just by increasing the sample size n and making sure the samples are drawn as independent as possible. A further drawback is the independence of the samples. Usually it is not easily feasible to draw independent samples from ψ, yet this distribution can have any possible non-standard shape. However, this fact can be circumvented by creating a process that draws samples throughout the support of ψ. One possibility is to create a Markov chain having ψ as its stationary distribution.

2.4.3 Markov chains and Markov Chain Monte Carlo

A Markov chain is a stochastic process having the Markov property. Most of the time Markov chains are discrete in time and space. It generates a series of random variables such that the probability distribution of future states is completely determined by the current state of the chain. Knowing just a part of the history of the chain is as good as knowing all of its history to make predictions about the future. This is what is called the Markov property. Formally, let $X_n, n = 0, 1, 2....$ be a series of random variables, be it scalars or vectors, and having a joint distribution such that

$$P(X_n|X_{n-1}, X_{n-2}, ..., X_0) = P(X_n|X_{n-1})\forall n \tag{2.3}$$

As a consequence, at each time point n ≥ 0, the next state X_{n+1} is sampled from a distribution $P(X_{n+1}|X_n)$. The sequence of random variables is called a Markov chain and $P(.|.)$ is called the transition kernel of the chain. As n increases the chain will gradually forget its initial state. Under certain conditions the distribution of X_n given the starting state X_0, described by $P^{(n)}(X_n|X_0)$, will converge to a unique stationary distribution $\phi(.)$ which is not dependent on X_0 or on n anymore. The more iterations the chain is being run, the less the sampled points X_n seem to be affected by the starting states and thus more and more look like being sampled from $\phi(.)$. The amount of iterations, say m, after which the sampled points can be assumed to be dependent samples from the stationary distribution is called burn-in. Often a chain is said to be in an equilibrium state after having reached the stationary distribution. As described in section 2.4.2, one possibility in Monte Carlo methods to calculate equation 2.2, is to construct a Markov chain having the distribution of interest as its stationary distribution.

The first proposal to do this came from Metropolis et al. in 1953 [99]. This so called *Metropolis Hastings* algorithm can be used to construct a Markov chain creating samples from a required posterior distribution. Basically, the algorithm can be divided into three steps [2]:

1. If currently at any state X_n, propose a move to a candidate point Y according to a transition kernel q(X→Y), which is a probability of moving from state X to state Y
2. Now calculate the probability $\alpha(X_n, Y)$, that the candidate point Y will be accepted as a new state X_{n+1} of the chain (also called acceptance or hastings ratio):
$$\alpha(X,Y) = min\left(1, \frac{\psi(Y)q(X|Y)}{\psi(X)q(Y|X)}\right) \tag{2.4}$$
Since $\psi(.)$ is the posterior distribution of interest,
$\psi(Y) = P(Y|D) = \frac{P(D|Y)P(Y)}{\int P(Y)P(D|Y)dY}$
3. Now move to state $X_{n+1} = Y$ with probability α or stay at X_n and go back to step 1

If the proposed candidate is not accepted, $X_{n+1} = X_n$ and the chain does not move further. It can be shown that once a candidate Y_n was sampled from the stationary distribution, Y_{n+1} will be also. The crucial point of the acceptance ratio is that knowledge of likelihood multiplied by the prior is sufficient for implementation, since the often high dimensional integrals in the denominator, known as the normalizing constant, can be reduced by applied math.

The algorithm results in a sequence of sampled points which can be assumed to be dependent on an underlying distribution of interest. Now, equation 2.2 can be used to infer the desired features from that distribution but having the characteristic of dependent samples.

2.4.4 Approximate Bayesian Computation

Although Markov Chain Monte Carlo can be a very efficient approach to generate observations from a posterior distribution (equation 2.1), its dependence on knowing a likelihood function can be a disadvantage in some cases. For more complex probability models and an increasing number of nuisance parameters, likelihoods are either impossible to derive or not computable in a tolerable amount of time. To bypass these kind of problems, observations from a posterior distribution can be made without relying on likelihoods. The idea is to replace the full data with summary statistics by summarizing a large amount of the data into a few representative values. This is what is called *Approximate Bayesian Computation* or MCMC without likelihoods [8, 95].

---

[2] adopted from [42]

One of the first implementations came from Tavaré et al. 1997 [130] and was one of the first *rejection sampling* methods to simulate an approximate posterior random sample. There exist many variations on rejection sampling themes. The basic idea as well as the simplest approach for a rejection method is [3]:

1. Generate θ from its prior distribution $\pi(.)$
2. Accept θ with probability $h = P(D|\theta)$ and return to step 1

In this scheme a likelihood still needs to be calculated and this can be impractical or even impossible in many cases. In several steps this simple scheme was improved, resulting in an approximate Bayesian computation scheme for data D and summary statistics S[4]:

1. Generate θ from $\pi(.)$
2. Simulate D' from stochastic model M with parameter θ, and compute the corresponding statistics S'
3. Calculate the distance p(S,S') between S and S'
4. Accept θ if p $\leq \epsilon$ and return to step 1

This improved rejection scheme can be combined with MCMC and the algorithm can be divided into 6 steps [5]:

1. If currently at state X, propose a move to a new candidate point according to a transition kernel q(X$\rightarrow$Y), which is a probability of moving from state X to state Y
2. Now generate some new data D' using an underlying model M with parameters Y
3. Calculate some summary statistics S' of D' and calculate the same statistics S of data D
4. If p(S,S') $\leq \epsilon$, for a given distance measure p and a given threshold value ϵ, go to step 5, otherwise stay at state X and go back to state 1

---

[3] adopted from Marjoram 2003 [95]

[4] adopted from Marjoram 2003 [95]

[5] adopted from Plagnol and Tavaré 2002 [109]

5. Now calculate the probability $\alpha(X_n, Y)$, that the candidate point Y will be accepted as a new state X_{n+1} of the chain :
$$\alpha(X,Y) = min\left(1, \frac{\pi(Y)q(X|Y)}{\pi(X)q(Y|X)}\right), \tag{2.5}$$
where π(Y) is the prior of state Y
6. Move to Y with probability $\alpha(X_n, Y)$ or stay at X and return to step 1

The difference to the previous MCMC schemes is that the likelihood $P(D|Y)$ does not contribute to the calculation of the acceptance ratio. Knowledge of the prior π(Y) is sufficient to run an ABC algorithm and the stationary distribution is $P(Y|p(S, S') \leq \epsilon)$.

The general advantages of these rejection methods are that they are easy to code and offer the opportunity for parallel computation since they produce independent samples. Otherwise, the more complex an underlying demographic model and with it the underlying probability model is, the less effective it is to solely sample from the prior, since prior and posterior can be rather different [109]. Another practical drawback of rejection-sampling methods, as described above, is that the number of summary statistics to be used is limited. It is also in general hard to anticipate the effect of different summary statistics, leading to the need of intuition. The more statistics are included, the lower the acceptance rates become. As a consequence the tolerance ϵ must be increased which can negatively influence the approximation of the posterior distribution. The approach that is used in chapter 5 is from [8] where the authors introduce two improvements, smooth weighting and regression analysis which shall overcome the described sensitivity of the approximation to ϵ. As will be explained later in this work, the amount of data and the time needed to simulate evolution is large. However, Approximate Bayesian Computation still seems to be suitable, although it has some not negligible practical issues. Details of application will be described in section 5.

2.5 Definitions

Definition 1 (Allele) *An allele is one of several possible forms of a gene or DNA sequence that is located at a certain locus on a chromosome.*

Definition 2 (Allele frequency) *The frequency of an allele (e.g. within a population).*

Definition 3 (Effective population size) *Wright:"The number of breeding individuals in an idealized population that would show the same amount of dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration."*

Definition 4 (Evolution) *The modification of the inheritable characteristics of a population of living organisms from generation to generation.*

Definition 5 (Genealogy) *A genealogy usually represents the inheritance relationships between alleles. It's similar in form to a phylogenetic tree.*

Definition 6 (Generation) *Regarding the Wright-Fisher model of reproduction, a generation is the time period from conception to reproduction (often assumed to be 25 years for humans; depending on the type of data and species). Additionally, reproduction and death are simultaneous for all individuals and synchronous for all individuals.*

Definition 7 (Genetic drift) *Genetic drift is one of the basic mechanisms of evolution. An allele occurs with a certain frequency in a population. Due to the fact that alleles in the offspring are a random sample of those in the parents, some individuals possibly leave more descendants than others. Hence, it may cause genetic variants to be removed from a population and reduces the genetic variability due to chance variations.*

Definition 8 (Haplotype) *A haplotype in the understanding of this thesis is a set of single nucleotide polymorphisms, i.e. the combination of allelic states of a set of polymorphic markers. They are all lying on the same chromosome or region of a chromosome.*

Definition 9 (Identity by descent) *Two or more alleles are identical by descent (IBD) if they are identical copies of the same ancestral allele.*

Definition 10 (Identity by state) *Two or more alleles are identical by state (IBS) if they just share the same mutational expression.*

Definition 11 (Infinite sites mutation model) *A mutation model which assumes that each new mutation changes a different nucleotide. Therefore the number of mutations on the coalescent genealogy is similar to the number of segregating sites. No recurrent mutations are allowed.*

Definition 12 (Linkage Disequilibrium (LD)) *A non-random association between alleles in a population. Alleles that are observed together more often than would be expected by chance, are said to be in LD. They are co-inherited due to reduced recombination between them.*

Definition 13 (Migration) *Describes the process of an individual or a population moving from one inhabited region to another.*

Definition 14 (Most recent common ancestor) *The most recent individual from which all individuals of a set of organisms are directly descended.*

Definition 15 (Mutation) *Any changes in a DNA sequence of an organism. Hence, mutations create variation in the complete set of unique alleles in a species or population.*

Definition 16 (Natural selection) *The varying contribution of individuals to the next generation based on their power to survive and reproduce.*

Definition 17 (Population) *The summation of all individuals of the same group or species, that live in the same geographical area and are able to interbreed, is called population.*

Definition 18 (Recombination) *The exchange of DNA between the members of a chromosomal pair. This process usually takes place in meiosis.*

Definition 19 (Selective sweep) *The rapid process where an advantageous allele, and other alleles linked to it, increases in a population until all other alleles go extinct and the locus only has one allele.*

Definition 20 (Site frequency spectrum (SFS)) *Given the infinite sites mutation model, for a sample of size k,* $F_i(k)$ *denotes the number of sites at which exactly i individuals carry a mutation. A vector* $\mathrm{V}=(F_1(k), F_2(k), \ldots, F_k(k))$ *is then called the site frequency spectrum of the sample.*

Definition 21 (Zygosity) *The degree of identity for the alleles of a trait in an organism. In the case of a diploid individual, the individual is heterozygous at a specific locus, if both alleles are different. If both alleles are the same, the organism is homozygous at that locus.*

Chapter 3

Software, algorithms and fundamental concepts

"Essentially, all models are wrong, but some are useful."

George E. P. Box (1919 - 2013)

The function of this chapter is to give a summary of the software that was used, the algorithms behind it, and the concepts that can be derived from that. The profound description of these basic ideas will allow for a sufficient understanding of the results of this work.
If not explicitly stated, all theoretical facts and information are based on [5, 20, 42, 47, 54].

3.1 Software

In the last few years the need for flexible and efficient software to simulate more or less large DNA fragments, evolving under complex evolutionary models, yielded the development of a variety of simulation programs, be it coalescent or forward simulators. As already mentioned in section 2.2, the data simulations in this work are all based on a coalescent approach allowing to test new ways of summarizing data to detect deviations from neutral demography.

In order to uncover the single effects of different demographic scenarios on various summary statistics the simulated demographic models need to be as simple as possible at the beginning. The more complex the scenarios are, the more difficult it is to eliminate side effects caused by combined demographic factors. Among others, this fact simplified the decision process and supported the choice of coalescent as a simulation tool for this work. Coalescent is relatively easy and straightforward to use, computationally faster than forward simulation, and the variety of different programs, each with its own pros and cons, fulfills the needs of a large amount of distinctive problems and makes it easy to find the appropriate program for ones own purpose. In [14, 31, 59] authors give a good survey of not all, but a large number of coalescent and forward simulators.

Hudsons *ms* is one of the first and most classical programs [62]. It can be used to generate many independent replicate samples including factors like migration, recombination, and population size changes. It uses a standard coalescent approach assuming an infinite sites model of mutation which does not allow multiple-hits or back mutations to occur. Since the output of *ms* is very easy to process, it perfectly fits into the entire workflow pipeline for AF-IBS and the 2 point spectrum method. For each simulated population the output contains the position for each polymorphic site (*SNP*) on a scale of (0,1). Also, the haplotypes of each of the sampled chromosomes are given, consisting of a string of zeros and ones. An ancestral state is coded with a 0 and the derived, also called mutant state, is coded with a 1 (see the output of *ms* in figure 3.1). The details of how such samples are generated by the coalescent process are given in the next section 3.2. The Approximate Bayesian Computation analyses crucially rely on a software called *abcEst_2* [30] whose application will be described in chapter 5.

Regarding the AF-IBD/IBS method, the processing of the *ms* output as well as the workflow pipeline are coded in Perl [1], a stable cross platform programming language that is, among others, widely used in the field of bioinformatics. One reason for using Perl was Bio::PopMX, a large Perl package developed by Kun Tang, Marc Bauchet, and Christoph Theunert [2], that allows handling and conversion of a huge variety of different genetic data formats as well as calculation of lots of different population genetic algorithms, tests, and statistics. The workflow pipeline is embedded in the Bio::PopMX environment, permitting a comfortable way of analyzing the data.

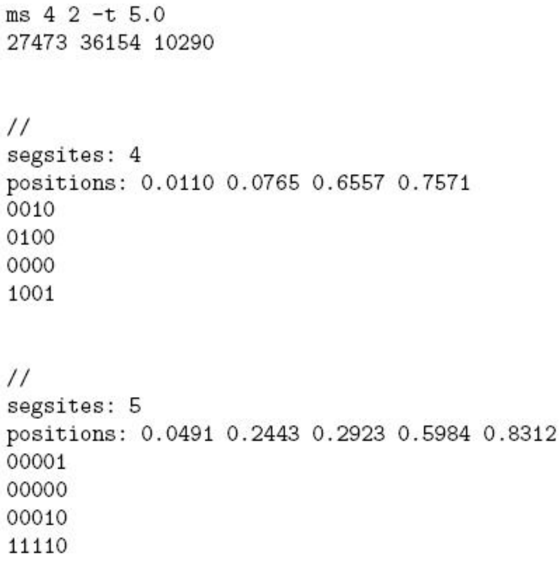
For the second project, the 2 point spectrum method, the underlying rjMCMC framework (see chapter 6 for more details) was implemented in the C language, one of the most widely used programming languages of all time. Many online recourses are available for details of practical implementations and extensions [3]. One of the reasons C better fits to the requirements of the second project is that C is a rather low level language and C scripts need to be compiled before run, i.e. the source code needs to be translated into machine code. This is a fundamental difference

---

[1] www.perl.org

[2] At the time of this thesis, Bio::PopMX was not yet published. Please contact christoph_theunert@eva.mpg.de

[3] see for example http://cm.bell-labs.com/who/dmr/chist.html, last visited on 14/12/2013



```
ms 4 2 -t 5.0
27473 36154 10290


//
segsites: 4
positions: 0.0110 0.0765 0.6557 0.7571
0010
0100
0000
1001


//
segsites: 5
positions: 0.0491 0.2443 0.2923 0.5984 0.8312
00001
00000
00010
11110
```

Figure 3.1: Shown is the output of *ms* for two samples ("populations") each containing four sampled chromosomes (haplotypes are shown). The first two lines are the command line and the random number seeds used to generate the data. Each new sample is indicated by "//". The first sample has four segregating (polymorphic) sites, so there have been four different mutations on the entire coalescent tree, since *ms* uses an infinite sites mutation model. The second sample has five segregating sites.

to languages like Perl where interpreters use a step-by-step execution of the source code and no pre-runtime translation takes place. Due to the compilation process the scripts of the C language are quicker at runtime when computationally intensive algorithms need to be run.

3.2 The basic coalescent

This section is not meant to give a complete introduction to the process known as the coalescent. Rather its purpose is to explain only the details necessary to understand the following ideas, statistics, and results.

In order to construct and analyze random genealogies, the coalescent has become the standard model for this purpose [116] which describes the connection between demographic history and genetic data and provides a framework for extracting information from samples of DNA sequences. Since all data in this work are based on the level of genes, the following meaning of population is best understood as a population of genes rather than individuals. The coalescent is a stochastic process providing good approximations to the distribution of ancestral histories, resulting from (e.g.) the Wright-Fisher or the Moran model. Three basic ideas make up the process of coalescent [80]. In short:

1. Tracing back the ancestry of a gene backward in time, building up a tree of genes, at a particular locus, in a population sample back to some point in the past where they have their single common ancestor
2. For a variety of demographic models, the stochastic structure of the genealogy does not depend on the detail of the reproductive mechanism, assuming no selection and finite population size
3. The process of adding mutations to a genealogy is independent from the genealogy itself

To be able to make predictions about ancestral events, dating of events, likely parameter values, etc., probability models of mutation, geography, and reproduction structures are indispensable. Without such, questions like: “Is there a sign of population structure, recombination, selection, or population growth in the data ?” could not be answered.

The central approach of analyses focusing on genealogies is a stochastic characterization of the genealogies that relate sequences. Following that, evaluation of the probability of a given data set consists of modeling reproduction in the population, leading to a probabilistic description of the genealogical relationship of sampled data and second, a genealogy will produce data with a specific probability if combined with a mutation model. That means, (1) simulate the genealogy of n genes and (2) add mutations to the genealogy according to the chosen mutation model.

3.2.1 The Wright Fisher model

As already mentioned in section 2.1, the first population model was introduced by Wright and Fisher. It provides a description of the evolution of an idealized population and how genes from one generation are transmitted to the next (where the term gene can refer to any material transmitted from generation to generation) [141]. Two versions of the model exist, namely the diploid version with N diploid individuals and the haploid version with 2N haploid individuals, both assuming a population size of 2N genes [54]. There are some simplifying assumptions which are made in the basic Wright-Fisher model of reproduction according to [141] and [54]:

1. Discrete and non-overlapping generations
2. Haploid individuals or two subpopulations, namely males and females
3. Constant population size

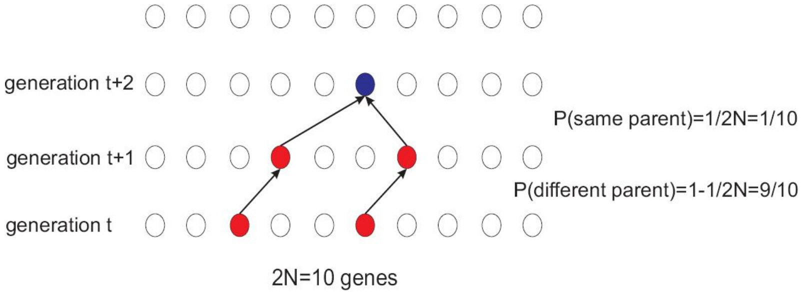


Figure 3.2: A present day sample contains 2N=10 different genes in generation t (indicated by points). Each gene in generation t needs to find a parent in generation t+1. The probability for 2 genes from generation t to find a different parent in generation t+1 is p=9/10 (indicated by the red points). This "sampling" is performed 2N times with replacement. In this case 2 genes from generation t+1 find the same parent in generation t+2 with probability p=1/10 (indicated by blue point).

4. Assumption of no recombination
5. Populations have no geographical or social structure
6. All individuals are equally fit, meaning no selective pressure

Most of these constraints are not realistic at all, since real world populations are not likely to behave like this and the process becomes much more mathematically complex if these assumptions are relaxed. But with these simple constraints the basic math can be derived.

Starting with a present day population at time t, each gene needs to find a parent in the previous generation t+1. If a gene in generation t+1 was not chosen as a parent, its lineage dies out.

Since this sampling scheme is applied in each generation, there is either success (finding the same parent) or failure (finding different parents) (see figure 3.2). Each time two lineages find the same ancestor it is called a *coalescent event*. This is an example of a binomial distribution Bi(m,p) with parameters v_i being the number of descendants of gene i in generation t, m=2N and p=1/(2N).

$$P(v_i = k) = \binom{2N}{k} \left(\frac{1}{2N}\right)^k \left(1 - \frac{1}{2N}\right)^{(2N-k)} \tag{3.1}$$

The mean of this distribution is 1 which ensures the population size being constant. As shown in figure 3.2 at the blue point, the 2 lineages from the red genes combine or coalesce in this point. For a set of n genes this point is called the most recent common ancestor, or *MRCA*, representing the most recent gene, or more generally individual, from which all genes are descended. If 2N is large, then v_i is almost Poisson distributed. If the number of ancestral genes is small, these calculations are not valid anymore, since for the Wright-Fisher model they are based on large sample size properties. This is the reason for the coalescent process to assist. In almost all possible cases the coalescent process explains real data significantly better than the Poisson prediction [48].

The coalescent was here introduced in the setting of the Wright-Fisher model of neutral evolution, i.e. assuming no selection. But it applies to a variety of models and it can be shown, that many different neutral models converge to the original coalescent[5].

3.2.2 Generating genealogies

There are two ways of measuring time. The discrete-time coalescent is based on the time being measured in generations on a discrete scale approximated by the geometric distribution. On the other hand, time can be continuous and the analogous distribution is the exponential distribution, being a limit distribution of a set of geometric distributions measured on a finer scale of time points. Both systems are coexisting and have their own mathematical representations and peculiarities. As mentioned in section 3.2.1, time in the Wright-Fisher population model is measured in discrete units, namely generations. The side effect of using a geometric distribution to represent properties of such a genealogy is approximation so that, e.g. the true probability that two genes will find their common ancestor j generations ago can slightly differ from the calculated, approximated value. Therefore it is often conceptually and also computationally advantageous to use the exponential distribution, since it is easier to use its properties to calculate important quantities of a genealogy and, the larger N, the more accurate the approximations are. In the continuous time coalescent one unit of time refers to the average time for two genes to find their common ancestor which is 2N generations. Different implementations can have different time scalings, e.g. N or 4N instead of 2N, but results are similar up to a factor of 2. If necessary it is easy to switch from continuous, denoted by t, back to discrete time, denoted by j, just with j=2N*t (see figure 3.3).

As already mentioned in section 3.2.1, real populations do not follow the constraints made to simplify the process. When the basic coalescent is used to model real populations, the population size (2N) in the haploid model does not represent the size of the real population. 2N is just the size of a Wright-Fisher population that best approximates the real population and is called the *effective population size* and is denoted by N_e. With these basic rules a genealogy for a set of sampled genes can be constructed.

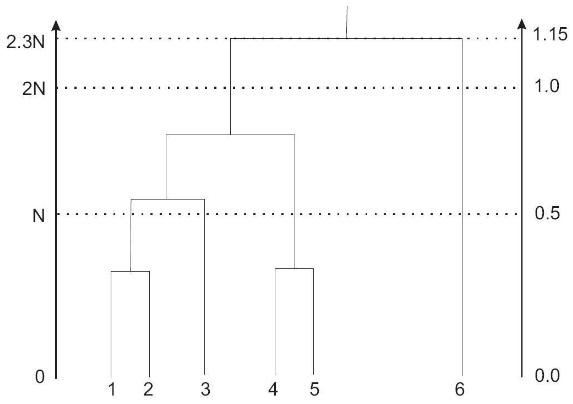


Figure 3.3: A genealogy with time measured in units of generations (left) and the corresponding time measured in units of 2N generations (right), whereas 2N is the average time for two genes to find their common ancestor.

3.2.3 Adding mutations to genealogies

Now that the relationships between genes, or in general sequences, in a population are constructed and, therefore, they share a common ancestry, mutations need to be added to the genealogy in order to cause changes in the DNA and to model real world data. Therefore, mutation models need to be considered. Historically the infinite alleles model [75] appeared first, followed by the infinite sites model [74], and the finite sites model by Jukes and Cantor [71]. Since the program *ms* uses an infinite sites model, this section will explain the basics behind this mutation model. The assumption of selective neutrality ensures that the process of adding mutations to the branches of the genealogy is independent from the transmission of genes from one generation to the next.

The infinite sites model can be interpreted as adding mutations to a very long string of DNA having a low mutation rate at each position. Therefore, each site, or position, mutates at most once which is the same as each mutation happens at a new position so that there are only two possible states per site (at least in most of the cases), an ancestral one and a derived (mutated) one [74]. The biological explanation is that the number of variable sites in a real sample is generally smaller than the number of sites which are identical in all sequences. All mutations are recoverable in contrast to the infinite alleles model because no back mutations, e.g. $A \to T \to A$ can occur. Hence, it is convenient to represent sequences as a string of “0” and “1”, not considering positions that are identical [47].

One fundamental fact is that mutations are superimposed on the branches of the coalescent tree, assuming this process follows a Poisson distribution of rate $t\theta/2$. θ is called the population mutation rate or scaled mutation rate and t is the length

	0.082	0.256	0.583	0.901	0.991
A	1	1	0	0	0
B	1	1	0	0	1
C	0	0	0	0	0
D	0	0	1	0	0
E	0	0	1	1	0

Table 3.1: Five Sequences are shown, each with five segregating ("polymorphic" or "variable") sites. The first row represents the position of each site (columns), relative to the length of the sequences on an interval from 0 to 1. Positions that are invariable are not displayed. Letters A-E (rows) represent the name of each sequence consisting of a string of "0" and "1", being the ancestral and mutant state.

of a branch. θ can be seen as the expected number of mutations that separate a sample of two sequences. 2N is the expected time for two sequences to coalesce and thus $\theta = 2N\mu$ mutations can be expected on each branch, with μ being the chance of a mutation occurring in an organism or gene in each generation (whether θ is defined by 2N or 4N only depends on the scaling factor of the actual coalescent implementation, the meaning stays the same). The following algorithm explains how mutations are actually tossed onto the genealogy after the genealogy has been simulated [4]:

1. Simulate the genealogy of n sequences according to the coalescent process with rate $\binom{k}{2}$ while there are k lineages
2. For each branch draw a number, M_t, from a Poisson distribution with intensity $t\theta/2$, where t is the length of the branch
3. For each branch, the times of the M_t mutation events are chosen randomly on the branch

Step 2 of this algorithm represents an important point, being one of the basic ideas that lead to the results of this work, namely that the longer the length of a branch, the more mutations can occur. The direction of the mutation process is inverse to the direction of generating the genealogy. Starting at the root of the coalescent tree moving forward in time the type of genes is modified as mutations are encountered.

Table 3.1 shows a small data set consisting of five different sequences A-E, containing five different segregating sites, each site with a different position. The principle of the infinite sites model is represented in figure 3.4. For instance sequence A has two segregating sites, since on the path back to the root of this tree-like sequence

---

[4] adopted from [54] p. 42-43

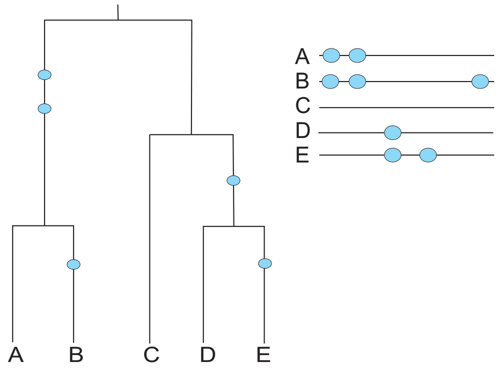


Figure 3.4: The data set of table 3.1 represented in a different way according to the infinite sites model. Each dot depicts a mutation at one position, resulting in the mutant state "1". Each position mutates at most once. The ancestral string at the root of this genealogy would be "00000". Letters A-E represent the name of each sequence.

representation only two mutations occur that can affect sequence A. Now that two models are available, the Wright-Fisher reproduction model and the infinite sites mutation model, these two can be combined. Since both models are stochastic, i.e. genealogies are treated as random and mutations are treated as random, if the whole process of coalescent (or "evolution") was repeated, the outcome of different runs would have different genealogies, which is very important to take into account when analyzing coalescent data.

Each gene that is going to be passed on to the next generation is subject to a mutation event with probability u. The chance of being copied without being changed occurs with probability 1-u. Hence, a position along the sequence is chosen randomly and the type of that position is changed from 0 to 1. Therefore, at this position the ancestral allele mutated to the derived allele.

3.3 Expanding the basic coalescent

The basic coalescent acts as a foundation for a wide field of more complex adaptations. Since real data is not as simple as required by the Wright-Fisher reproduction model (3.2.1), the process of coalescent needs to be extended. Constant population size, absence of selection, or random mating are restrictions that are hardly ever present in real world data. This section is supposed to give a short introduction of how to include simple deviations from these basic assumptions. However, since the prevalent results of this work and the newly developed statistic are mainly based on population size changes, emphasis will be on the inclusion of non-constant population size to the basic model.

3.3.1 The coalescent with changes in population size

It is well known that human populations vary in size over time (e.g. [2]). Such fluctuations can be caused by a variety of factors. A population bottleneck, an evolutionary event in which a significantly large number of individuals is either completely removed from the population or is not able to contribute to the process of reproduction anymore, can occur due to changes in the environment of a population, for instance a natural disaster or diseases like polio or measles. But also a slow change over time can happen as in the case of humans, starting at the time of the neolithic revolution 10,000 BC [49], with 6 million individuals and growing with constantly increasing growth rates per year up to over 6 billion individuals [13].

When considering such events the size of a population is not constant anymore but (simplified) a function of time. At any point t the population size N is N(t). Hence, equation 3.1 does not hold anymore, since the probability that two genes coalesce (find the same parent in the previous generation) is now $p(t)=1/(2N(t))$ because N(0) might differ from N(t). This simply means that in each generation the rate of coalescent, i.e. how many genes find the same parent in the previous generation, is totally dependent on the size of the population. The more individuals (i.e. genes) N in generation t, the smaller the probability $p=1/(2N(t))$, so the probability of a coalescent event decreases. The understanding of this fundamental concept is crucial for the rest of this work, since it acts as the basic idea that is used to detect population size changes in the history of a population. Thus, the smaller the population size, the more quickly the MRCA (see figure 3.2) is found.

To simulate genealogies under changing population sizes time is the crucial point. Still the modeling can be done by using the constant size coalescent process, but then time needs to be stretched or compressed. If p(t) is larger than p(0), i.e. N(t) is smaller than N(0) by e.g. a factor of 2, then time needs to be stretched locally by a factor of 2. As already mentioned in figure 3.3, 2N is the general time for two genes to coalesce and one unit in the continuous time coalescent model accounts for 2N generations in the discrete time coalescent model. The time scale, therefore, directly reflects the rate of coalescent and it can be shown that the coalescent model with variable population size converges to a coalescent process with a non-linear time scale [47]. Metaphorically speaking, stretching the time by a factor of 2 in this case implicates that the constant size model would need twice as much time to generate the same amount of coalescent events as the reduced population size model (because this model has an increased coalescence rate). Thus, the smaller or larger the population size is in generation t, the more or less *coalescent time* (time in units of 2N) passes. The genealogies will look the same in the basic constant size coalescent model and in the varying population size coalescent model, just the branch lengths are different. This idea is essential for the remainder of this work.

3.3.2 Genealogical effects of variable population size

The general effect of varying population size over time was briefly described in the previous subsection. However, for a complete understanding, especially for the 2 point spectrum method, a more detailed explanation of the way how coalescent times and genealogies as a whole are affected is of utmost importance. Following the previously mentioned assumptions that need to be made when taking non-constant size models into account, another condition is that N(t) is given in terms of continuous time (in units of 2N generations) and that N(t) does not need to be an integer value.
Let

$$\Lambda(t) = \int_0^t \frac{1}{\lambda(u)}, \tag{3.2}$$

where $\lambda(t) = N(t)/N$, the relative size of N(t) to N. $\Lambda(t)$ is then the accumulated coalescent rate over time measured relative to the rate at time t=0 ($\lambda(0) = 1$). Furthermore $\Lambda(\infty) = \infty$ in order to ensure that a sample of genes always finds a MRCA. Also let T_2,...,T_n be the waiting times while there are 2,...,n ancestors of the sample and let $V_k = T_n$+...+T_k be the accumulated waiting times from there are n genes until there are k-1 ancestors (see also [47]). So the distribution of times T_k given $V_{k+1} = v_{k+1}$ is

$$P(T_k > t|V_{k+1} = v_{k+1}) = exp\left\{-\binom{k}{2}(\Lambda(t + v_{k+1}) - \Lambda(v_{k+1}))\right\}, \tag{3.3}$$

and v_{n+1}=0. In order to calculate the time to the next event T_k it is important to keep track when the last coalescent event occurred (V_{k+1}). The times of earlier coalescent events are not required, which is a result of of the Markov property. Now times need to be distinguished between the basic constant size coalescent (T^*_k) and the variable size coalescent (T_k).

Since this concept will later be picked up again, a general algorithm is given (adopted from [54] p. 98):

1. Simulate T^*_2,...,T^*_n according to the basic coalescent (see previous sections), where T^*_k is exponentially distributed with parameter $\binom{k}{2}$. Simulated values are denoted by t^*_k
2. Solve $\Lambda(t_k + v_{k+1}) - \Lambda(v_{k+1}) = t^*_k$ for v_k, k = 2,...,n and $v_{n+1} = 0$
3. The values $t_k = v_k - v_{k+1}$ are an outcome of the process, T_2,...,T_n, described in equation 3.3

This algorithm can be used to generate coalescent genealogies with arbitrary N(t). Simply speaking, simulate times according to the basic coalescent until the start of a size change and from there on, until the next size change happens, simulate times from the basic coalescent with all times shortened or extended by a factor of f (with N1 the time before and after the event and N2=N1*f the time during the event). At the end of the event simulate again according to the basic coalescent. This will be important for calculating the tuples of inter-coalescence times further on in this study. The population size variation analyzed in the current work exclusively focuses on instantaneous size changes, turning the function N(t) into a piecewise constant function.

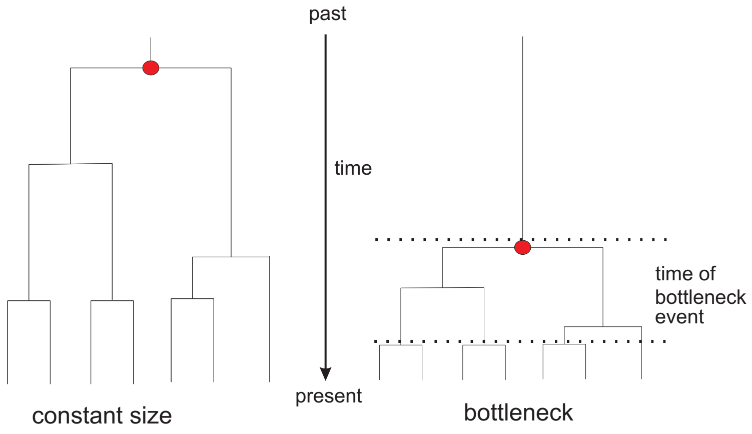


Figure 3.5: Shown are 2 coalescent trees, one generated under a constant size model (left) and one under a bottleneck model (right). The red dots represent the most recent common ancestors of each of the set of sequences. Time is running from past to present and the dotted lines indicate the time at which the bottleneck occurred.

Figure 3.5 shows the comparison of a genealogy under a constant size model and under a model that incorporates a population bottleneck. As can be seen, if the population size was reduced in the history of a population, the coalescent rates are increased during this time. Hence, more lineages than expected under a constant size model will join and more genes will, therefore, find their common ancestor. In the case of figure 3.5 all lineages coalesce during the bottleneck and the MRCA of the entire sample is located within this time. At the time where two lineages coalesce they are completely identical. The more recently this happens, the less time there is for mutation and recombination to occur and to consequently change the sequences until the present day. The effect of such a bottleneck is strongly dependent on the strength and how long ago it occurred.

Another very important way of size variation is exponential growth, the simplest and most natural way of steady population growth, which can also be incorporated into the basic coalescent. For the sake of completeness a brief summary of the effects of exponential growth follows. Figure 3.6 shows the same genealogy generated under an exponential growth model (left) and under a constant size model (right). As indicated by the red points the MRCA is much more recent in the constant size model. The stronger the rate of growth, the more star shaped the trees tend to look and the sequences are expected to be equally diverged from each other [118], and because of more time (on average) for mutation and recombination events to occur, the sequences are expected to be more diverged as compared to a constant size model. An interesting fact that can be seen for exponentially growing populations is an increase in the amount of low frequency polymorphisms, especially singletons.

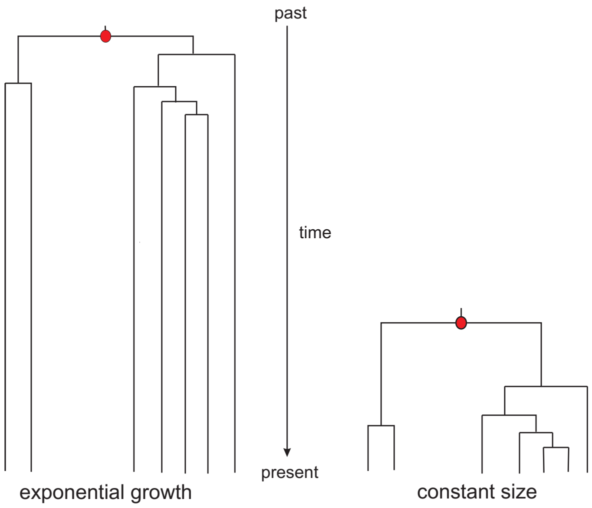


Figure 3.6: Shown are two genealogies, the one on the left generated with exponential growth and the one on the right generated with a constant size model. The red points indicate the MRCA of each set of sequences.

3.3.3 The coalescent with recombination

The models and genealogies described so far have one thing in common: they do not include recombination. Since the math behind the coalescent with recombination is complex and a detailed understanding of this would not contribute to a better understanding of the rest of this work, the details are just explained as necessary. The structure needed to describe the relationship of a set of recombining sequences turns into a complicated graph. Hence, calculations of tree properties, e.g. tree

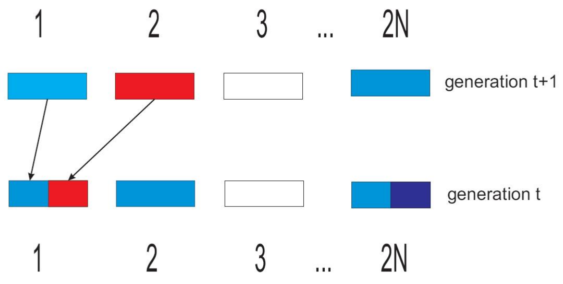


Figure 3.7: Shown is the haploid (2N genes) Wright-Fisher model with recombination. As already mentioned in the text, the genetic material of each individual from generation t is transmitted by two parents from generation t+1 by recombination. If the recombination rate is low, there is a higher chance for genes not to be affected by recombination and being a direct copy of one of the parent genes.

height, branch lengths, etc., are not as simple as for the basic coalescent model, actually even impossible to derive in some complex cases.

Recombination can occur by quite different mechanisms depending on whether you look at eucaryotes, bacteria, or viruses. In humans, i.e. eucaryotes, recombination is realized by sexual reproduction. In the basic Wright-Fisher model of reproduction (see section 3.2.1) each gene in generation t finds exactly *one* parent in generation t+1. With recombination the genetic material that a gene, or more generally an individual, is made of, can be transmitted from two parent individuals (see figure 3.7). That recombination can be included into the framework of coalescent was first shown by Hudson in [61]. The disadvantage that needs to be accepted is that there is no longer just one single underlying tree relating a set of sampled genes. Linked sites can have different genealogical trees, so the lineage of a segment of the sequence splits into two. Single points on the sequence of each gene (or individual) are transmitted by only one parent from the previous generation. In this case the relationship of the different sequences for each single point can still be represented by a single tree, the so called *local tree*. Hence, the genealogy for the entire sequence can be characterized by a collection of local trees, one for each point. Thus, different parts of the sequence can have a different, even younger MRCA than the MRCA of the entire sample.

Coalescent and recombination are two rival events, with coalescent merging two lineages and with recombination splitting two lineages as time running backwards (see figure 3.8). Like the scaled mutation rate θ (see section 3.2.3), the scaled recombination rate is defined as $\rho = 4Nr$, with r being the probability of a recombination event in a sequence in which case a point is chosen uniformly along the paternal and maternal sequence and they recombine in that point. The scaled time until the first recombination event is exponentially distributed with a rate of $\rho/2$ for the

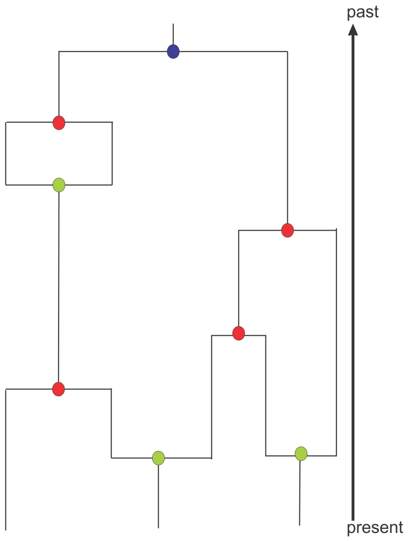


Figure 3.8: Shown is a very basic example of how coalescent and recombination can be combined to build up the underlying graph structure. Red dots indicate coalescent events and green dots indicate recombination events. The dark blue dot represents the MRCA of the entire sample.

continuous time approximation if N is large. Thus, the more time passes, i.e. the deeper the genealogy of a particular region is, the more time there is for multiple recombination events to occur. This fact, among others, is one important insight that has led to some of the results of this work. The effect of recombination is, as well as for mutation, that the similarity of the sequences is decreased with each event, resulting in a sample with more variation.

The combined effect of coalescent and recombination still leads to a MRCA (called *grand-most recent common ancestor GMRCA* if recombination implicates a graph like structure), since the coalescent intensity is proportional to the square of the number of ancestors and the recombination intensity is linear to the number of ancestors.

Chapter 4

Importance of population size changes

> *“Wonder is the seed of knowledge.”*
>
> *Francis Bacon (1560 - 1620)*

The present chapter explains why inferring population size changes is of high importance and briefly introduces commonly used methods in this field. Its special focus is on approaches based on the site frequency spectrum (SFS) or based on linkage disequilibrium (LD). LD and SFS are the two main principles that have been used to develop the two approaches described in this work. This section is loosely structured according to a recent publication that gives a comprehensive introduction into the field of population size inference approaches [40] and to [113, 114].

4.1 Overview

As already mentioned the patterns of genomic diversity of a population or species are shaped by a complex interplay of a large variety of demographic events like population size changes, the exchange of genetic material between different populations, and population splits or divergence. Trying to infer the demographic history is important for understanding evolutionary trajectories of populations. Reconstructing temporal population size changes and understanding their causes can give insights into how populations responded to historical events such as climate changes, glacial cycles, or human driven events. For example, in the field of conservation biology, gaining insights into how the size of a population is going to evolve might assist in detecting past or ongoing bottlenecks. This seems crucial, since the reduction in size also correlates with a reduction in genetic diversity which in turn might negatively affect the evolutionary potential of a species. Furthermore, the effects of bottlenecks and founder effects (the loss of genetic variation that occurs during the settlement of a new population by a very small number of individuals from a larger population) can increase the presence of deleterious alleles. This effect has been shown to have occurred in humans during the Out-of- Africa bottleneck [92].

What we see in today's genetic data is a record of the past and, therefore, multiple approaches to infer population size changes from genetic information have been designed. Probably one of the most well studied events in human history is the already mentioned Out-of-Africa bottleneck (e.g. see [51]) with the human expansion that began approximately 45 kya to 60 kya [56]. Often new methods were particularly designed to investigate such specific events in species as the fruit fly *Drosophila melanogaster* (e.g. [134]) or the thale cress *Arabidopsis thaliana* (e.g. [36]). Especially the fruit fly shares a similar demographic history with humans, as it also has an African origin and over time populated new areas outside the continent [88].

4.1.1 Classical population size inference approaches

The site frequency spectrum (SFS) is the distribution of allele frequencies at a large number of variant sites. Given the ancestral allele is known, it is then possible to recover the SFS that contains the probabilities that an allele is carried by i individuals p_i=p(i|n), i=1...(n-1) for a sample of n sequences. In cases where the allele frequency is identified experimentally by counting the two alternative alleles within this sample of n sequences it is not clear which of the two is the mutant allele. In such situations it is common to work with the less frequent (minor) allele, which is also called *folded spectrum* [96]. Various statistical tests have been suggested to calculate one dimensional summaries of the SFS and the current section will elaborate on commonly used methods. In section 3.3.1 effects of population size changes on the coalescent genealogy were explained and algorithms to incorporate such changes into methods of parameter inference were given. Additionally I will now give a brief summary of how such events can affect the SFS and lead to signatures that can, therefore, be detected. One of the most common effects of a recent population bottleneck is the reduction of low-frequency polymorphisms [103]. For a recent moderate bottleneck the SFS would show a slight deficit of singletons (mutations that are only observed once), a more significant reduction of other low frequency variants, and an increase of high frequency variants. The more ancient a bottleneck is, the less of a signal of singleton reduction, but the reduction of other low frequency variants remains long after the end of the bottleneck. Strong bottlenecks imply that only few or even only one lineage can escape and, as a result, genealogies tend to have a star-like shape which in turn creates an excess of singletons immediately after the bottleneck event. These deviations from a standard neutral SFS (i.e. from a population following a standard Wright Fisher population) can be detected by many statistics. However, single summary statistics are often not sufficient to capture the specific patterns of deviations from the basic coalescent.

Classical tests can be classified into three different categories based on the information they use. Class I tests are based on the frequency spectrum of mutations, class II tests on the haplotype distribution, and class III tests on the distribution of pairwise differences. Class I statistics often make use of the differences between different estimators of the population mutation rate ($\theta = 4N_e\mu$), thereby detecting distortions in allele frequency spectra. Commonly applied tests are Tajima's D [129], Fu and Li's D, F, D* and F* [38], Fay and Wu's H [33], and R_2 [114]. Class II tests use information from the haplotype distribution and are measures of linkage disequilibrium. Hence, they are not completely independent of mutation frequencies and thus not independent of class I statistics. Commonly applied class II tests and statistics are Fu's F_s [37], Nei's *Dh* [102], EHH (extended haplotype homozygosity) [117], Wall's B [138], Kelly's Z_{nS} [73], Roza's Y_A, the haplotype partition test (HP) [63], the haplotype number statistic K [126], and the haplotype diversity statistic H [27]. Class III statistics make use of fact that population size changes can leave a particular signature in the distribution of pairwise mismatches. Examples are the raggedness statistic *rg* [52], the MAE between observed and expected mismatch distribution [115], and the *ku* test [114].

In [114] the authors describe a number of statistical tests (class I, II and III) for detecting population growth and compare them with other available tests in the literature. Soriano et al. study several statistics (class I and II) to detect demographic expansions, contractions and bottlenecks [113]. In [118] authors investigate the properties of test statistics of neutrality (class I) under population growth and Cornuet et al. investigate statistics that are sensitive to population bottlenecks, trying to detect the signature of a transient heterozygosity excess in bottleneck populations [24]. A test for the detection of recent population bottlenecks was proposed in [94], designed to detect the allele frequency distortion after a bottleneck. In [26] departures from a simple model, caused by population bottlenecks and hitchhiking effects, are investigated. The authors evaluate the power of different tests (class I and II) when faced with "severe" and "moderate" bottlenecks as a function of age and strength. Summarizing the mentioned studies, class I and II statistics are powerful for detecting population growth and bottlenecks. Class III statistics often perform poorly to accept or reject a null hypothesis of growth or contraction as they depend largely on one highly stochastic coalescent event leading to the root of the genealogy. In particular, moderately old severe bottlenecks can best be detected by statistics relying on the frequency spectrum of mutations. On the other hand, haplotype tests are mainly useful to detect recent and more moderate bottlenecks. Such conclusions massively depend on the details of demographic events and on factors like mutation and recombination rate, population structure, gene flow, etc. For example, class II statistics are strongly affected by recombination. Over- or underestimating the recombination rate can decrease the power of such statistics and result in biased or completely wrong conclusions.

These classical approaches applied such test statistics to reject null hypotheses of population growth, bottlenecks, etc. Besides those, there are alternative approaches, e.g. maximum-likelihood based or MCMC methods to study population size changes [6, 7, 39, 85], or different Bayesian approaches [5]. Furthermore, summary statistics like the ones mentioned above can be included in an Approximate Bayesian Computation framework in order to infer the timing, duration, and severity of bottlenecks as in [134] (in this case for Drosophila melanogaster). The authors consider the variance of nucleotide diversity ($\hat{\theta}_\pi$) per locus, the number of haplotypes K, and a standardized summary of the site frequency spectrum H as informative to detect changes in population size. In [15] the number of segregating (polymorphic) sites and the nucleotide diversity are assumed to be informative of a population bottleneck. Both statistics are included in an Approximate Bayesian Computation framework (in this case for ancient DNA).

Despite using one dimensional SFS summary statistics, considering the entire SFS can also easily be incorporated into likelihood frameworks. A variety of approaches was suggested like analytical solutions of diffusion equations and fitting a simulated SFS gradually to approach the observed SFS, by searching the constrained parameter space of given demographic models (e.g. see [36, 44, 51]).

Further approaches are based on IBD segments and runs of homozygosity. As for the AF-IBD/S statistics (chapter 5), the idea for such methods is that homologous long segments that are IBD from a common ancestor provide clues about past population sizes. Model based approaches showed that population growth and subdivision can strongly affect the expected length of an IBD tract [16]. Long runs of homozygosity contain insights into (among others) recent ancestry such as effective population size for recent time periods. These segments can be calculated for two homologous chromosomes within the same individual (e.g. [81]).

4.1.2 Recent genome-wide approaches

Approaches shown so far all depend on parametric demographic models. Hence, an a-priori demographic model needs to be assumed and inference can then only be based on this model. However, since population history is much more complex than those described by parametric models, semi- and non-parametric methods have been developed. The present work contains both kinds of approaches and provides a comparison between the two principles.

One of the first and later commonly used non-parametric methods developed to estimate historical patterns of population size from a genealogy without the need for too many a priori restrictions on possible demographic models was the *skyline* *plot* framework [112] from 2000. Later various methodological extensions have been suggested, generating a whole family of skyline plot methods (see [28, 55, 100, 107, 112, 125]) and see [58] for a comprehensive review of those. All of these methods are based on the coalescent theory trying to reconstruct and estimate the coalescent genealogies. As previously described in section 3.3.1, population size changes affect the rate of coalescent, i.e. the smaller the effective population size, the more coalescent events take place on average. This principle is the key of all skyline plot methods, since the rate of coalescent events (or the frequency of coalescent events in a given time period) gives insights into the population size at a certain time point. The first step includes estimating the genealogy, which can be done using standard phylogenetic methods. The second step is the reconstruction of population history based on the genealogy taking into account the relationship between population size and expected length of the coalescent intervals, producing a piecewise reconstruction of the demographic history. This whole process involves considerable uncertainty, often referred to as the *coalescent error*. The coalescent is a stochastic process and a single genealogy, given a specific demographic model, is only a single realization of this process. These coalescent errors increase towards the root of each genealogy, since population size here is only estimated from the last remaining lineages (i.e. only two lineages in the last coalescent interval). The first published method is the classic skyline plot [112] which is known to produce noisy reconstructions of the demographic history owing to the number of free parameters and short branches in the genealogy. For each coalescent interval in the genealogy a separate population size is estimated and the genealogy is assumed to be known without error. The generalized skyline plot [125] tries to circumvent the noise produced by short coalescent intervals by grouping them with their neighbors if they are below a certain length. Finding the optimal threshold length involves a good balance between the reduction of noise and the information provided by the structure of the coalescent tree. Again, the genealogy is assumed to be known without error. The Bayesian multiple-change-point method was developed by Opgen-Rhein et al. in 2005 [107]. Purpose of this reversible jump Markov Chain Monte Carlo approach was to smooth the piecewise demographic function of the generalized skyline plot and to estimate the coalescent error. Sudden and drastic population size changes are assumed to be unlikely which is implemented as a spline, a piecewise function comprising a number of polynomial curves. The 95% credible interval for the population size at each time point can be used to estimate the coalescent error. According to its Bayesian nature priors can be given for each parameter and the method tends to produce relatively smooth plots. The Bayesian skyline plot [28] co-estimates the genealogy, demographic history, and substitution model parameters in a single analysis. The credibility intervals also represent the combined phylogenetic and coalescent error. Since this method roots in the generalized skyline plot, the number of groups that combine neighboring intervals need to be chosen. However, this decision needs to be made a priori and choosing extreme numbers can increase error and be problematic for analyses of uninformative data sets. In the Bayesian skyride method [100] differences between population sizes of successive coalescent intervals are penalized using a smoothing prior. Until then all methods were only capable of analyzing individual genealogies, i.e. single loci. One drawback of this approach is the coalescent error that is associated with estimates from a single locus because each genealogy only represents a single realization of the stochastic coalescent process. Therefore, the extended Bayesian skyline plot [55] allows the analysis of multiple independent unlinked loci, reducing the uncertainty of the coalescent, which can lead to an improved reliability of demographic inference and reduction in estimation error. For a simple comparison of the mentioned methods see figure 4.1.

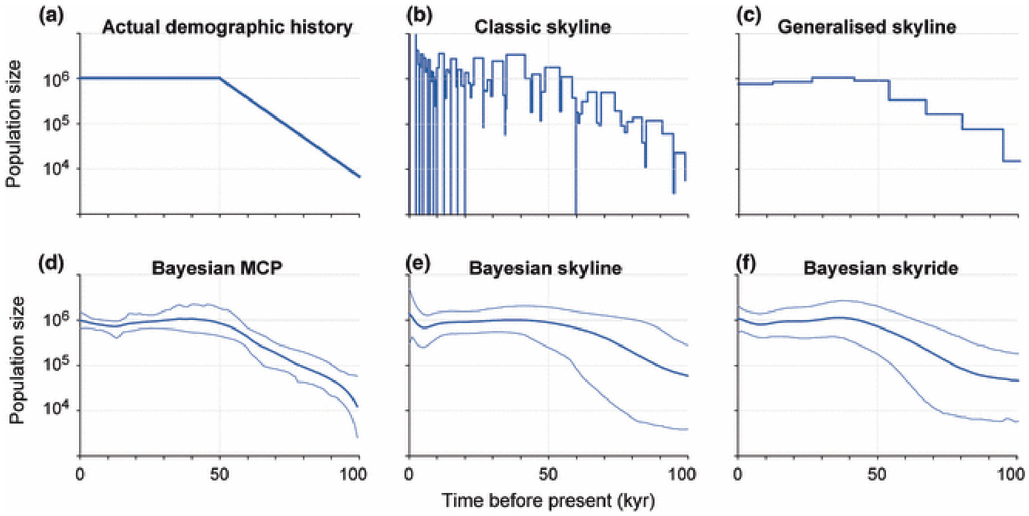


Figure 4.1: Performance of different skyline plot methods for a simulated data set. The actual demographic history is shown in panel A, with the faint vertical line at 50 kya indicating a change-point in the demographic function. The remaining panels show the reconstructions of demographic history by five skyline plot methods. Note the logarithmic scale on the y-axis. Figure was adopted from [58]. Further details of the simulation model are given in their Appendix.

It is known that the distribution of time since the most recent common ancestor (TMRCA) between two alleles in an individual provides information about the history of change on population size over time. Previous studies have reconstructed the TMRCA distribution, analyzing large samples of individuals at non-recombining loci like mtDNA [3]. Taking only a single locus into account might not provide a sufficient resolution of inferences and power decreases with moving back in time. On the other hand, a diploid genome sequence contains a huge amount of independent loci, each with its own TMRCA between the two alleles carried by a single individual. Therefore, it should be possible to use this information to make demographic inferences. Based on this idea Li and Durbin recently developed a coalescent-based hidden Markov model for a pair of chromosomes (or one diploid individual) to estimate past population size changes. This method is called pairwise sequentially Markovian coalescent (PSMC) [87] and the approach turned out to be useful and efficient, but its accuracy in the very recent past is disturbed by the fact, that only very few coalescent events occur in that time period. This is because of the small sample size. *PSMC* takes linkage information into account and efficiently models recombination between sites. It, therefore, uses the sequentially Markov coalescent [98] for a pair of sequences to estimate an arbitrary piecewise constant population size history. The HMM's hidden states at a given position represent the coalescent time of the two lineages at that position (TMRCA), whereas the observed state represents the observed genotype (homozygous or heterozygous). Moving along the sequences, coalescence time for the two lineages varies as a result of recombination. Each region that can be described by a single genealogy, i.e. a region between two recombination events, has a single TMRCA (see figure 4.2). These coalescent times can be inferred by taking a mutation rate and sequence diversity in each region into account. Transitions from one region to another can be detected by changes in the spatial distribution of heterozygous and homozygous sites and TMRCA are informative about past population sizes. The demonstrated accuracy for simulated scenarios shows that this approach can produce reliable population size estimates even from a single diploid genome. However, as mentioned before, for very recent time periods this method can not be recommended due to the fact that the small number of samples causes only few coalescence events in that period of time. Metaphorically speaking, the genetic variation information content from a pair of sequences is not sufficient to infer very recent demographic events.

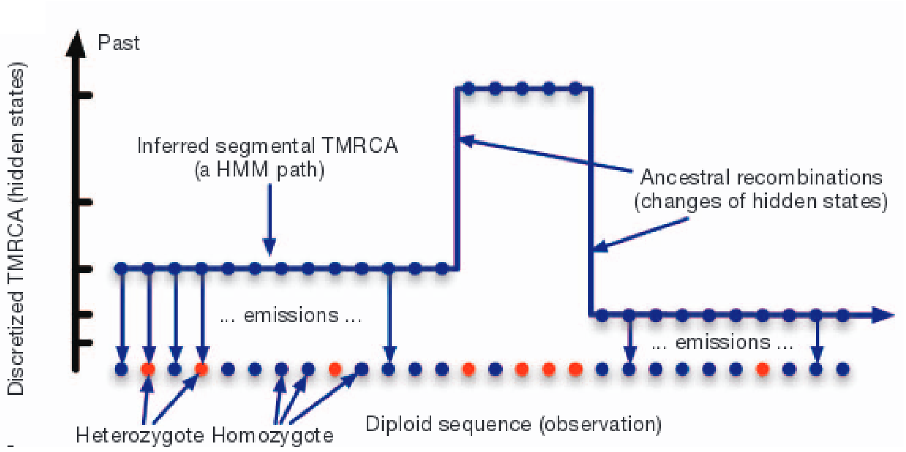


Figure 4.2: Illustration of the PSMC model. The PSMC infers the local time to the most recent common ancestor (TMRCA) on the basis of the local density of heterozygotes, using a hidden Markov model in which the observation is a diploid sequence, the hidden states are discretized TMRCA and the transitions represent ancestral recombination events. Figure was adopted from [87] Figure 1.

Following the principles of PSMC, Sheehan and Harris developed an alternative method that incorporates multiple sequences while retaining the key generality of the previous approach [121]. As Sheehan et al. state, increasing the number of sequences also increases the number of coalescence events during the recent past, thereby enhancing the resolution to infer size changes in those time periods. The main problem when trying to incorporate multiple sequences in this inference framework is the explosion in the state space, i.e. the number of coalescent trees grows enormously with the number of leaves (sequences). This method is called demographic inference using composite approximate likelihood *diCal* and the computational complexity depends quadratically on the number of sequences and can easily be parallelized, thereby achieving a computationally efficient approach. For a direct comparison between PSMC and diCal see figure 4.3. The aim of this figure is to show the performance differences when using diCal and PSMC to infer a simulated demographic history. As already mentioned, the power of PSMC in recent time periods decreases, clearly shown by the deviation of the green line from the true simulated red line, whereas diCal shows a perfect match for the recent constant size period of the history. Lines shown here are the average of ten data sets.

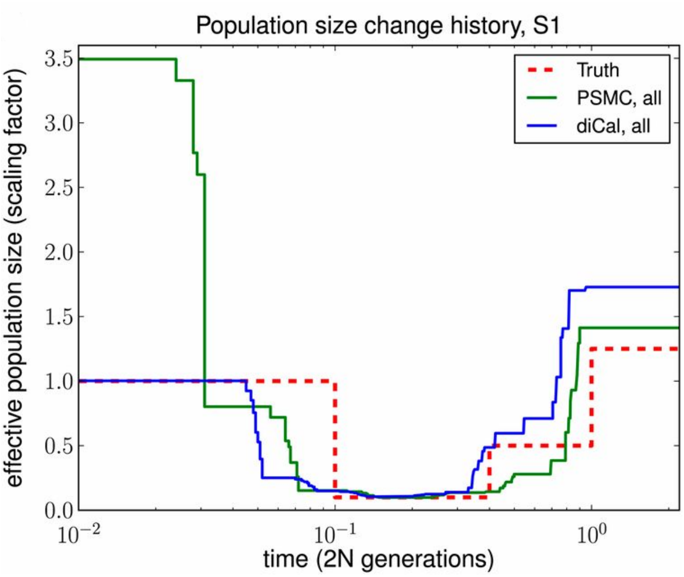


Figure 4.3: Results of PSMC and diCal for simulated data sets under a specific demographic history with sample size n=10 and four alleles (A,C,G and T). PSMC significantly overestimates the most recent population size, whereas diCal obtains good estimates up until the very ancient past. Results are shown as an average over ten data sets (runs). Figure was adopted from [121] Figure 5.

4.1.3 Perspectives

The current section tried to give an introduction into the field of population size inference approaches and enumerated various methods from single statistics to comprehensive likelihood inference approaches. Simple test statistics like Tajima‘s D, etc., provide valuable information and proved to be sensitive to deviations from a standard constant size demographic model. They have been applied in a wide range of diverse studies trying to gain insights into the demographic past of populations. However, single number statistics can only capture specific parts of the underlying data and often the effect of certain demographic events is quite specific on different statistics. Therefore, genome-wide approaches that include the calculation of a likelihood function, made up of several aspects of the data, more and more develop into powerful methods of demographic parameter estimation. Detecting departures from equilibrium conditions was usually done by assuming that populations can be approximated by a Wright- Fisher model (i.e. no selective pressure, assuming panmixia, demographic stationarity, etc.). However, natural populations are often part of spatial networks and the assumption of completely isolated populations is as well unrealistic, since they are often interconnected by gene flow. Hence, so called confounding factors can lead to false positive results during demographic parameter estimations when not accounted for. Biasing factors can also influence and skew the estimates of demographic parameters in wrong directions. As an example, in [18] Chikhi et al. simulated stationary populations not subject to any population size change. Varying the levels of gene flow between populations, the mutation rate, and sampling scheme, they showed that those factors were able to create false bottleneck signals (detected with an MCMC method). Confounding and biasing factors can therefore be classified into genetic factors like mutation and recombination rate, and demographic factors like hidden population structure and the sampling scheme (i.e. how multiple individuals are sampled from multiple populations). False assumptions of the underlying mutation and recombination rate can massively affect the outcome of inference methods. If not accounted for these factors, the quality of the results can be significantly decreased. For example, class II statistics are sensitive to recombination events, since recombination can break down the haplotypes into smaller chunks creating false signals if not modeled correctly. Often, such methods are only applied together with a known recombination map of the genetic regions of interest. Pooling samples from different populations can produce a shift in the SFS toward low-frequency polymorphisms which in turn is a signature of population expansion [111] and, as also shown in Chikhi et al., population structure can produce a similar SFS as a bottleneck event. Class I statistics can, therefore, heavily be affected by factors influencing the site frequency spectrum of populations. Despite the briefly mentioned factors, many more demographic and genetic influences like recent admixture or ascertainment bias (see chapter 5), etc., can negatively affect the results. Trying to take these things into account requires a complex framework and often includes modeling a variety of different peripheral parameters. The more parameters are involved, the more sensitive to specific parts of the demographic model the applied summary statistics need to be. Hence, the use of only a single summary statistic is, therefore, not sufficient enough and a specific interplay of data summaries and likelihood-like approaches is needed. But even if having controlled for all these caveats, the interpretation of potential signals is still challenging, since distinguishing between multiple potential explanations can be difficult and a variety of scenarios can have similar outcomes. Sophisticated non-parametric methods (e.g. [87, 121]) that allow for a high degree of flexibility, facilitate the fitting of demographic processes, and at the same time become more and more convenient to use. However, since these methods recently emerged, little is known about the effects of confounding and biasing factors on these approaches. Interpreting such demographic estimates of different methods must not be underestimated and is specific from case to case.

The current study introduces two different methods, a parametric and a non-parametric approach. Both methods represent a new way of summarizing information from underlying data while being sensitive to past population size changes.

Chapter 5

The AF-IBD/ AF-IBS method

"Everything must be made as simple as possible. But not simpler."

Albert Einstein (1879 - 1955)

As already stated, the overall goal of this thesis is to analyze genome-wide data to infer demographic parameters. Two distinct methods were developed for this purpose and the following chapter will therefore give a detailed description of the AF-IBD/IBS method. As mentioned, it can be read as a complete study in its own, providing an introduction, methodology, results and discussion section. Most of the current chapter is adopted from [132] and introduces a new set of summary statistics that are sensitive to past population size changes. These statistics are applied within a framework of Approximate Bayesian Computation to infer parameters of interest from underlying demographic models of different population size histories.

5.1 Introduction

As I already introduced in chapter 2.3.1, large parts of genetic data are from the recombinant autosomal genome, especially in the form of SNPs (see [23, 89, 93]). As previously explained, a lot of the existing studies infer the demographic parameters of interest by examining the allele frequency spectra, which are first corrected for ascertainment bias and later fit to the best demographic models using maximum likelihood computation. Since likelihoods of SFS can be numerically derived or approximated by simple simulations, such methods are computationally efficient. However, such methods based on the SFS are often sensitive to different sources of ascertainment bias and are usually applied under highly simplified demographic models. Another class of methods make use of summary statistics that capture information from different aspects of the data and evaluate how well they fit different demographic scenarios, which often involves the simulation of large amounts of genome-scale SNP data, and therefore, are highly computationally intensive. Another limiting fact is their mutual dependency and diverse sensitivities toward simulation assumptions make it difficult to evaluate the inference accuracy. Improvements in simulation efficiency and novel statistics systematically designed for demographic inference without making too strong assumptions are much needed (e.g. PSCMC, diCal etc.). Methods based on haplotype or linkage disequilibrium (LD) patterns should in theory be less affected by ascertainment bias ([21]). LD is a property of a genetic region, whereas ascertainment bias influences particular SNPs. Effective population sizes have recently been estimated from LD ([57, 127]). However, LD based statistics often suffer from limited resolution as either N_e is estimated as an average over long periods of time or the models studied are too simplistic. In [124] the authors reported that measurements of intra-allelic variability can be used to test neutrality and to infer population growth. Intra-allelic LD might also provide information for inferring more complex demography. I show that the statistics suggested here are informative about ancient population size trajectories and can be used in the framework of ABC to accurately estimate demographic parameters from simulated data. Finally I applied the ABC-based method to genome-wide SNP data for the Yoruba and French population from the CEPH-HGDP panel [89].

5.2 AF-IBD & AF-IBS

In this section I propose two statistics to infer ancient population size changes under neutrality. It is known that the intra-allelic variability and the allele frequency are two different measurements of allele ages, with the former revealing age at the absolute time scale, for example, in generations [97], and the latter at the rescaled coalescent time scale [123]. In [124] Slatkin and Bertorelle proposed, that the contrast of these two measurements can be used to test neutrality or to make inferences about population growth. All times in the current chapter are given in units of generations, in order to be able to directly compare the results with previous studies. Therefore, times in generations need to be divided by 4*10,000 in order to obtain times in units of $4N_e$ for simulation in ms. Nordborg and Tavaré [106] suggested that the intra-allelic LD can be informative about different aspects of demography, such as ancient population size and population structure. We propose that the intra-allelic LD measurement, when conditioned on allele frequency, may indeed be very informative about complex demographic trajectories. This is because when allele age in absolute time scale is compared with the age in coalescent scale, their ratio actually measures the N_e in each time interval (see 8.1.1 for detailed discussion).

The two statistics proposed here are both related to the haplotype sharing for a given derived mutation. Studies have investigated the extension of the ancestral (identical) haplotypes from a derived mutation, and its use in disease/quantitative trait locus (QTL) mapping and neutrality tests [64, 97, 122, 123]. Our statistics are similarly constructed. The first statistic is the extended length of identity by descent (IBD) conditioned on derived allele frequency (AF-IBD). Here the literal meaning of IBD is taken, which is the identity of sequences that descend from a single ancestral sequence, without any change in status from either mutation or recombination. IBD quantities usually have to be indirectly estimated, as tracts of IBD cannot be directly observed [97, 127]. However, to directly study how AF-IBD varies under different demographic scenarios, I start by assuming that IBD can be directly observed, and I later relax this assumption.

Assume that the genome is continuous, and all recombination and mutation events can be detected and exactly positioned. For any variant s of derived allele frequency j in a sample of n haplotypes (2≤j≤n-1), denoted as $l_{n,j,s}$, the length of the identical haplotype extending from s to either side until the first detectable event (mutation or recombination) occurs (see figure 5.1A and B). The AF-IBD for allele frequency j is then defined as the expectation over all variants of frequency j:AF-IBD$_{n,j}$=E(l$_{n,j,s}$). To study empirical sequence or SNP data, the statistic AF-IBS is proposed, similarly defined as AF-IBD: for a sample of n sequences, for each site s with derived allele frequency j (2≤j≤n-1), the distance up to which the carrier chromosomes are identical by state (IBS) is calculated in either direction, that is, up to one site before the first breakpoint, here denoted as x$_{n,j,s}$ (see figure 5.1C). The maximum distance x$_{n,j,s}$ is limited to 500 kbp in the simulations and any distance larger than 500 kbp in either empirical or simulated data is taken as 500 kbp. The AF-IBS of allele frequency j is then taken as the average of x$_{n,j,s}$ over all sites of allele frequency j:AF-IBS$_{n,j}$=Mean($x_{n,j,s}$).

The underlying idea is not new, i.e. that at the root of a genealogy the sequences are identical. Metaphorically speaking, adding mutations to a coalescent tree is a process proceeding from the root to the tips and the root has the ancestral state, i.e. a string of zeros. The more time there is for mutation and recombination to change the state of a position from “0” to “1”, the more the sequences, related by the underlying tree, will differ from each other. In the absence of mutation and recombination, all genes would be identical, starting from the root to the tips of the tree. Furthermore, the more ancient a mutation or recombination event is, the more sequences are affected by it, see figure 3.4 for details. In section 3.3.1 the effects of changes in population size are explained.

The distances calculated with the above algorithm should be, at least on average, larger for smaller derived allele frequencies. This is rather intuitive, since the probability that a set of haplotypes is identical at the same position, decreases the more haplotypes are considered at once. The main concepts that are expected are given in table 5.1.

I first study the properties of AF-IBD under different demographic scenarios, and then the performance of AF-IBD in demographic parameter estimation, using an ABC approach on simulated data, is examined. I then analyze the relationship be-

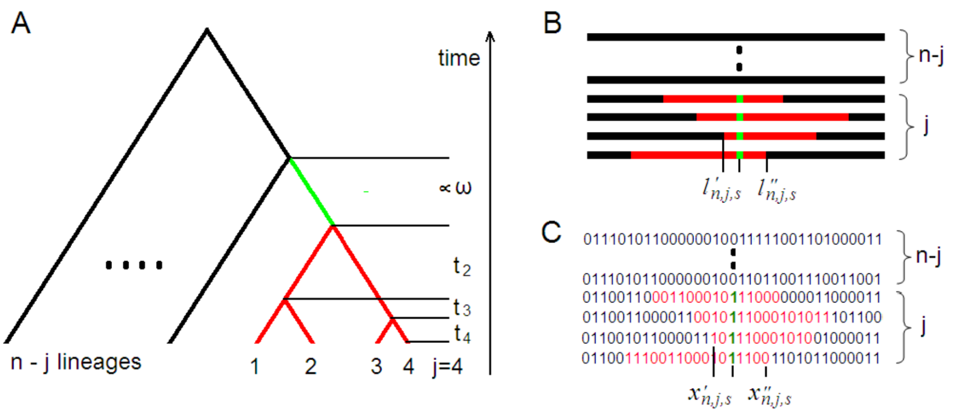

A
time
∝ω
t_2
t_3
t_4
n - j lineages
1
2
3
4
j=4
B
n-j
j
$l'_{n,j,s}$
$l''_{n,j,s}$
C
0111010110000000100111111001101000011
011101011000000010011011001110011001
01100110001100010111000000011000011
011001100001100101110001010111101100
01100101100001110111000101001000011
01100111001100010111001101011000011
n-j
j
$x'_{n,j,s}$
$x''_{n,j,s}$

Figure 5.1: A) The coalescent of n individuals, where j (here j=4) lineages from a subtree J are shown, colored red, before joining the other lineages by a root edge colored in green. The total length of subtree J, $T_{n,j}$ is the sum of the red edges measured in generations. B) The extension of the ancestral haplotypes in red is shown for multiple sequences from a core mutation of frequency of j=4 (shown in green). Mutation and recombination are taken as equivalent events that terminate the extension of the original ancestral haplotype. The ancestral shared haplotype is, therefore, the overlapping red segment that ends at the first event among all the sequences. The length of this segment is taken as a measurement of $l_{n,j,s}$ which when averaged over all sites of frequency j defines AF-IBD for j. C) As the counterpart of $l_{n,j,s}$ in empirical sequence/polymorphism data, $x_{n,j,s}$ is taken as the length of the shared haplotype extending from the core mutation up to the first observed site that varies among the j haplotypes. $x_{n,j,s}$ averaged over all sites of frequency j gives the estimation of AF-IBS for sequence/polymorphism data.

tween AF-IBS and AF-IBD and establish an efficient ABC approach for relating AF-IBS to AF-IBD. Finally, the AF-IBS-based ABC method is evaluated in simulations and applied to the estimation of demographic histories from empirical SNP data for human populations.

5.2.1 Examining AF-IBD under various demographic scenarios

I first examined how AF-IBD behaves under different scenarios of population size changes, by analyzing the mathematics and generating simulations.

For a mutation s of allele frequency j in the n sampled sequences, when the coalescent tree is given, it occurs on the root edge (shown in green in figure 5.1A) of a subtree J with j lineages (shown in red in 5.1A). The recombination and mutation events (hereafter referred to as "events") can then be superimposed onto the tree with rates ρ and μ $base^{-1}$ generation^{-1}, respectively. As in equation (4) of [124],

	Amount of derived alleles	Similarity	AF-IBD/-IBS Distance
old event	high	small	short
young event	low	large	long

Table 5.1: The main concepts that are *expected* under a neutral constant size model are given. Events refer to mutation or recombination events. Statements refer to a specific site (position) in a set of haplotypes, related by an underlying coalescent tree.

l$_{n,j,s}$ follows an exponential distribution with the rate parameter as the event rate, integrated over time and lineages in the subtree J:

$$l_{n,j,s} = \sim Exp(T_{J,s}(\mu + \rho)) \tag{5.1}$$

where T$_{J,s}$ is the total length of the subtree J in generations, defined by the mutation s. AF-IBD$_{n,j}$, which is the expectation of l$_{n,j,s}$ over all sites of frequency j out of n, can be integrated over all sites of j out of n as:

$$\begin{aligned} AF - IBD_{n,j} = E[l_{n,j,s}] = \int_{L=0}^{\infty} \int_{\tau=0}^{\infty} P(T_{J,s} = \tau) \\ P(l_{n,j,s} = L | T_{J,s} = \tau) d\tau dL \end{aligned} \tag{5.2}$$

where E(T$^{-1}_{J,s}$) is the expectation of the inverted total length of the subtree across all mutations of frequency j. Denote the absolute time as τ and the variable population size as a function of τ, N(τ). The distribution of AF-IBD can be derived by simulating a large number of coalescent trees as proposed previously [124]. Details of calculating the distribution of AF-IBD can be found in chapter 8.1.2. This procedure is referred to as the tree sampling method, and it was used in the current work to study different models of population size changes.

To understand how AF-IBD responds to population size changes, I simulated models of various demographic scenarios including constant size, bottleneck, exponential growth and complex models. A total of 1,000 coalescent trees were generated for each model. The sample size was set to be 100, so AF-IBD for j=2. . . 99 were calculated. The mutation and recombination rate were both set to the arbitrary value of 2.5×10^{-8} gen^{-1} site^{-1}. The constant size models assumed different population sizes of 1,000, 5,000, and 10,000. The scenario of expansion was examined by assuming that the population size grew exponentially from an ancestral population size of 500, 1,000, and 10,000 to a present population size of 10,000, 50,000, and 100,000, starting at a time point between 40 and 2,400 generations ago. A series of models of single bottlenecks were simulated with the event occurring sometime between 200 to 3,200 generations ago, with the reduction factor being 0.3, 0.1, or 0.01, and the duration ranging between 10 and 100 generations. Finally, a series of complex models were also simulated with an expansion event following a bottleneck event, or two or three consecutive bottlenecks. Combinations of events of various times of onset, durations, and magnitudes were examined. To quantify the effects of population size changes on AF-IBD, the AF-IBD vectors from various models were compared with that of a standard constant size model with N_e of 10,000.

5.2.2 Parameter estimation with AF-IBD using ABC

To further analyze the properties and information content of AF-IBD, ABC was applied. The underlying idea of ABC is that observed and simulated data sets are summarized into several representative values, which are then compared to find the simulations which best match the observed data. I implemented the ABC approach as described previously [30]. The aim was to investigate whether underlying demographic parameters can be estimated, if only AF-IBD is used to summarize a data set. Here, I assumed that AF-IBD can be calculated from the observed (simulated) data. Later I developed a procedure to relate AF-IBD to the statistic AF-IBS, which is directly calculated from the observed data. All data were generated by simulating coalescent trees as described in the previous section. The sample size of 100 was assumed but only AF-IBD for j=2, 3. . . 41 were considered as summary statistics for the ABC calculation. Pseudo-observed (i.e., simulated data sets for which the true values of the parameters were known) were generated for 300 parameter sets from each of three different demographic models. One million ABC simulations, with parameters drawn from the uniform parameter prior distributions, were then compared with the pseudo-observed data to calculate the posterior parameter distributions. See chapter 8.1.4, for details concerning the ABC settings.

The first model assumes a constant size with a single parameter, the effective population size N_e. The second model was a 2-parameter sudden-growth model, in which the ancestral population size is fixed to 10,000 and starts growing exponentially at time T_1 ago until reaching a present day population size of β∗10,000, and the third was a 3-parameter single-bottleneck model of a fixed ancestral population size of 10,000, whose population size declines by a factor β at time T_1 and then recovers to 10,000 at time T_2.

The accuracy and performance of this AF-IBD-ABC approach were evaluated by the relative root mean square error (RMSE, which is the square root of the mean square error divided by the true value), the mean absolute error (MAE, a weighted average of the absolute errors, with the relative frequencies as the weight factors), and the 95% and 50% coverage (proportion of times in which the true parameter value is inside the equal tailed 95% or 50% credible interval [CI]). These measurements were calculated by taking the mode of the posterior distribution as a point estimate. In table 5.2, for both the sudden-growth and the bottleneck model, ancient N_e was fixed to 10,000. In the bottleneck model, the population recovered 100% of its original size after the bottleneck event. Each estimation was based on the comparison between one pseudo-observed AF-IBD and one million simulated

Model	Parameters	Prior~U	RMSE	MAE	95% Cov	50% Cov
Constant	N_e	1,000-10,000	0.0498	0.0425	0.97	0.72
Sudden growth	T_1	200-800	0.1920	0.1392	0.94	0.68
	β	2-10	0.0851	0.0624	0.97	0.63
Bottleneck	T_1	200-800	0.6761	0.4457	0.93	0.61
	T_2	200-800	0.5412	0.4414	0.95	0.55
	β	0.01-0.3	0.4311	0.3259	0.94	0.54

Table 5.2: Measures of accuracy for AF-IBD-ABC parameter estimation. Prior gives the ranges of uniform priors. Cov represents the 50% and 95% coverage.

AF-IBD statistics. The ranges for the uniform prior distributions for each parameter are given as well.

5.2.3 Use of AF-IBS for sequence or polymorphism data

When considering realistic polymorphism data, it is not easy to estimate AF-IBD, as the status of IBD is not directly observable. Although there exist various methods to estimate IBD-based statistics from sequence or SNP data ([11, 97]), these methods are too computationally intensive to apply to genome-wide data. I, therefore, use AF-IBS to replace AF-IBD. Other than IBD, IBS can also result from recombination among homologous haplotypes or back mutation, or simply lack of polymorphic sites [64]. When the SNP density is high, in theory the length of IBS should be mainly accounted for by IBD. Therefore, I test whether AF-IBS has similar sensitivity as AF-IBD toward ancient population size changes. I simulated sequence data from the same models as for AF-IBD (see model cartoons in appendix figure 8.2). For the sequence simulation, sets of 100 haplotypes of length 2 Mb were simulated for 1,000 replicates for each set of demographic parameters. Simple ascertainment schemes were applied in which only sites variable within a parallel discovery panel of 5, 7, 10, or 15 haplotypes were kept to compose the polymorphism data and used to calculate AF-IBS. Throughout these simulations, the mutation rate and recombination rate for sequence data were assumed to be 2.5×10^{-8} and 1.3×10^{-8} gen^{-1} site^{-1}, respectively, which are the reported genome averages [22, 142]. Results for different demographic models were then contrasted to a constant size model of N_e=10,000 to examine whether AF-IBS shows similar demographic sensitivity. I then examined how different AF-IBD and AF-IBS are for the same demographic history. I introduce the ratio between AF-IBS and AF-IBD (hereafter referred to as SD ratio) for the same demographic model. The SD ratio is defined as a vector of index j where SD ratio_j=AF-$\text{IBS}_{n,j}$/AF-$\text{IBD}_{n,j}$ for each frequency j.

Model	Parameters	Prior∼U	RMSE	MAE	95% Cov	50% Cov
1 Parameter	N_e	1,000-20,000	0.080	0.0714	0.96	0.69
3 Parameter	T_1	100-2,000	0.195	0.131	0.94	0.57
	β	0.01-0.9	0.179	0.171	0.94	0.52
	N_e	5,000-40,000	0.153	0.131	0.93	0.61
5 Parameter	N_e	15,000-50,000	0.257	0.213	0.93	0.59
	T_1	50-2,000	0.391	0.314	0.92	0.49
	β_1	0.01-0.5	0.516	0.467	0.90	0.52
	T_2	10-510	0.314	0.201	0.91	0.48
	β_2	0.1-0.4	0.402	0.357	0.89	0.46

Table 5.3: Measures of accuracy for AF-IBS-ABC parameter estimation. Prior gives the ranges of uniform priors. Cov represents the 50 and 95% coverage

5.2.4 Parameter estimation with AF-IBS using ABC

I established an ABC method using AF-IBS to estimate demographic parameters and evaluated its performance in the simulated scenarios.

The accuracy and performance of this AF-IBS-ABC approach were similarly evaluated as already described for the AF-IBD-ABC approach (see table 5.3). I simulated 300 random data sets for each of three different models, which have one, three, and five parameters, respectively. The 1-parameter model is similar to the previous constant size model. The 3-parameter model is a sudden-growth model in which an ancestral population size increases instantly by a factor of β at time T_1. The 5-parameter model assumes a population size reduction from an ancestral size at time T_2 by a factor of β_2 and a population expansion at time T_1 by a factor of β_1 to a current size of N_e. For all models I sampled the pseudo-empirical parameters from a uniform prior on each parameter space. For each simulation, 250 10-Mb segments, each composed of 42 haplotypes, were generated with maCS [17]. For all analyses of AF-IBS-ABC, I used the software recosim [119] to simulate a random map of variable recombination rates across 10-Mb regions. I used the same recombination parameters as in the “best fit” model of [119], and the basal recombination rate is set according to the autosomal deCODE distribution [82]. I generated 250 of such 10-Mb maps covering the whole genome, and each simulation takes one of them.

A simple ascertainment scheme was applied to match the SNP densities of all allele frequencies to that of the empirical data, as similarly applied before [119]. Briefly, the empirical allele frequency spectrum was determined for both Yoruba and French, and then the simulated SNPs of a certain derived allele frequency (DAF) were repeatedly removed until the SNP densities in simulations equaled that of the empirical data. AF-IBS was then calculated for the simulated SNP data.

Theoretically, the ABC method based on AF-IBS can be done by randomly generating large amounts of SNP data and calculating their AF-IBS as described earlier.

However, this is computationally non-feasible as the SNP data simulation at genome scale is very time consuming, and the required number of samplings in ABC is usually very large, for example, 10^6. Here I developed a new ABC approach to solve this problem. Note that the simulation of AF-IBD is very efficient as only coalescent trees are sampled. If AF-IBS is calculated from AF-IBD, then the AF-IBS values can be efficiently obtained by tree simulations. Noting that AF-IBS and AF-IBD are closely related, and their SD ratios are relatively robust against changes in demographic parameters (supported by the analysis, shown in section 5.3.4), I constructed a SD ratio grid on which AF-IBD can be efficiently converted to the corresponding AF-IBS. The SD ratio grid approach is implemented as follows: First the ratios of AF-IBS/AF-IBD were obtained for a predefined grid of parameter values, by simulating both coalescent trees and SNP data. SD ratios for any arbitrary parameter sets are then imputed based on this grid, assuming local linearity along the parameter values (details about the construction of the SD ratio grid and the SD ratio imputation method can be found in chapter 8.1.3). Based on this, the ABC method using AF-IBS is briefly summarized as follows:

1. 10^6 random parameter sets are sampled from the priors
2. AF-IBD is calculated for each parameter set
3. The AF-IBS/AF-IBD ratio is imputed from the SD ratio grid, and AF-IBS is calculated from AF-IBD
4. The simulated AF-IBS is compared with the empirical AF-IBS to give the best-fitting model

5.2.5 Model misspecification

The real populations may have hidden population structures that are not represented by these simple models. It is, therefore, important to evaluate whether such hidden population structure will influence the AF-IBS calculation. I analyzed the AF-IBS behavior under certain model misspecifications. To see the effects of potential hidden population structure, I simulated an ancestral population of size N_e=10,000 that split into two populations, 200, 500, and 1,000 generations ago (constant size demography). I analyzed the effect on AF-IBS of the two daughter populations having sizes 50/50 or 30/70 percent of the ancestral population, respectively (50 samples each). After that, I additionally simulated gene flow (0.1% and 0.5% per generation) between the two populations.

Empirical data are usually obtained as unphased genotype data, which is subject to an additional statistical calculation of phase reconstruction to infer the haplotype composition. As AF-IBS essentially measures how long a homologous segment extends, it may be sensitive to switching errors during the phase reconstruction.

Therefore, I evaluated the effect of errors in the phase reconstruction on the AF-IBS calculations. I applied the program fastPHASE [120] to various SNP data sets, simulated under different demographic scenarios (1-, 3-, and 5-parameter models with different parameter sets). The parameter values for the demographic models were chosen to cover a broad range of possible scenarios, with ancestral N_e ranging between 5,000 and 30,000 and recent N_e ranging between 5,000 and 40,000 and times of expansion or bottleneck events ranging between 50 and 2,000 generations ago. The parameters for fastPHASE were set to the same values used for the phasing of the empirical data. I then analyzed the ratio of AF-IBS before and after the phasing. The ratios indicate that the phasing errors do have an impact on the AF-IBS calculation, especially for the lower DAFs (figure 8.3). As the effects are similar for different demographic scenarios, I calculated the average ratios across all the simulations. The AF-IBS values calculated for the empirical data were then corrected by multiplying the inverses of these average ratios, for the lower DAFs 2-12. AF-IBS values for higher DAFs do not seem to be affected by the phase errors and are, therefore, not corrected.

5.2.6 Parameter estimation for empirical data

I applied the ABC method using AF-IBS to the empirical SNP data. The genome-wide SNP data from the CEPH-HGDP panel was used [89]. The data were phased with the fastPHASE program and then corrected for the effects of phasing error, as described before. Statistics were calculated for 42 randomly chosen chromosomes from each population. For the calculation of AF-IBS I considered only sites at least 5 Mb away from the chromosome ends, which resulted in AF-IBS values for ~490,000 sites, covering a genomic length of ~2.2 billion bp. I tested the same three models as for the pseudo-empirical SNP data described earlier. To decide which model performs the best, I performed a model selection using a Bayes factor analysis [9, 65]. The same number of simulations was chosen for each model, so that they were a priori equally likely, and I computed the ratio of acceptance rates for each pairwise model comparison. The posterior probability of a given model is then approximated by the proportion of accepted simulations given this model. The approach used here is implemented in the R package *abc* (http://cran.r-project.org/web/packages/abc/index.html[1]). I additionally performed a test based on a logistic regression method [32], where a multinomial logistic regression is fit with the model being the categorical dependent variable. The regression is local around the observed summary statistics vector (as in the parameter estimation). Finally, the model probability is assessed at the point corresponding to the observed vector of summary statistics. For this method I used the "calmod" function written by Beaumont (available from the *popabc* package at http://code.google.com/p/popabc/[2]).

---

[1] last visited on 30/11/2013

[2] last visited on 25/11/2013

Model selection was based on 1 million simulations for each model.

On the basis of this model choice approach, I additionally analyzed the power of this procedure to accurately recover the true model using the AF-IBS-ABC approach, following previous methods [32]. I used the 300 simulated ascertained and phase-corrected data sets from the prior distribution for each model considered (1-, 3-, and 5 parameters) and analyzed them using the same simulations and pipeline as for the empirical data. Each of the 300 data sets then refers to one of the three models with the highest posterior probability. I then counted how many times this approach was able to identify the true model.

5.3 Results for AF-IBD & AF-IBS

5.3.1 Properties of AF-IBD

When AF-IBD is plotted against the allele frequency, it can be seen that AF-IBD decreases monotonically with increasing allele frequency (see figure 5.2A and 8.1A). This is easily understood, as variants of higher allele frequencies are on average older, and their intra-allelic IBD, therefore, has decreased more over time. Parameters given in the legends of figure 5.2 represent the start and end of the bottleneck in generations before present, as well as the reduction factor during the bottleneck, the number of generations for each period of growth lasting to the present day as well as the ancient and present day population sizes, respectively. As explained, different demographic histories have distinct effects on the outcome of AF-IBD, which clearly shows the sensitivity of this statistic to population size changes.

When AF-IBD values are compared between constant size models of different population size, it can be seen that the ratio is constant across different allele frequencies, and it is the inverse of the ratio of population size (figure 5.2B). This is expected given that coalescent rescales with population size.

AF-IBD is essentially contrasting two different measurements of allele age. Each allele frequency defines a time range on the coalescent time scale, for example, in the unit of inverse of effective population size (appendix equation 8.1). For the same time range in coalescent scale, when the effective population size is big, then the absolute time span is long, resulting in shorter average IBD length. Otherwise, the average IBD length becomes longer. This suggests that smaller AF-IBD indicates a bigger effective population size and vice versa. Therefore, the AF-IBD curve along the allele frequency spectrum reflects the details of population changes.

The observations from simulations are consistent with the above statements. I contrasted AF-IBD values for different demographic models with that of a constant size model of N_e=10,000. Figure 5.2C shows the comparisons among four bottleneck models. All ratio curves are elevated above 1, with a single peak at different allele frequencies and magnitudes. The most recent bottleneck has a peak around allele frequency 10 with the highest ratio approximately 2.1. The intermediate-aged bottleneck is shifted to the right to around frequency 15 with a peak height of 1.6 and even the relatively ancient bottleneck event, starting 1,000 generations ago, also resulted in elevated ratios around the frequency 20-30. It is obvious that AF-IBD has higher sensitivity to more recent events than older ones of the same magnitude. On the other hand, strong ancient events can also induce big changes in the relative AF-IBD curve. This can be clearly seen in the fourth model, where the duration of the size reduction was increased to three times that of the third model (figure 5.2C).

For the scenarios of expansion, figure 5.2D shows that the ratios of AF-IBD started from 0.3-0.4 at the lower allele frequency range, much lower than the value of 1 expected under a constant population size. The ratio curve recovers quickly back to close to 1 for the recent expansion. The increase of the ratios along the allele frequency is progressively slower and to a lower maximum when the expansion starts earlier in time (figure 5.2D). Finally, the AF-IBD ratio is also sensitive to complex models where multiple events shaped the population size trajectory. Figure 5.2E shows the AF-IBD ratios for two complex models, one defined by a recent weaker bottleneck (200-210 generations ago, 100 times size reduction) following an old strong bottleneck (1,000-1,100 generations ago, 100 times size reduction; colored in black), and the other defined by a recent expansion (population size from 10,000 to 100,000, starting at 500 generations ago) after an intermediate-aged bottleneck (1,000-1,100 generations ago, 100 times size reduction, colored in red). The two curves clearly differ from each other: for the case of two bottlenecks, the ratio starts above 1 and increases to a first turning point around frequency 10, then rises to the second turning point around frequency 40. For the case of expansion following bottleneck, the ratio starts from below 1 as expected for large population size and keeps ascending above 1 until reaching a maximum at the highest frequency. The increase in the AF-IBD ratio is clearly due to the bottleneck.

5.3.2 AF-IBD-ABC

I first tested an ABC framework assuming that AF-IBD can be directly observed. The purpose was to first analyze how accurate underlying demographic parameters, connected to population size changes, can be estimated in the absence of any complications introduced by the type of empirical data (e.g., ascertainment bias). In table 5.2, shown are several calculated measures of precision, which represent the differences between preset parameter values and estimated parameter values. I calculated the RMSE, the MAE and the 95% and 50% coverage (see 5.2.2). Results from table 5.2 show that this method of inference is highly precise for the single parameter constant size model. This can be explained by the underlying mathematical features of AF-IBD. As shown in figure 5.2B, the reverse ratio of AF-IBD for different constant size models coincides with the population sizes. The estimation for the 2- and 3-parameter models, although slightly less accurate, still provides estimates that are sufficiently close to the true values. The reduced accuracy is expected, as the same AF-IBD curve might result from different but equivalent demographic histories. For example, the general effect of a strong but short bottleneck can be very similar to that of a weaker but longer bottleneck. However, in most cases, I could estimate the true underlying parameter values with a high level of accuracy (table 5.2), demonstrating the validity of the AF-IBD-based ABC approach.

5.3.3 Properties of AF-IBS

Presenting the comparisons between AF-IBS and AF-IBD for models of three different scenarios: constant size, expansion, and bottleneck. Specifically AF-IBD and AF-IBS of the bottleneck and expansion models were contrasted against those of the constant size model, and the ratios were plotted together (figure 8.2A and 8.2B). It can be seen that the ratio curve of AF-IBS is close to that of AF-IBD. In the bottleneck scenario, the AF-IBS ratios are shifted slightly below the AF-IBD ratios, but the position of the peak is well conserved. For the expansion scenario, AF-IBS curves are slightly above the AF-IBD curve although the general shape is unchanged. Comparisons for additional population size change models are shown in figure 8.1B. Overall, AF-IBS curves for different ascertainment schemes are very similar to each other, which suggest that the AF-IBS ratio is generally robust to the ascertainment bias schemes implemented here.

5.3.4 The IBS/IBD Ratio

I showed in the previous section that the relative AF-IBD curve is very similar to the relative AF-IBS curve for the same demography, despite different ascertainment schemes. This suggests that AF-IBS is related to AF-IBD in a way that is not affected by the changes in population size. I checked the robustness of the SD ratio between AF-IBS and AF-IBD in various demographic scenarios including constant size, bottleneck, and expansion. Figure 8.2C shows the SD ratio curves for AF-IBS. In figure 8.2C the SD ratio starts at a low level and rises steeply above 1.0 for the first few frequency bins. This is an artifact due to the fact that the maximum length of AF-IBS is 0.5 Mb (see section 5.3.1), whereas AF-IBD estimation from tree simulation theoretically can be infinitely long. The subsequent values range between 1.5 and 3, and the curves for the two different models have a similar shape. In fact, what can be seen is that the SD ratio curve distributes within a rather defined interval, across a large parameter space, and the values for each bin in general are in a roughly linear relationship with the parameters (data not shown).

5.3.5 AF-IBS-ABC

I constructed a fast ABC pipeline that applies to the observed AF-IBS values. I first checked whether correct estimations can be obtained for simulated pseudo-observed

	1-Parameter Model	3-Parameter Model	5-Parameter Model
1-parameter model	0.84	0.13	0.03
3-parameter model	0.09	0.78	0.13
5-parameter model	0.06	0.20	0.74

Table 5.4: Power of AF-IBS ABC to recover the true model.

SNP data. Three models (constant size, sudden growth, and expansion-after bottleneck) were tested, which contain one, three, and five parameters, respectively. The entire workflow of ABC for AF-IBS is shown in figure 8.4.

Figure 5.3 shows the estimated posterior distributions for some parameters of interest from the 1-, 3-, and 5-parameter models. As presented in table 5.3, inference based on AF-IBS-ABC is relatively accurate and precise for the 1- and 3-parameter models and still reliable for the most complex 5-parameter model. I also analyzed the power to correctly recover the true model based on the logistic regression procedure. As described earlier, I counted how many times the method correctly assigned the true model in a set of 300 simulated data sets from the prior distributions of each model. As presented in table 5.4, data sets are properly assigned in most of the cases. However, the more complex the model, the less power this approach has. Also, the inferred empirical Bayes factors are in good agreement with the ones I simulated.

5.3.6 Application to genome-wide data

I then applied the currently explained approach to the genome-wide data set of the CEPH-HGDP panel [89]. Figure 5.4A shows AF-IBS for the first 34 DAF bins calculated from 42 randomly chosen chromosomes from each of 11 worldwide populations. As high DAF values reflect old mutations and low DAF values reflect more recent mutations, variation in AF-IBS values indicates population size changes at different times in the past. The AF-IBS values for higher DAF for African populations are clearly smaller than for all non-African populations, indicating much reduced ancient population sizes for non-Africans compared with Africans. Furthermore, populations show continental or areal clustering, which suggests similar demographic histories for populations within the same cluster. All non-African populations show higher variability in the tails of the curves. This is due to the fact that fewer sites with high DAF are present in these populations, probably because of severe bottlenecks. I then analyzed two representative populations in more detail: Yoruba from Africa and French from Europe.

5.3.7 Model misspecification

I calculated AF-IBS for a standard constant size model and the models assuming different population structure and migration. The ratios of AF-IBS between the standard model and the models with complete population structures were approximately 1, ranging from 0.96 to 1.14, indicating that hidden population structure does not significantly influence the results. Adding migration between the daughter populations further reduces the difference between the standard and alternative models (figure 8.3).

On the other hand, the phase reconstruction error seems to have an impact on the AF-IBS calculation. I contrasted the AF-IBS values of different scenarios both before and after phasing (figure 8.3). The ratios are rather consistent among different scenarios, starting at approximately 0.8 for the DAF=1 and recovering back to 1 after DAF=12. This suggests that the method might underestimate AF-IBS for the lower DAFs, due to the phasing errors. Such bias is corrected by multiplying the empirical AF-IBS values with the phasing error correction ratios.

5.3.8 ABC analysis for Yoruba

Table 5.5 lists the results for the estimated demographic models for Yoruba. The logistic regression analysis was done before the actual ABC analysis. Among all model comparisons, the 3-parameter model of sudden expansion was the best fitting model (Bayes factor 4.1 and probability of 0.63), followed by the 5- and 1-parameter models. The most likely constant population size was estimated to be approximately 8,850. The inferred parameter ranges for the 3-parameter model suggest a constant recent population size of approximately 22,915 (95% CI: 21,706-24,110) followed by a population-size decrease (backward in time) to approximately 0.57 of the recent N_e (95% CI: 0.53-0.62) to an ancestral size of 13,061 at 806 (95% CI: 685-1,030) generations ago. The 5-parameter model had a probability of 0.25. The inferred parameter ranges suggest a recent population size of approximately 28,000 followed by a bottleneck between 1,005 and 1,302 generations ago, with an ancestral size of approximately 18, 600 and a bottleneck size of approximately 8,000. Results from figure 5.4B show that the ratio between the observed AF-IBS and the best 1 parameter model simulations AF-IBS (black line) is approximately 1 for most DAF, which supports a relatively stable ancient population size, followed by a more recent expansion (ratio below 1 for the first bins). Therefore, the 3-parameter model of a simple expansion seems to best explain the data.

5.3.9 ABC analysis for French

To analyze the French data, only the first 2...36 AF-IBS values for 42 randomly chosen chromosomes were used, as there are not enough high-frequency DAF cases to get a reliable genome-wide average for their AF-IBS values. Among all model

Population	Model	Parameters	Prior∼U	Regr. Est.	95% CI
Yoruba	Constant size	N_e	1,000-41,000	8,850	7,825-13,617
	Sudden growth	N_e	5,000-40,000	22,915	21,706-24,110
		T_1	100-2,000	806	685-1,030
		β	0.01-0.9	0.57	0.52-0.63
	(Bottleneck + sudden growth)	N_e	15,000-50,000	28,310	27,081-32,506
		T_1	50-2,000	1,005	780-1,436
		β_1	0.01-0.5	0.28	0.12-0.35
		T_2	60-2,500	1,302	895-1,498
		β_2	0.11-0.9	0.81	0.73-0.85
French	Constant size	N_e	1,000-41,000	6,311	4,753-8,623
	Sudden growth	N_e	5,000-40,000	5,043*	*
		T_1	100-2,000	351	*
		β	0.01-0.9	0.21*	*
	(Bottleneck + sudden growth)	N_e	15,000-50,000	18,300	16,116-22,082
		T_1	50-2,000	1,300	987-1,520
		β_1	0.01-0.5	0.18	0.14-0.25
		T_2	60-2,500	1,580	1,410-1,805
		β_2	0.11-0.9	0.55	0.32-0.68

Table 5.5: ABC estimation results for empirical data for Yoruba and French. Prior gives the ranges of uniform priors. Regr. Est. is the regression estimate.[3]

comparisons, the 5-parameter model of a bottleneck followed by sudden expansion was the best fitting (Bayes factor 3.9 and probability of 0.71), followed by the 3- and the 1-parameter models. Table 5.5 presents that the most likely constant population size was estimated to be approximately 6,300, which is smaller than for the Yoruba population. Again, note that this estimate cannot directly be compared with usual measurements of N_e. Trying to fit a 3-parameter model of sudden growth did not yield any reliable parameter estimates. I also analyzed the 5-parameter model with ABC (table 5.5). The results suggest a recent population size of approximately 18,300 (95% CI: 16,115-22,082). The ancestral population size was estimated to be 10,065 (95% CI: 5,856-12,444). The timing of the bottleneck was estimated to be between 1,580 and 1,300 (95% CI: 1,410-1,805; 987-1,520) generations ago with a population size of approximately 3,300 (95% CI: 2,562-4,575) during that time. Importantly, the CIs of the parameters seem to be rather narrow compared with the priors, except for the two time parameters. The ratio curve of the best 5-parameter simulation against the best constant size simulation also matches closely with that of the empirical data against the best constant size simulation (see figure 5.4C). Therefore the 5-parameter model of a bottleneck with an expansion is the best fitting model for the French.

---

[3] *For this model I was not able to reliably infer parameter values from the French data.

5.4 Discussion AF-IBD & AF-IBS

Using genetic data to make inferences concerning the demography of populations (especially population size changes) has long been of interest [47, 84]. As genome-wide SNP and full sequence data are becoming increasingly abundant for human populations and other species, it is of great interest to make efficient use of such data to infer ancestral demographic history with high accuracy. In this chapter, I introduced two potentially very useful statistics, AF-IBD and AF-IBS, which make use of haplotype configuration changes resulting from both mutation and recombination events. As was shown, both have some desirable mathematical properties, which determine their high sensitivity to population size changes even for complex demographic histories over a wide time range.

The high sensitivity of AF-IBD and AF-IBS toward ancient population size changes results from contrasting two types of age estimators: the intra-allelic LD inferring the absolute age and the derived allele frequency surrogating the coalescent scale age. In this study, the ABC approach to estimate the trajectory of population size was used, by minimizing the distance between the summary statistics calculated from simulated and observed data. On the other hand, if a closed form equation can be found that defines the AF-IBD/AF-IBS as a function, say G(j,N(τ)) (assuming a one-to-one map between N and G), of allele frequency j and N(τ), it is possible to analytically derive N(t) by solving the reverse function G_{-1}. In the perspective of the coalescent, the AF-IBD/IBS statistics are similar to AFS: they are all conditioned on the derived allele frequency. Although the AFS measures the length of the root edge of a j-node subtree (corresponding to the green edge in figure 5.1) by counting the number of mutations, the AF-IBD/IBS measures the total subtree length (the red subtree in 5.1). In principle, the subtrees should be more informative about the population size changes than their root edges. This is because the subtrees of the same DAF coalesce in the same time interval and are responsive to the same population size changes. The root edges on the other hand do not necessarily overlap in time for a given DAF and thus are less responsive to a particular population size change. In [91], authors proposed the HCN statistics, which also make use of the haplotype distributions. By summarizing the local haplotype frequency distribution, the HCN essentially makes use of both recombination and mutation events to reflect the properties of the coalescent trees within windows of fixed recombination size. The current statistics are similar to HCN in the use of both mutation and recombination information, but AF-IBD/IBS focus explicitly on the tree defined by the central SNP. The recently proposed PSMC method directly estimates the time of most recent common ancestor (TMRCA) on a pair of genome sequences [87]. By evaluating the coalescent density over the stepwise time intervals, this method revealed many details of the population size trajectories. However, the pairwise comparison by design provides less information on very recent history and is sensitive to recent population structure. The AF-IBD/AF-IBS statistics are based on multiple haplotype comparisons and, therefore, may help complement the PSMC for recent history.

In this work, AF-IBS is associated to the AF-IBD statistic by a correction ratio. In fact, it might be possible to express AF-IBS as functions of AF-IBD in explicit mathematical form. The greater AF-IBS than AF-IBD values at higher frequencies are mainly due to the undetected recombination events (figure 8.2). I introduced the SD ratio as one potential way of transforming AF-IBD to AF-IBS. However, remember that this is an approximate way of solving this issue, and there is room for improvement. Nonetheless, the ABC estimation based on AF-IBS already shows promising accuracy on the pseudo-observed SNP data. The ABC parameter estimation results show that even for quite complicated models such as the 5-parameter model, parameters of interest can be estimated accurately.

Current results show that the AF-IBS ratios are relatively robust against very different ascertainment schemes (figure 8.2). This suggests that possible misspecifications of the ascertainment scheme should not affect the inference very much. Some SNP data are censored for the lower minor allele frequencies. This will certainly cause losses of information for very recent or ancient demographic events. On the other hand, the switching errors during the phase reconstruction from the empirical genotype data do seem to cause a slight underestimation of AF-IBS for lower DAFs. This is not difficult to understand: phasing errors can be seen as a low level of artificial recombination. When this fraction of recombination rate, say ρ_{phase} is added to the term $\mathrm{T}_{J,s}(\mu+\rho)$ in equation 5.1, it tends to reduce AF-IBD/AF-IBS when $\mathrm{T}_{J,s}$ is small, which corresponds to lower DAFs. However, the effect of ρ_{phase} can be negligible when $\mathrm{T}_{J,s}$ or DAF is big. This problem can be minimized by using phase certain SNP data, such as those genotyped on trio samples.

In the application of the AF-IBS statistic to the CEPH-HGDP Yoruba and French data, it was seen that neither of the two data sets can be fully explained by the constant size model. The three parameter model with a recent population expansion provides a slightly better fit to the Yoruba data than the more complex 5-parameter model. For the French, the 5-parameter model featuring both a bottleneck and an expansion is needed to explain the observed data. This result is in general agreement with previous studies. Most of the existing studies showed that a simple expansion is sufficient to account for the African demography [1, 72, 96, 135], whereas Schaffner et al. [119] suggested a minor bottleneck for the Yoruba (inbreeding coefficient F=0.008), and Li et al. showed a mild reduction between 20,000 and 100,000 years ago [87]. Moreover, all studies infer that European populations had at least one bottleneck before the recent expansion [1, 72, 91, 96, 119, 135, 139]. For the specific parameter estimation, the comparisons among different studies are summarized in tables 5.6 and 5.7. Current result show that the Yoruba had an ancient population size (N_{anc}) of ∼13,000 recovering to a present size (N_{cur}) of ∼22,900. This is in good agreement with previous studies (N_{anc} 9,069-12,500; N_{cur} 16,233-31,000, tables 5.6 and 5.7). The time of expansion T_{exp} varies considerably among different studies.

Studies	N_{anc}	N_{cur}	T_{exp}(gen)	T_{exp}(kya)
Adams and Hudson (2004)	10,000	19,000/31,000	1,080	27
Marth et al. (2004)	10,000	18,000	7,500	187.5
Voight et al. (2005)	10,625	21,304	1,000	25
Keinan et al. (2007)	9,069	16,234	7,440	186
Schaffner et al. (2005)	12,500	24,000	17,000	425
Fagundes et al. (2007)	12,722	206,920	-	-
This method	13,601	22,915	806	20.15

Table 5.6: Estimated African demographic parameters compared among different studies.

Although our estimate of 806 generations (∼20 thousand years ago [kya]) is close to previous estimates of 27 kya [1] and 25 kya [135], other studies gave much older estimates (186-425 kya). Results from Li et al. revealed two waves of expansions (or bottlenecks depending on the perspectives), one earlier (200-600 kya) and one later (∼20 kya) [87]. This suggests that different methods may have captured either of the two inferred periods of growth. The more recent expansion given by our result coincides with that of [87] and the last glacial maximum.

For the European demography, our estimates of the ancient population size (N_{anc} ∼10,000) and current population size (N_{cur} ∼18,300) are also similar to those from previous studies of N_{anc} 8,000-10,065 and N_{cur} 10,000-20,000 (tables 5.6 and 5.7).The time when the bottleneck starts (T_{bot}) and the time of recovery (T_{exp}) are surprisingly consistent among most studies, although these two values are usually considered difficult to estimate. Other than one study [96] with older time estimates (T_{bot} ∼87.5 kya, T_{exp} ∼75 kya), the other studies estimated the T_{bot} to be approximately 31-50 kya and T_{exp} approximately 27.5-40 kya (tables 5.6 and 5.7). The current estimations of 39.5 kya and 32.5 kya fall into these two ranges. This bottleneck probably corresponds to the Out of Africa dispersion. Estimates of the population size of the bottleneck (N_{bot}) vary considerably among studies. Our estimate of ∼3,300 is larger than many such estimates (tables 5.6 and 5.7). When the inbreeding coefficient F is calculated [72], the current estimate (0.042) is close to previous estimates of 0.085, 0.02 [119], and 0.032 [139], although much smaller than other studies of 0.125-0.364 (tables 5.6 and 5.7). Li et al. showed a much reduced population size of approximately 1,200 between 40 and 20 kya. These suggest that the currently proposed method may have underestimated the intensity of the bottleneck. The precise reason is not clear, but the 95% lower bound of our N_{bot} is approximately 2,500, suggesting a lower bottleneck size is also possible.

Remember that this is a preliminary study to demonstrate the usefulness of the AF-IBD-related statistics. There are various ways in which the inference can be improved. For example, I used the mean AF-IBD/IBS as the inference statistics in this method. In fact, the distribution of each AF-IBD/IBS for a given DAF is

Studies	N_{anc}	N_{bot}	N_{cur}	T_{bot}(gen)
Marth et al. (2004)	10,000	2,000	20,000	3,500
Adams and Hudson (2004)	10,000	1,500	20,000	1,500
Wall et al. (2009)	-	625	-	1,240
Voight et al. (2005)	10,695	1,065.9	-	2,000
Keinan et al. (2007)	8,712	-	-	1,280
Schaffner et al. (2005)	-	-	-	-
Lohmueller et al. (2009)	8,000	550	10,000	1,500
This method	10,065	3,300	18,300	1,580
Studies	T_{bot}(kya)	T_{exp}(gen)	T_{exp}(kya)	F
Marth et al. (2004)	87.5	3,000	75	0.125
Adams and Hudson (2004)	37.5	-	-	-
Wall et al. (2009)	31	1,200	30	0.032
Voight et al. (2005)	50	1,600	40	0.19
Keinan et al. (2007)	32	-	-	0.151
Schaffner et al. (2005)	-	-	-	0.085, 0.02
Lohmueller et al. (2009)	37.5	1,100	27.5	0.36
This method	39.5	1,300	32.5	0.042

Table 5.7: Estimated European demographic parameters compared among different studies. Because of the limited available space, estimated parameters are split horizontally.

also sensitive to population size changes (data not shown). This is easy to understand: subtrees of the same DAF span different lengths of the coalescent time scale, therefore may be perturbed by the fluctuating demography at different times or intensities. The power of the inference methods may be further improved by using the full distributions of AF-IBD/IBS.

Moreover, the current computational approach still offers room for improvement. Although coalescent simulators are capable of simulating a wide range of demographic scenarios within a rather short time, simulating full genomes with an underlying variable recombination map is still computationally quite intensive, especially when every full sequence simulation needs to be ascertained and corrected for phase reconstruction error. Although the simulations I carried out provide support for the overall effectiveness of this approach, further work should improve the accuracy of the parameter estimates, especially for more complex (and hence realistic) models.

In conclusion, it was shown that quite accurate estimates of demographic parameters can be obtained from ascertained genome-wide SNP data, even for complex underlying population histories. Improved inference may also be achieved by applying more elaborate methods of parameter estimation, especially when adding more parameters to underlying demographic models. For example, combining the advantages of ABC and MCMC can lead to improved estimation results [140]. Moreover, with full sequence data sets becoming available, the limitations of SNP data will no longer apply. With further work, it might be possible to find the closed forms of AF-IBS and AF-IBD as functions of population size change N(τ), and non-parametric methods could potentially be used to infer more realistic demographic trajectories through time.

Interestingly, after completion of this project and quite recently before the completion of the entire dissertation, Harris and Nielsen published a method for inferring demographic history from genome-scale data based on IBS tract lengths [53]. The method is able to predict the IBS distribution for various demographic models and therefore might be very helpful for the AF-IBS method. Given a computationally inferred IBS distribution, the simulations of coalescent trees and the whole idea of the SD ratio grid might become redundant, resulting in a potentially more convenient to use application and eventually more stable results. This way of improving our method is one of the most promising and will definitely be further investigated.

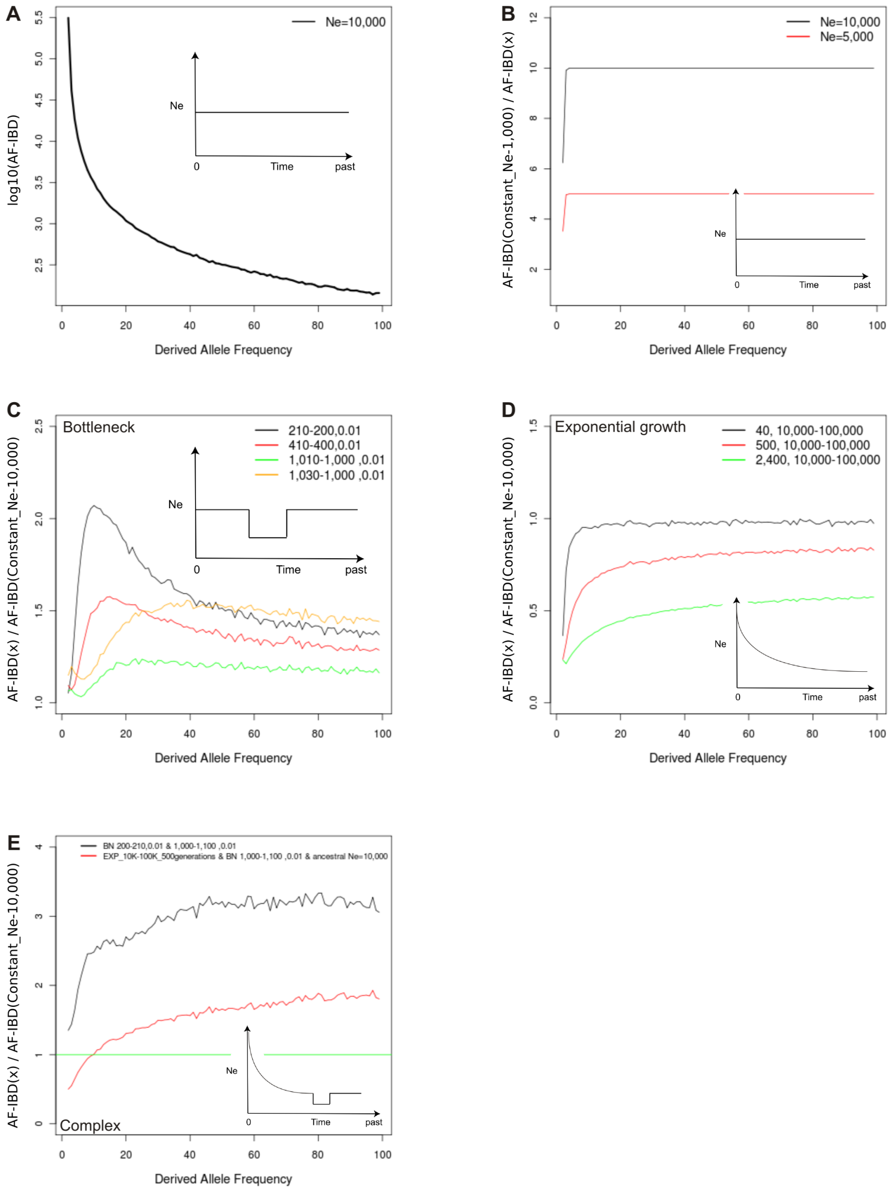


Figure 5.2: AF-IBD was calculated from data sets simulated as coalescent genealogies under various demographic models of interest (see subfigure cartoons) and for a constant size reference model. A) AF-IBD curve calculated from a constant size population of N_e = 10,000. B) AF-IBD ratios between two different models of constant population size (N_e=5,000, N_e=10,000) and one constant N_e=1,000. C) AF-IBD ratios between various bottleneck models and one constant population size model of N_e=10,000. D) AF-IBD ratios between various exponential growth models and one constant population size model of N_e = 10,000. E) AF-IBD ratios between a two-bottleneck model and one constant size model of N_e=10,000, and between a complex bottleneck followed by sudden growth and a constant size model of N_e=10,000.

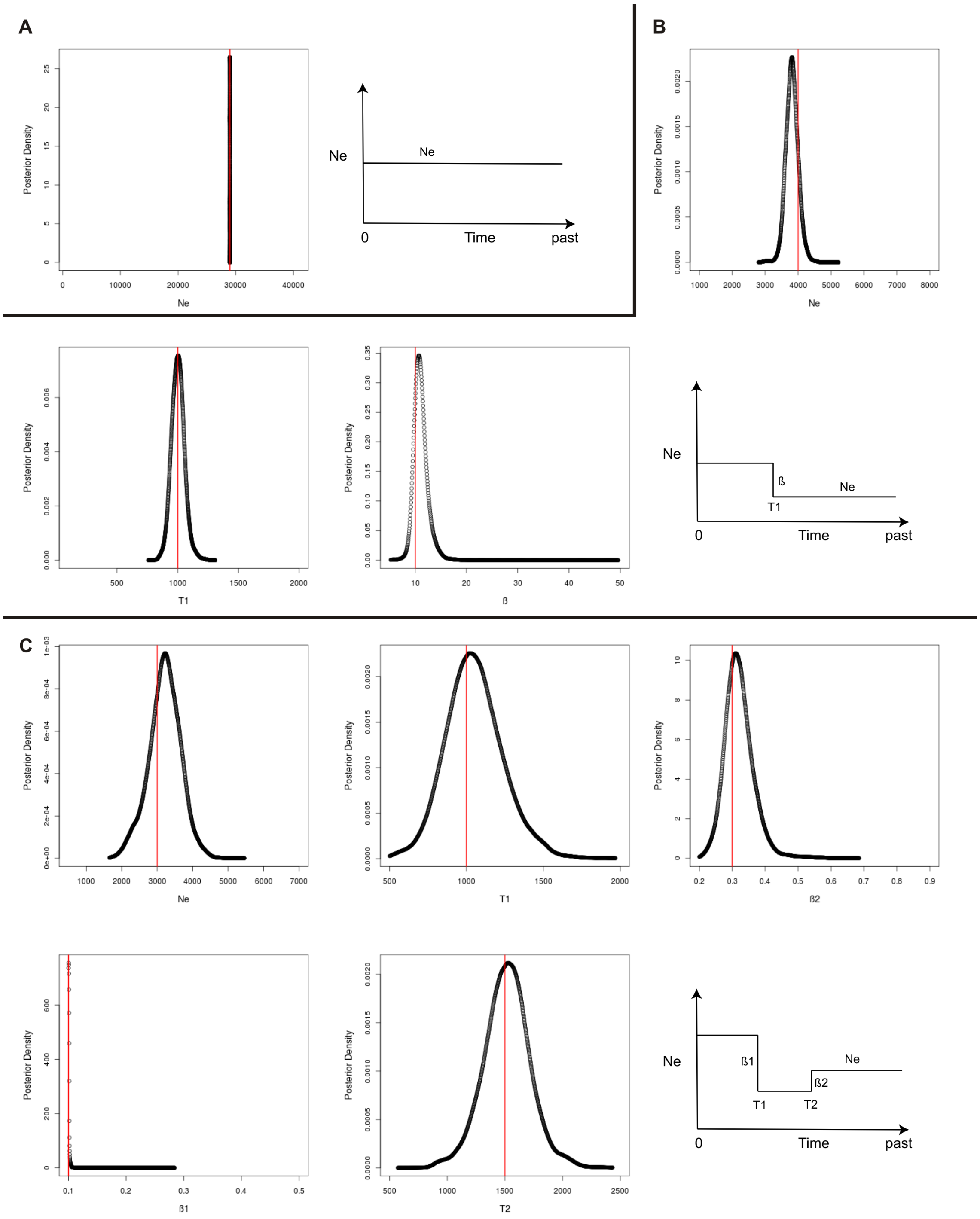


Figure 5.3: The posterior densities from ABC parameter estimation for 1-,3-,5-parameter models are shown. Simulated polymorphism data were used as pseudo-observed data. Vertical red lines represent the true underlying parameter values. For each panel, a cartoon of the underlying model with all parameters that were estimated is shown. A) Results for the single constant size model parameter N_e. B) Results for three parameters of a demographic model of sudden growth. C) Results for five parameters of a model of an ancient bottleneck followed by more recent sudden growth. Prior ranges for each uniformly distributed prior are equivalent to the x-axis ranges.

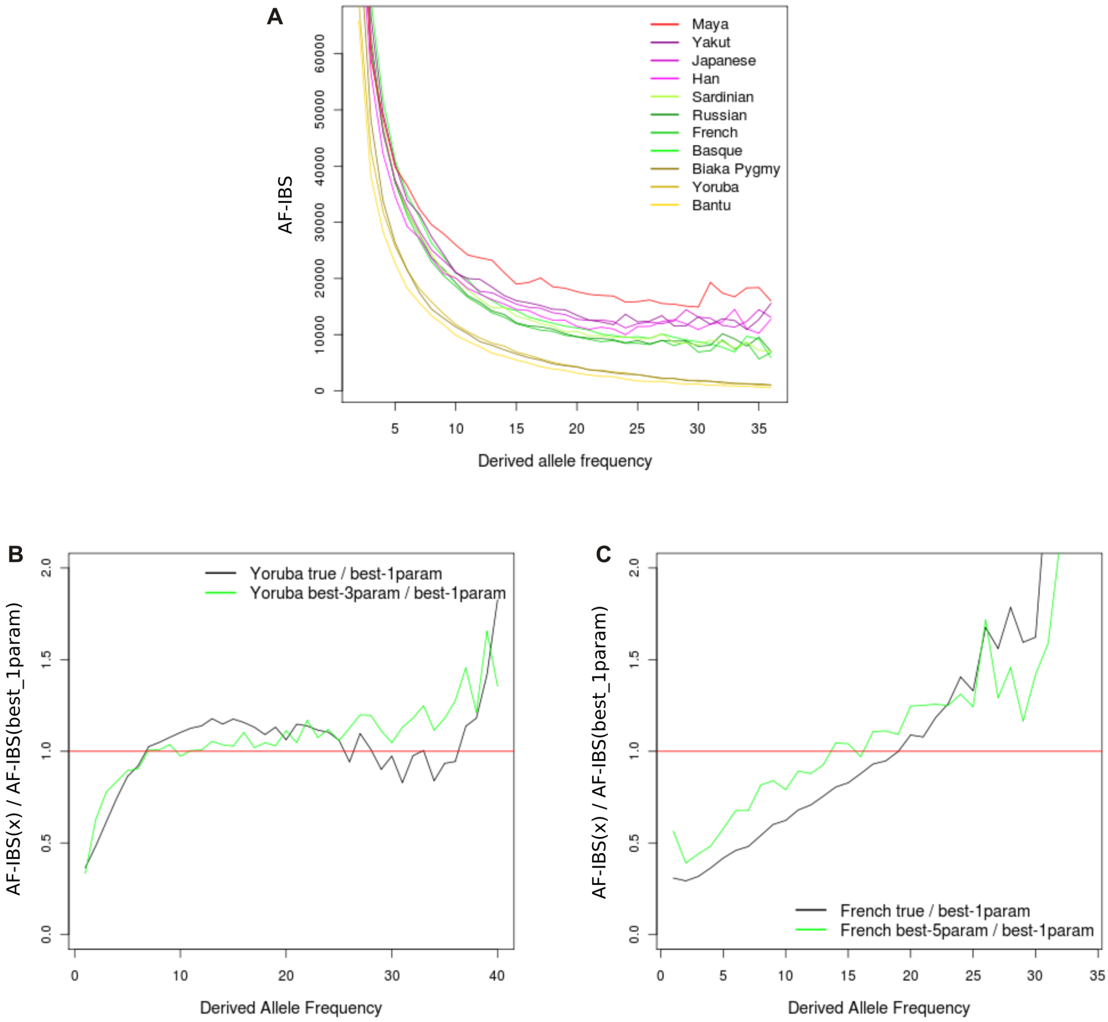


Figure 5.4: A) AF-IBS calculated for various populations from the CEPH-HGDP panel. B) Two ratios between the observed Yoruba AF-IBS and the AF-IBS of the best constant size model simulation and the ratio between AF-IBS from the best 3-parameter simulation and the best constant size model simulation. C) Ratio between the observed French AF-IBS and the AF-IBS of the best constant size model simulation and the ratio between AF-IBS from the best 5-parameter simulation and the best constant size model simulation.

Chapter 6

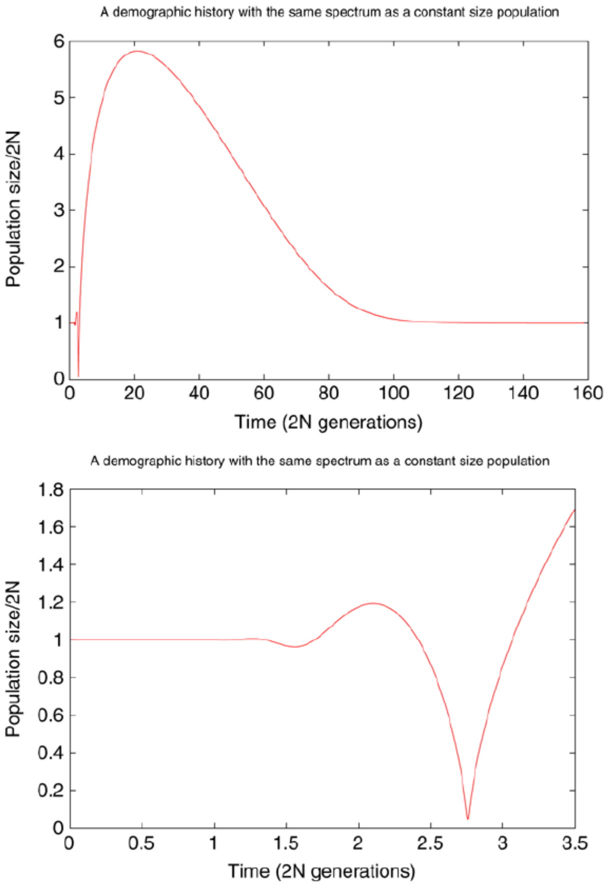
The 2 point spectrum method

> *"Science may be described as the art of systematic oversimplification."*
>
> *Karl Popper (1902 - 1994)*

As previously described, by using a variety of pseudo observed simulated data we were able to show that AF-IBD/AF-IBS are informative about past population size changes. The new insights we gained about underlying patterns of IBD and IBS provide means for a better and thorough understanding of how evolutionary events affect the underlying genetic patterns of individuals and populations. Hence, the initial goal of this project was accomplished. However, one of the limitations that always caught the center of my attention was the dependency on a certain type of demographic model. Using Approximate Bayesian Computation as a model choice approach involves the realization of multiple ABC runs, each based on a different demographic model with a fixed number of parameters, e.g. a constant size or bottleneck model. Furthermore, a thoroughly and well designed validation later needs to be performed in order to choose the best fitting model. If the tested models do not include the true underlying model, results can be biased in various ways. Therefore, approaches that realize the inference of underlying model parameters of interest, while allowing for a variable number of model parameters, can be advantageous.

6.1 Introduction

Additional to my personal interest in achieving more flexibility in the actual inference process itself, the site frequency spectrum, known to be sensitive to past population size changes (see chapter 4), was successfully incorporated as part of the summary statistics in the previous project. Hence, this is a statistic we wanted to investigate further. We came across a more general issue that goes beyond the scope of population genetics. Assuming two different processes that can generate exactly the same outputs for certain summary statistics, then these processes cannot be distinguished anymore. As a result, depending on the way the output is summarized or analyzed, equivalence classes can contain outcomes from different processes. As was shown for the case of the neutral allelic frequency spectrum of a population, two different demographic history processes can produce the exact same site frequency spectrum [101]. Even if the exact frequency spectrum is known, the history of past population size changes is not fully determined. As a consequence, any demographic inference approach just based on the neutral site frequency spectrum alone is biased in a way that conclusions about potential demographic parameters might only account for one out of many possible demographic histories that can explain the observed data. Figure 6.1 shows the example from Myers et al. in which two quite distinct demographic processes produce the exact same frequency spectrum.


A demographic history with the same spectrum as a constant size population

A demographic history with the same spectrum as a constant size population

Figure 6.1: This figure was adopted from [101]. Shown is the population size for a history corresponding to $N \sim (\tau)$ of their figure 2. Most of the interesting structure is for relatively small times, so also shown is an expansion of the figure for time t $\leq$ 3.5.

Combining the previously studied features of the SFS and results from Myers et al. made us think of potential new ways to use the frequency spectrum for demographic inference. One idea among others was to not only look at the frequency information for each single mutated site but for pairs of sites. Taking this pairwise information into account may capture more valuable information from the underlying data as is the case for the standard SFS. We call this statistic the “2 point spectrum” method and the following section describes the methodological steps from the first tests to the final implementation within a rjMCMC framework.

6.2 Idea of the 2 point spectrum

Figure 6.2 describes the basic implementation of the 2 point spectrum. The following pseudo algorithm explains how the initial approach can be calculated:

1. Based on a previously obtained recombination landscape, define a core region $r \in$[a,b] with no or very low levels of recombination
2. For a site j, with j=1,..,s, with s being the total number of polymorphic sites in region r, calculate the derived allele frequency F
3. For the set of haplotypes that are affected by a mutation at site j, calculate how many sites in a region $r2 \in$ [j+1, j+e] have the sub frequencies f with f=1...F and e being an arbitrarily set physical distance in bp
4. If physical position of j$\leq$b, j=j+1 and go to step 1. Else end

The derived allele frequency information is averaged for the entire region r and stored in a triangular matrix. These matrices represent 2-dimensional arrays with s-1 rows and s-1 columns. Rows represent frequencies F for each site j (calculated in step 2) and columns represent the sub frequencies f (calculated in step 3). Each entry [m][n] is the average over all cases with F=m and f=n.

This summary (called sub-sfs) can be obtained from genetic data simulated with ms under different demographic scenarios. In order to compare two competing models of interest, each entry from matrix 1 is divided by the corresponding entry in matrix 2 and differences are represented by a heat plot in R.

The number of mutations on a certain branch of a coalescent genealogy is the outcome of a Poisson process taking the branch-length into account. Hence, the frequency information that can be calculated with numbers of mutations can easily be represented by the actual length of the branch of interest. For example, assuming the infinite sites model, each mutation on branch e in figure 6.2 would result in a new polymorphic site with derived allele frequency F=3 (red mutation), since sequences a, b, and d all share the same mutation and e is said to have a size of 3. The relative length of e then indicates how many possible mutation events could happen on this branch, independent of the actual mutation rate. Incorporating the mutation rate would give an absolute number of mutations. In order to obtain the pairwise information, the lengths of all branches in the subtree defined by e represent the sub frequencies *f*. In the above example the subtree of e is defined by branches a, b, c, and d with sizes 1,1,2, and 1, respectively. Hence, the sub-sfs statistic is calculated without the need to simulate mutations, only by considering lengths of branches. This approach is computationally more efficient than simulating actual mutations, allowing to average over a large number of coalescent genealogies.

To investigate how informative the pairwise frequency spectrum is about past population size changes I used *ms* to simulate pure coalescent genealogies under various demographic models for 20 diploid individuals. Since the coalescent is a random process, the outcome for each demographic model was averaged over $1*10^6$ independent trees. The branch length information from each simulation was then stored in a triangular matrix as described above. I assumed a null model of constant population size over time. Contrasting the matrix from a constant size to the matrix of a non-constant size model then allows to see deviations from the null expectation. The non-constant size models were represented by a variety of bottleneck and expansion models.

Furthermore, I analyzed whether this new statistic is able to distinguish the previously mentioned demographic models that show the exact same allelic spectrum in [101]. As shown in figure 6.1, population size over time is a continuous function and makes it impossible to accurately simulate without further information. The authors provided 5,000 data points, enabling the simulation of the histories as discrete piecewise constant functions with *ms*. Each history (constant and non-constant size) was simulated for 20 diploid individuals (40 haplotypes). The pairwise information was then averaged over $1*10^6$ independent trees per demography and matrices were compared with each other. However, graphically analyzing the results is rather difficult and imprecise, so that we had to explore different methods to make use of the obtained information.

The initial implementation (the sub-sfs method) was thought to capture information from sites that are nested within the same subtree. We later decided to extent this method to all possible pairs of segregating sites (denoted as the 2 point spectrum). This approach is very similar, since a similar information content was captured from the data (i.e. based on the relative frequency information of pairs of sites, the underlying core structure of the genealogy was captured with both methods). The second approach is closer to a study from Jenkins et al. [68] (details follow in the remainder of this chapter). The initial idea was, therefore, used for the first tests in order to see the potential sensitivity of this method. Later, the second approach, based on all possible pairs of sites, was used for the implementation of the MCMC framework.

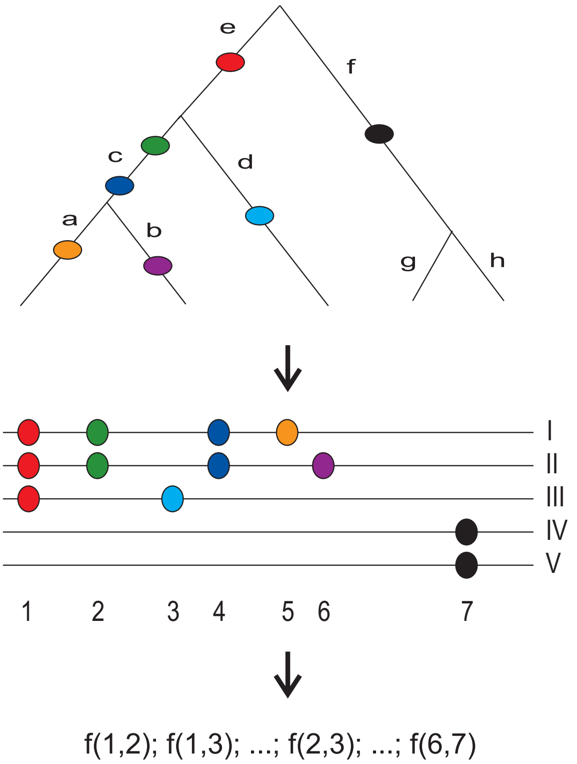


Figure 6.2: The basic principle behind the 2 point spectrum statistic. Going from top to bottom, a coalescent genealogy relates five sequences (I-V). Each mutation (colored circles) generates a new segregating site (1-7), with the amount of sequences affected being dependent on the position of the mutation in the genealogy. Using the commonly defined SFS, seven frequency values would be available. However, the final 2 point spectrum uses pairwise frequency information for all pairs of sites (denoted as function f(i,j) for sites i and j), resulting in 21 potentially informative frequency values. Increasing the number of informative values from the same data set, potentially increases the power for parameter inference. The initial idea of the 2 point spectrum (the sub-sfs method) would summarize the frequency information in a slightly different way, just focusing on sites from the same subtree (hence, only sites that are located on the same haplotypes are summarized in pairs, resulting in a reduction of the total number of pairs).

6.3 Implementation of the 2 point spectrum

As described in the previous section, the graphical output of the sub-sfs method was sufficient to get first insights into the behavior and information content of this data summary, but had limited capabilities to directly infer potential parameters of interest.

Approximate Bayesian Computation was the method of choice for AF-IBD/S. One reason for using ABC was the lack of a proper way to express the likelihood of a given data set, given a certain demographic model. This fact was easy to circumvent by using statistics summarizing a full data set by a number of representative values. However, in order to make use of the flexibility of MCMC approaches a likelihood function $P(D|\Theta)$ is indispensable. To the best of our knowledge two studies have contributed significant insights into this particular field of research. Hobolth and Wiuf obtain the joint and marginal sample frequency spectrum of two mutant alleles when the mutations are genealogically nested. They also obtain the age of the younger and older mutation [60]. Jenkins and Song [68] extended these results to non-nested mutations which is an important case, since, as they show, the probability of two mutations being nested approaches 1.0 with increasing sample size. Both studies investigate the effect of a second mutation on the sample frequency spectrum of a segregating site where the model of Wiuf and Hobolth can be included as a special case in the model of Jenkins and Song. Given a coalescent genealogy and two segregating sites (implying two mutations being analyzed at once), four possible topological events can be observed (see figure 6.3):

1. Two mutations are nested
2. Both mutations happen on the same branch
3. Two mutations are non-nested (excluding basal branches)
4. Both mutations happen on the same basal branch (the branches closest to the root of a genealogy)

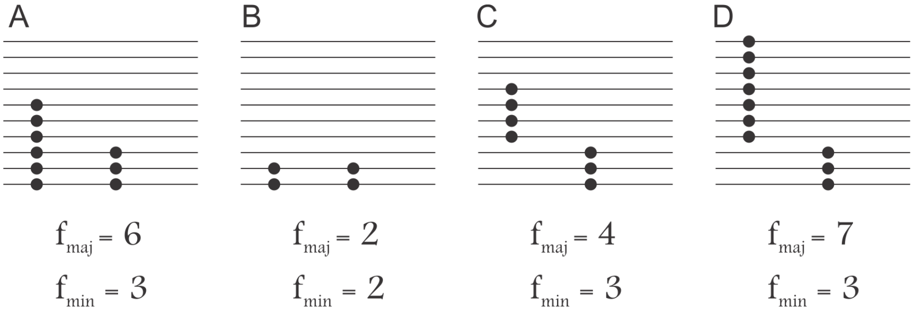


Figure 6.3: Four possible types of observation of two segregating sites. On the coalescent tree the two mutation events are either A) nested, B) on the same branch, C) non-nested, or D) on the two basal branches.

and authors in [68] were able to obtain their relative probabilities. However, their results only hold for models of constant population size over time. After personal communication, Jenkins and Song, in particular Paul Jenkins, were able to generalize the results of their study to account for deterministically varying population size. The following explanations and formulas are closely based on [66, 68]. Given an observed frequency configuration of a pair of segregating sites, let $\mathbf{n}$=(n_a,n_b,n_c) with n_a being the number of haplotypes carrying the ancestral allele, and n_b, n_c being the numbers of haplotypes carrying the older and younger derived alleles respectively, the total sample size is n = n_a+n_b+n_c. This configuration can be recored in the form (f_{maj},f_{min}), the sample count of the major (more common) derived allele, and the sample count of the minor (less common) derived allele. The four topological cases are represented by the following symbols: ($E_{2N}^{(b,c)}$) with $f_{min} < f_{maj} < n$, ($E_{2S}^{(b,c)}$) with $f_{min} = f_{maj} < n$, ($E_{2NN}^{(b,c)}$) with $f_{min} + f_{maj} < n$, and ($E_{2B}^{(b,c)}$) with $f_{min} + f_{maj} = n$ respectively (see listing above). As this classification is rather important for the understanding of this method, see figure 6.3 again for a graphical representation. Since we assume the infinite sites model, pairwise observations can only be of one of the two forms: (1) two different haplotypes carrying a single mutation are observed, with the set of haplotypes carrying one mutation disjoint from the set carrying the other mutation or (2) some singleton and some doubleton mutant haplotypes are observed, with the set of sequences carrying one mutation a subset of the other. Furthermore, let $\mathbf{T}$=(T_2, T_3, ...,T_n) be the random vector of inter-coalescence times in the coalescent genealogy (see figure 6.4 for a graphical representation). Times T_j are the time periods during which the genealogy has j=n,...,2 lineages respectively. The vector $\mathbf{T}$ fully describes the entire genealogy and summarizes the demographic effects on the coalescent tree quite well. Since in this setting the population size is a function of time, times T_j are not independent with $T_j \sim \exp\binom{j}{2}$ on the coalescent time scale anymore. Hence, they cannot easily be integrated out and are expressed in terms of the joint moments of $\mathbf{T}$.

$$E[N|E_{2N} \cup E_{2S}] = \begin{cases} \frac{1}{E} \sum_{k=3}^{n-f_{min}+1} \sum_{j=2}^{k-1} C_{j,k}^{(n-f_{maj}, f_{maj}-f_{min})} E[T_k T_j] & \text{if } f_{min} < f_{maj}, \\ \frac{1}{E} \sum_{k=2}^{n-f_{min}+1} \sum_{j=2}^{k} D_{j,k}^{(n-f_{maj})} E[T_k T_j] & \text{if } f_{min} = f_{maj}. \end{cases} \tag{6.1}$$

$$E[N|E_{2NN} \cup E_{2B}] = \begin{cases} \frac{1}{H}\sum_{k=3}^{n-f_{min}+1}\sum_{j=2}^{k}\frac{\left[F_{j,k}^{(n_a,f_{maj})}+F_{j,k}^{(n_a,f_{min})}\right]}{1+\delta_{f_{maj},f_{min}}}E[T_kT_j] & \text{if } f_{min}+f_{maj} < \text{n}, \\ \frac{1}{H}\sum_{k=2}^{f_{maj}+1}\frac{G_k^{(f_{min},f_{maj})}}{1+\delta_{f_{maj},f_{min}}}E[T_kT_j] & \text{if } f_{min}+f_{maj} = \text{n}. \end{cases} \tag{6.2}$$

where

$$\begin{aligned} C_{j,k}^{(n_a,n_b)} &= \sum_{l=j-1}^{k-2}\binom{n_a-1}{l-1}\binom{n_b-1}{k-l-2}\binom{k-j}{k-1-l}\binom{n-1}{k-1}^{-1}\binom{k-1}{k-l}^{-1}j(j-1), \\ D_{j,k}^{(n_a)} &= \binom{n_a-1}{k-2}\binom{n-1}{k-1}^{-1}\frac{j(j-1)}{k-1}\frac{1}{1+\delta_{j,k}}, \\ E &= \sum_{k=3}^{n}\sum_{j=2}^{k-1}j\left[k\binom{k-2}{j-1}-(j-1)\binom{k-1}{j}\right]\binom{k-1}{j-1}^{-1}E[T_kT_j]+\sum_{k=2}^{n}\sum_{j=2}^{k}\frac{j(j-1)}{k-1}\frac{E[T_kT_j]}{1+\delta_{j,k}}, \\ F_{j,k}^{(n_a,n_b)} &= \sum_{l=(j-2)\vee 1}\binom{n_a-1}{l-1}\binom{n_b-1}{k-l-2}\binom{k-j}{k-2-l}\binom{n-1}{k-1}^{-1}\frac{j(j-1)}{1+\delta_{j,k}}, \\ G_k^{(n_b,n_c)} &= \left[\binom{n_b-1}{k-2}+\binom{n_c-1}{k-2}\right]\binom{n-1}{k-1}^{-1}\frac{2}{k-1}\frac{1}{1+\delta_{k,2}} \end{aligned} \tag{6.3}$$

Provided the population size N_e(t) is such that the moments E[T_jT_k] can be calculated, equations 6.1 and 6.2 provide the likelihood of an observed pair of sites under a given demographic model without the need to simulate a full coalescent genealogy. Equations 6.1, 6.2 and 6.3 are provided by Paul Jenkins ([66]).

6.4 Obtaining the joint moments

In order to obtain the joint moments for pairs of inter-coalescence times E[T_jT_k], given a demographic model of interest, two possible approaches are available. The first is by Monte Carlo simulations to obtain tuples (T_2, T_3,...,T_n) of inter-coalescence times. This process must not be seen as the classical way of generating a full genealogy (see section 3.3.2 for explanation), but more as drawing random numbers from an exponential distribution taking the current number of lineages and population size N(t) into account. The following lines will explain how this process can be implemented.

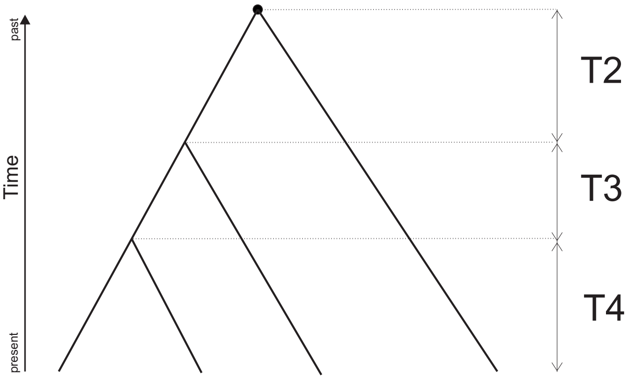


Figure 6.4: Shown is a coalescent genealogy relating four individuals. The times between the coalescent events are the inter-coalescence times v=(T_4,T_3,T_2) (from present to past). The black circle denotes the *MRCA*.

Every time a new demographic model is proposed, a new set of inter-coalescence times needs to be calculated. Following algorithm 3.3.2, after drawing random numbers from an exponential distribution, times need to be modified according to the population size at time t. Hence, for each demographic model M multiple instances x of (T_2, T_3,...,T_n) are generated. These tuples are then directly used to obtain E[T_jT_k] for all possible pairs of inter-coalescence times. Within this setting tuples are stored in a 2-dimensional array A with n-1 columns and x rows. Hence, T_j is a vector of length x, denoted v_j=[T_j^1,...,T_j^x]. Given that, the joint moment E[T_jT_k] is calculated as the mean of the inner product of the two vectors v_j and v_k:

$$E[T_jT_k] = \frac{\sum_{i=1}^{x} v_j[i]v_k[i]}{x}, \tag{6.4}$$

However, averaging over a certain number x of tuples is only an approximation to the true E[T_jT_k]. Therefore, second order moments approximated with this method are denoted $E'[T_jT_k]$ from now on. As a consequence, calculating the second order moments for the exact same demographic model can result in two slightly different outcomes. The number of tuples x should be chosen large enough in order to get a more stable approximation (see figure 6.6 for more details).

The second possibility is to directly calculate the joint moments without the need to draw times from an exponential distribution and generate tuples. To the best of our knowledge there exist two studies that have investigated this problem, Polanski et al.[110] and Živković et al.[136]. One of the main differences between the two studies is that [110] define T_k, k=2,3,...,n as coalescence times from sample of size n to sample of size k-1, whereas [136] uses the same definition as figure 6.4. Živković et al. derive all second order moments by using conditional expectations and all details are well explained in their supplementary material. Given these theoretical results it is possible to derive a direct expression to calculate E$[T_jT_k]$ (adapted from [136]):

$$\begin{aligned}\alpha_{n,j,k} &= \frac{(2j-1)n!(n-1)!(k+j-2)!}{(j-k)!k!(k-1)!(n-j)!(n+j-1)!} \\ g_{j,i}(t,t') &= \frac{\binom{j}{2}\binom{i}{2}}{\lambda(t)\lambda(t+t')} exp\left\{-\binom{j}{2}\int_0^t \frac{1}{\lambda(u)}du\right\} exp\left\{-\binom{i}{2}\int_t^{t+t'} \frac{1}{\lambda(u)}du\right\}\end{aligned} \tag{6.5}$$

$$E(T_k^2) = \begin{cases} \sum_{j=k+1}^{n}(-1)^{j+k+1}\frac{\binom{k+1}{2}}{\binom{j}{2}}\alpha_{n,j,k+1}\int_0^\infty\int_0^\infty t'^2 g_{j,k}(t,t')dtdt', & \text{if } 2 \leq \text{k} \leq \text{n-1} \\ \int_0^\infty t^2 g_n(t)dt & \text{if k=n} \end{cases} \tag{6.6}$$

$$E(T_{k'}T_k) = \sum_{j=k'}^{n}\sum_{i=k}^{k'}(-1)^{i+j+k+k'}\frac{\binom{j}{2}-\binom{i}{2}}{\binom{j}{2}}\alpha_{n,j,k'}\alpha_{k',i,k}\int_0^\infty\int_0^\infty tt'g_{j,i}(t,t')dtdt' \tag{6.7}$$

Following the notation of [136] and equations 8.3 and 6.5, for the mean waiting times E$[T_j]$, the first waiting time T_n is not dependent on the following waiting times, so the density function g(t_n) is given within formula 8.3, when setting j=n. First, the expectation E$[T_n]$ of time T_n is calculated and n is set to j. Then E$[T_j]$ can be iterated via a harmonic sum through the respective coalescent events to finally get the expression for all waiting times. Regarding the calculation of the expectation of the product of two waiting times, the joint density of T_n and T_{n-1}, g$(t_n$,$t_{n-1})$ does not depend on the subsequent waiting times. Hence, when setting j=n and i=n-1, g$(t_n$,$t_{n-1})$ is directly given by equation 8.5. Simply speaking, after calculating E$[T_nT_{n-1}]$, n is set to j and n-1 to i and one then iterates E$[T_jT_i]$ via the double harmonic sum and the respective coalescent events to finally get the expression for the products of all waiting times. Iterating over the respective coalescent events can be explained as follows: For each demographic model with a finite number of m intervals, each with a different constant population size, equation 6.7 can be explicitly solved by decomposing the double integrals according to the $\binom{m+1}{2}$ arrangements of coalescent events over these m time periods. As can be seen in figure 6.5, for a simple bottleneck model with m=3 intervals (reducing the population size from 1 to f), six possible arrangements have to be taken into account to calculate the joint density of two coalescent times. Although the example shows a genealogy with sample size n=3 and only the two coalescent times T_3 and T_2, increasing the sample size would not change the essence of this figure. The two coalescent times can be any pair of times within a full genealogy with arbitrary sample size n.

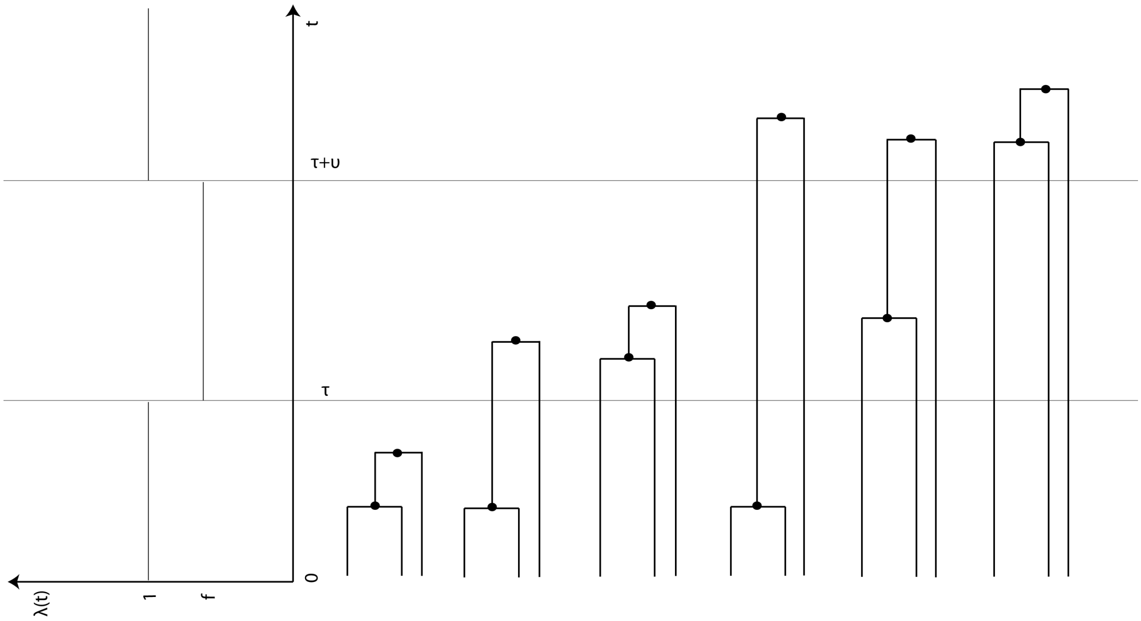


Figure 6.5: Possible coalescent trees for a three-phase bottleneck model, temporarily reducing the population size from 1 to f. $\lambda(t) = N(t)/N$. Assuming two coalescent events and m=3 intervals, results in six possible arrangements.

I used Mathematica[1] to simplify the formulas for the following pairwise coalescent cases:

1. Both events occur in the first interval
2. Both events occur in the last interval
3. Both events occur in a middle interval
4. One event occurs in the first, the other event in the last interval
5. One event occurs in a middle interval, the other in the last interval
6. One event occurs in the first interval, the other in a middle interval
7. Both events occur in different middle intervals

With the simplified expressions for each case all possible arrangements for any demographic model can be calculated (due to space constraints I omit the actual notation of the expressions, but they can be obtained on request). The same principle is applied for calculating $\mathrm{E}[T_j^2]$ (see equation 6.6). With the steps from the above listing, $E[T_jT_k]$ can be calculated for (almost) any desired distribution of past population size changes, obviating the need to specify a separate formula for every number m of N_e intervals.

---

[1] http://www.wolfram.com/mathematica/, last visited on 03/12/2013

The Monte Carlo approach to approximate $E'[T_j T_k]$ was used for the first tests and for the initial investigation of this method, since only the study of Polanski et al. was known to us at the beginning. However, after personal communication, the Polanski et al. method could not efficiently be applied in our practical computations for larger sample sizes and the fact, that their definition of T_j differed from ours (see previous explanation) complicated the use significantly. Therefore, the simulation approach that averages over a large number of values was used in the first place. Furthermore, obtaining the second order moments $E'[T_j T_k]$ can be efficiently implemented allowing for a large number of steps the MCMC chain can be run in an appropriate amount of time.

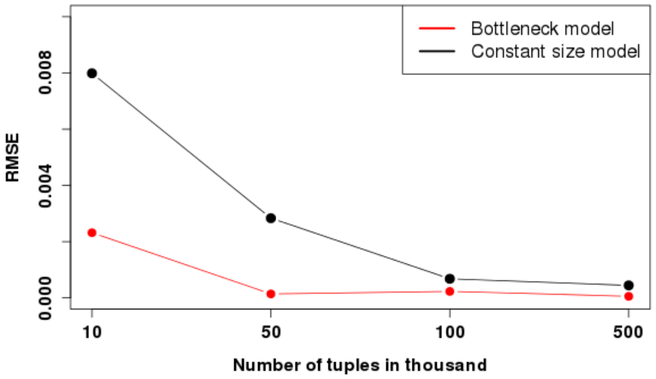


Figure 6.6: Statistical error due to approximating first- and second order moments of inter-coalescence times using simulations. For a constant population size (black) and a model with m=3 intervals (bottleneck model, red) the plots show the RMSE between the moments approximated by Monte Carlo simulations and moments calculated by using the closed form from [136].

Figure 6.6 shows the statistical difference (calculated as the RMSE) between the approximated and exact second order moments of inter-coalescence times for a constant size and a recent bottleneck model. Because this figure illustrates that using x=10^5 tuples results in negligibly small divergences from the true (closed form) $\mathrm{E}[T_j T_k]$ and further increasing this number does not result in a significant reduction of the RMSE, x is set to this value for the rest of this work.

6.5 MCMC approach

The previous sections introduced all theoretical means necessary to incorporate the setting into a MCMC framework. Starting with the new idea of the 2 point spectrum and its calculation from observed polymorphism or sequence data to the derivation of second order moments of pairs of inter-coalescence times, calculated from a full expression or as an approximation, to the possibility to calculate the likelihood of a pair of sites given a demographic model. The overall question is how to combine all aspects to finally infer parameters of interest from observed data sets.

As already mentioned, Markov Chain Monte Carlo methods are a class of algorithms that allow the sampling from probability distributions based on constructing a Markov chain having the distribution of interest as its equilibrium distribution. Our interest is a demographic model comprised of an unknown number of intervals that can potentially explain the observed data. The current section will explain the steps that lead to the realization of such a Markov chain.

See figure 6.5 for a graphical example of the current setting of the demographic model that is to be inferred. Since the overall aim is to allow a high flexibility without being limited to a certain model (i.e. a certain number of N_e intervals), the number of intervals, each with a different constant N_e, should be a part of the actual inference process. This fact complicates the whole framework. Classical Metropolis-Hastings MCMC approaches have an a priori fixed number of dimensions, i.e. the number of model parameters that need to be inferred is known beforehand and, therefore, are less suitable for Bayesian model determination problems. Green [45] has suggested a solution to this problem as he proposed a new framework for the construction of reversible Markov chain samplers that jump between parameter subspaces of differing dimensions. The name 'reversible jump' comes from the algorithm's ability to change the dimensions of the state space in a move known as a reversible dimension jump, or *reversible jump*. The shortcut of this extension to the classical approaches is *rjMCMC*. As the name implies, jumps between dimensions need to be reversible, i.e. it must be possible to revert back to the previous state in a later move.

6.5.1 RJMCMC overview

The Metropolis-Hastings-Green method according to [45] was used. Briefly, compared to MH-MCMC (see section 2.4.3), rjMCMC has a slightly different procedure of single steps that need to be considered. The algorithm starts in an arbitrary configuration known from the previous time step. Then a move type from a set of reversible moves is selected. Then a target object is chosen and the selected move is applied to generate a newly proposed configuration. In an acceptance test the new configuration needs to be evaluated. Based on the outcome of this test the previous or the new state is then added to the Markov chain. The following decisions about the detailed implementation of the Markov chain are mostly based on [45, 107].

Let x be the initial state of the chain and $\tilde{x}$ a newly proposed state with proposal density $q(\tilde{x})$. As explained in chapter 2.4.3, the acceptance ratio α of the MH-MCMC is a ratio of the likelihood of the newly proposed configuration to the likelihood of the previous configuration. However, since certain moves may change the number of model parameters, the previous and proposed configurations can be of different dimensions. Obviously, calculating the same ratio with these likelihoods would be pointless. Therefore, the acceptance rate is written as:

$$\alpha(x, \tilde{x}) = min\left\{1, L \cdot A \cdot P \cdot J\right\}, \tag{6.8}$$

with L being the likelihood ratio $P(D|\tilde{x})/P(D|x)$, A being the prior ratio $P(\tilde{x})/P(x)$, P the proposal ratio $q(\tilde{x})/q(x)$ and J being the determinant of the Jacobian that results from the potential change of dimension. In order to achieve a parameter inference based on MCMC, several components need to be specified:

1. A proper representation of the estimated demographic function $N_e(t)$
2. A prior distribution that specifies the a-priori knowledge for each parameter
3. A likelihood function
4. Conditions and rules to construct the Markov chain (e.g. acceptance probabilities, move types etc.)

The *demographic function* $N_e(t)$ is represented as a piecewise constant function. It consists of m distinct intervals with the first interval starting at position a_0=0 (time point 0, present day) and has a N_e of N_{e0} (representing the height of this interval). This first interval is then followed by k internal positions at (a_1,N_{e1}), (a_2,N_{e2}), ..., (a_k,N_{ek}) and the terminal node (a_{k+1},N_{ek+1}). If representing $N_e(t)$ as a spline, it is defined for all t∈ [0,T], with $T = \sum_{i=0}^{m} l_i$, with m=k+2 and l_i being the length of interval i.

The *likelihood function* L was described before (see equations 6.1 and 6.2). It depends on the allele frequency configurations of each pair of fully linked segregating sites in a genomic region with a zero-recombination rate. Since $N_e(t)$ is represented as a piecewise constant function, L can be calculated efficiently (with L being the likelihood ration, l' the likelihood of the newly proposed state and l the likelihood of the current state). Furthermore, for computational reasons, likelihoods are calculated on a log scale resulting in the likelihood ratio L=log(l')-log(l).

Choosing the *prior distributions* is a crucial step in Bayesian methods. As already mentioned and discussed in section 2.4.1, priors represent any knowledge about a specific parameter of interest prior to the actual inference process. In [107] a similar approach is used to infer past population size changes using a rjMCMC framework based on the product of densities of the waiting times between subsequent coalescent events. Since the final question and setting of their method is to a high degree very similar to ours, I chose to adopt the prior choices of [45] and [107] (I did validate that parameter and prior choices closely follow the principles of the original rjMCMC framework). The prior for the number of change points (i.e. representing the number of distinct N_e intervals) is represented as a truncated Poisson-distribution for m:

$$P(k) = \begin{cases} \frac{1}{c}\frac{\lambda^k}{k!}e^{-\lambda} & \text{for k} \leq k_{max} \\ 0 & otherwise, \end{cases} \tag{6.9}$$

where c is a normalizing constant to ensure that P(k) is a proper distribution. The maximum number of intervals (m_{max}) can be specified by the user, keeping in mind that increasing the number increases the computational burden for the calculation of $E[T_k T_j]$ when using the exact method from [136]. The smoothing parameter λ is set between 0.1 and 1.0.

The *positions* of the starting points of N_e intervals (inner nodes of the spline) are uniformly distributed on the interval [0,T]. Positions can later be modified if a specific move type is applied, thereby extending or reducing the length of an interval (i.e. shifting the start position) in either direction.

For the *height* N_{ei} of interval i a Gamma distribution Gamma($N_{ei}|\alpha_i$, β_i) is assumed, which ensures that the sampled heights are always positive, with α_i and β_i being the mean and variance of size N_{ei}.

In order to construct a Markov chain with the mentioned demographic function, likelihood functions, and prior distributions, the following reversible move types, designed to explore a variable-dimensional state-space, were chosen:

1. Update. Either increase or decrease the N_e in interval i
2. Extend. Extend the length of interval i to the left or right by some random number of generations
3. Join (death step). Join block i and i+1, thereby decreasing the total number of intervals by one
4. Split (birth step). Split interval i into two intervals at a random position, thereby increasing the total number of intervals by one

According to [45], the probabilities for the birth and death moves need to be synchronized to ensure a detailed balance of the Markov chain. Let η_m, π_m, b_m and

d_m the probabilities of the four move types, with $\eta_m + \pi_m + b_m + d_m = 1$. The synchronization can be achieved by:

$$b_k = c \cdot min\left\{1, \frac{P(k+1)}{P(k)}\right\} \tag{6.10}$$

and

$$d_{k+1} = c \cdot min\left\{1, \frac{P(k)}{P(k+1)}\right\} \tag{6.11}$$

with c chosen so that $b_k + d_k < 0.9$ for all k. The remaining section explains the respective methods to propose and accept one of the mentioned steps.

Update the N_e of an interval is done by first choosing an interval i out of the m existing intervals with probability $\frac{1}{m}$. Second, a new size is proposed by $N'_e = N_e \cdot exp(z)$ where z is a uniformly distributed random variable on $\left[-\frac{1}{3}, \frac{1}{3}\right]$. The new N_e is then accepted with probability

$$\alpha_U(x, \tilde{x}) = min\left\{1, L + (log(N'_e - N_e)^{\alpha}) + log(exp\left(-\beta\left(N'_e - N_e\right)\right))\right\} \tag{6.12}$$

Extending the length of an interval is done by first choosing an interval i with probability $\frac{1}{m}$. The new start position a'_i is chosen uniformly from $U \sim [a_{i-1}, a_{i+1}]$, resulting in the starting position being shifted to either side with probability $\frac{1}{2}$. The newly proposed step is then accepted with probability

$$\alpha_E(x, \tilde{x}) = min\left\{1, L + log[a_{i+1} - a'_i] - log[a_{i+1} - a_i] + log[a'_i - a_{i-1}] - log[a_i - a_{i-1}]\right\} \tag{6.13}$$

Split an interval into two (birth step) is done by first choosing an interval i out of the m existing intervals with probability $\frac{1}{m}$. Second, the new change point (start of newly generated interval) is denoted as a^*, which is between a_i and a_{i+1}. The new N_e that corresponds to a^*, is N^*_e and is generated by randomly modifying the current $N_e(a^*)$ on position a^*, according to $N_e(a^*)$+z$N_e(a^*)$, with z being a random variable, uniformly distributed on the interval $\left[-\frac{1}{3}, \frac{1}{3}\right]$. This step increases the number of intervals by one. The newly proposed step is then accepted with probability

$$\alpha_S(x, \tilde{x}) = min\left\{1, L + log[k+1] + log\left[\frac{(a^* - a_i)(N_{ei+1} - N_{ei})}{a_{i+1} - a_i} + N_{ei}\right]\right\} \tag{6.14}$$

Join an interval with its right neighbor (death step) is done by first choosing an interval i out of the m existing intervals with probability $\frac{1}{m}$. This represents the inversion of the birth step and consists of removing an interval. This step decreases the number of intervals by one. The newly proposed step is then accepted with the inverted probability of the birth step

$$\alpha_J(x, \tilde{x}) = -\alpha_S \tag{6.15}$$

The following algorithm briefly summarizes the necessary steps:

1. Begin with the state of the previous sample x = x_{n-1}
2. Select a move type from the set of reversible move types (by sampling from a distribution p_{move_type})
3. Apply the selected move by selecting a target object and proposing a new configuration $\tilde{x}$. Choosing the target i^* is done through a move-specific target proposal distribution q_{target}, and proposing $\tilde{x}$ is done through a move-specific proposal distribution Q
4. Compute the acceptance ratio, α, taking into account that it is defined differently for the various move types
5. Add the nth sample to the chain. If $\alpha \geq 0$, add the proposed configuration $\tilde{x}$. Otherwise add the proposed configuration with probability 10^{α}. If the proposed configuration is rejected, add the previous configuration x_{n-1}

6.6 Chain length

One of the most difficult steps of applying MCMC algorithms is to determine how long it takes for the chain to reach its stationary distribution (also called limiting distribution in a Markov chain). If the chain has not been run long enough, it may not give a good approximation of the target distribution (the distribution of interest). The number of steps it takes to reach that state is called mixing time. Mixing is one out of many topics of MCMC convergence and is a field of research of its own. Metaphorically speaking, it is directly connected with the speed of forgetting the initial value or distribution of the Markov chain. For a brief summary of possible statistical analyses see [10]. Although several authors have shown practical implications of these analyses, applying suggested tests does not guarantee satisfactory results and they often remain difficult to use and not reliable enough. Hence, in practice it is often advisable to experimentally determine an adequate chain length. One commonly applied method is to run multiple parallel chains. When convergence is slow this can be a serious practical limitation. Multiple chains increase the computational complexity, but can be very useful to diagnose non-ideal convergence behavior. For example, each chain may individually appear to have converged, but comparisons between them may uncover discrepancies in the apparent stationary distributions.

6.7 Performance optimization

The whole Markov chain algorithm is implemented in C, a general purpose programming language[2] and compiled with gcc version 4.6.3, the GNU compiler collection, on a x86_64 Linux environment. Furthermore, functions provided by GLib were used. GLib is a library written in C consisting of a variety of additional functionalities and data structures[3].

In order to get an estimate of $E'[T_jT_k]$, the calculation of $(n * (n-1))/2$ pairs needs to be done after every newly proposed step, and the Markov chain is usually run for several ten to hundred thousands of steps, which makes computational speed optimization an important point. For that reason the calculation of the inner products is implemented with the use of *intrinsic functions* from the Intel Streaming SIMD Extensions technology (SSE). Among others, this set of functions is often used to optimize the performance of an algorithm by vectorization where a computer program is converted from a scalar implementation, which processes a single pair of operands at a time, to a vector implementation which processes one operation on multiple pairs of operands at once. The compiler has a direct knowledge of the intrinsic function and can, therefore, better integrate it and optimize it for the problem at hand. The 128 byte $\_\_m128$ data type I used in this setting is able to hold four floating point values at once, reducing the time to calculate the scalar product by a factor of 4. Therefore, SSE is a set of instructions which allow to load the floating-point numbers to 128-bit registers, perform the arithmetic and logical operations, and write the result back to memory. SSE was introduced by Intel in 1999 for the x86 architecture and has ever since been improved and refined. Using this method in a program that is supposed to later be used on a variety of different computer systems is not problematic, since all modern Intel and also AMD CPUs support SSE, after AMD gave up on their own instruction set called 3Dnow. Various websites provide more detailed information about the use and implementation of this method (e.g. the Intel website[4]). Functions provided by the C header file emmintrin.h were used.

The second way of calculating $E[T_jT_k]$ can be achieved without averaging over a large number of simulated inter-coalescence times. However, the practical application of this approach contains some points that need special attention. The given expressions include, among others, the summation of both very large and very small positive and negative terms which leads to a numerical instable behavior of the calculation. Therefore, with increasing sample size n or increasing number of N_e intervals m results get less and less reliable. Even for a rather small setting with

---

[2] see http://cm.bell-labs.com/cm/cs/who/dmr/chist.html, last visited on 03/12/13

[3] see https://developer.gnome.org/glib/, last visited on 03/12/13

[4] http://software.intel.com/en-us/articles/using-intel-streaming-simd-extensions-and-intel-integrated-performance-primitives-to-accelerate-algorithms, last visited on 14/10/2013

$n = 10$ and $m = 6$, problems already emerge. For example, $E[T_{24}^2]$ with $n = 40$ and $m = 3$ (with a specific $N_e(t)$ function) results in $-1.3 \cdot 10^{-7}$, when calculated with standard C data types, but its true value is $2.678 \cdot 10^{-9}$. Hence, using the standard C floating point data types *double* and *long double*, whose implementations range between 64 and 128 bit precision, depending on the compiler and underlying computer architecture, is not sufficient to handle these calculations. I, therefore, decided to apply an extended precision library. Extended precision refers to floating point number formats that provide greater precision and more exponent range than the basic floating point formats. The *MPFR* library is a C library for multiple-precision floating-point computations with correct rounding (see http://www.mpfr.org/ [5] for more details). The basic principle of these libraries is that precision is not limited by the provided data types, but only by the available amount of machine memory. At the cost of a higher calculation time any desired precision can be achieved. The syntax of most libraries differs significantly from the usual C syntax. For example, an mpfr_t object needs to be initialized with $mpfr\_init(mpfr\_tx)$ before storing the first value in it. Furthermore, functions to assign values $mpfr\_set\_()$ and arithmetic functions like $mpfr\_sub()$, $mpfr\_mul()$, etc., exist to allow the application of any desired arithmetic. The set of provided functions is complex, however, easy to use for reasonably short expressions. However, trying to transform the aforementioned approach turned out to be practically unfeasible. Though, a number of interfaces and extensions exist for *MPFR*. An MPFR *C++* wrapper, written by Pavel Holoborodko (see *http://www.holoborodko.com/pavel/mpfr*[6]), is based on the *MPFR* and uses the possibility of operator overloading in C++ to replace $mpfr\_mul()$ with the commonly used operator $*$. This can be achieved by transferring the actual calculations to a C++ script and later linking both the C and C++ implementations together. The speed of this method is then much more dependent on the actual sample size n and number of N_e intervals m. With increasing numbers calculations get computationally more intensive and time consuming. It is, therefore, currently only feasible to use data sets with up to 5 diploid individuals (10 haplotypes) within the MCMC framework when using $E[T_jT_k]$. C header file gmp.h and the C++ header file mpreal.h were used.

In order to improve the runtime and hence the overall performance of the algorithm I also implemented multi-threading, a widespread programming and execution model that allows multiple threads to exist within the context of a single process. The resources of the program are shared but threads are able to execute independently. Users with the access to computers with mutli core processors or multiple CPUs benefit from this implementation, since it results in a significant reduction of computational time. Parts that can be parallelized by multiple threads are the calculation of the likelihood (equation 6.1), the calculation of $E'[T_jT_k]$ (equation

---

[5] last visited on 10/13/2013

[6] last visited on 10/13/2013

6.17) as well as the simulation of the inter-coalescence time matrix.

All mentioned steps were extensively tested and the current set of optimization steps turned out to be the most efficient algorithm so far. This implementation provides a significant reduction in computational time and overall accuracy.

6.8 Parameters and options

The software comes as a binary file that can run on 32 and 64 bit machines. It's an easy to use command line program without any graphical user interface. Although the rjMCMC approach does not have a fixed number of model parameters to allow for a high flexibility in terms of demographic inference, the user has to specify a number of parameters to control the behavior and performance of the chain. The following section will briefly introduce the parameters and their effects.

Parameters that need to be specified:

1. -file = A file containing the information about the observed population of interest, given in vector format
2. -nhaplo = A positive integer number representing the number of haplotypes (sample size). For example, 5 diploid individuals are 10 haplotypes
3. -mmethod = The method used for the calculation of $E'[T_jT_k]$ (approx) for the approximation method, or $E[T_jT_k]$ (exact) for the calculation based on [136]
4. -threads = A positive integer number representing the number of parallel threads used for the program
5. -n_time_sims = A positive integer number representing the number x of tuples of inter-coalescence times (only needed if -mmethod=approx)
6. -nsteps = A positive integer number representing the number of steps the chain should be run
7. -min_int_size = Minimum length a block can have (effects the smoothness of the final trajectory, default 0.0125)
8. -min_ne = Minimum N_e a block can have (default 0.01)
9. -max_ne = Maximum N_e a block can have (default 20)
10. max_num_blocks = The maximum number of N_e blocks allowed (default 10)
11. update_ne_interval = A number giving the range of z (see Update step, default is [-0.3,0.3])

6.9 Likelihood calculation and data sets

Since the current implementation of the 2 point spectrum does not take recombination into account, segregating sites that are analyzed need to be connected by a single coalescent genealogy, which would not be possible if recombination events happened between the sites. Therefore, it is currently not possible to analyze continuous autosomal chromosomes or entire genomes. For that reason genetic regions need to be chosen that are known to have a very low or no recombination rate at all. This is practically feasible, since fine scale recombination maps are available for an increasing number of species of interest (e.g. [50], etc.). In order to decrease the chance for recombination events to happen, the length of the chosen regions must be kept rather short. We decided to use regions not longer than 50 SNPs of length. Since the data under consideration, therefore, consists of independent recombination free regions, the likelihood can not be calculated for the whole data set at once, but each region (that is assumed to be independent from the remaining regions) needs to be calculated separately. Hence, the likelihood is calculated as a pseudo (or composite) likelihood. In many practical applications the joint distribution of the data may be difficult to evaluate or the data consists of smaller subsets whose internal dependencies are complex and difficult to specify. The basic principle of the class of pseudo likelihoods is that if computing likelihoods for certain subsets of the data is possible, a pseudo likelihood can be constructed by combining single likelihood objects using them as a surrogate for the ordinary likelihood. Often the computing cost of calculating the full-likelihood increases rapidly, even exponentially with the sequence length and hence, the composite likelihood takes substantially less time to calculate than the full-likelihood. A quite comprehensive discussion about the accuracy and potential disadvantages can be found in [4, 34, 90]. The composite likelihood implemented in the current work can be calculated as follows:

$$CL(\theta|D) = \Pi_{k\in K} L_k(\theta|D) \tag{6.16}$$

Therefore, all pseudo observed data used for the initial analysis of the 2 point spectrum method were simulated with *ms* with a large number of independent replicates (k) based on the same underlying demographic model. The replicates represent the genomic regions in empirical data sets. Mutation rate was chosen so that each replicate consisted of roughly 50 SNPs.

The simulated sequence output from ms is then summarized by a Perl script to obtain the input format for the Markov chain program (denoted vector format). Given n segregating sites per independent ms replicate, the allele frequency information for each of the $n(n-1)/2$ possible pairs need to be summarized. This results in a format that gives the pairwise information for each pair of sites (from each replicate) in a vector $v = (f_{maj}, f_{min}, model, n_a, n_b, n_c)$ (see section 6.3) with model=1 representing the cases for E_{2N} (two mutations being nested) and E_{2B} (two mutations located on basal branches). Model=2 represents the cases for E_{2NN} (two mutations are non-nested) and E_{2S} (two mutations are located on the same branch).

6.10 Results 2 point spectrum

I first tested whether single demographic events have an effect on the sub-sfs (the initial implementation of the 2 point spectrum) statistic. As described in section 6.2 I simulated different demographic histories and compared them to a constant size model. Figure 6.7A shows how the ratio between different demographic models, represented by a triangular matrix, is affected. The first important information is the data range, showing the extent of the ratio of the sub-sfs values for two different demographic models. The according heat plot color code is depicted by a transition from blue to light brown/white to dark brown, representing values being < 1, ~ 1 and > 1 respectively. First of all, the ratio matrix for two similar but independently simulated demographic models is investigated. Despite the random nature of the coalescent, the distribution of sub-sfs values, calculated from a large number of coalescent trees, should be rather similar, since genealogies were generated under a model with the same demographic parameters. As shown in figure 6.7A, the comparison between 2 constant size models shows no clear color pattern, but more the impression of random noise. As expected, the range of values is rather small, between 0.92 and 1.09, indicating a negligible deviation from the value of 1. When comparing a non-constant with a constant size model the difference should be more pronounced, since changes in population size directly affect the branch lengths of the underlying genealogy. Therefore, figure 6.7B shows the comparison between a constant size and a bottleneck model (event happened 0.05 - 0.0525 time units ago, decreasing the population size by a factor of 100). Times of demographic events are from now on given in units of $4N_e$ generations. In order to obtain the actual number of generations this number needs to be multiplied by $4 * N_e$. First of all there is a distinct color pattern, indicated by a dark blue vertical area, representing the first five to six low frequency bins on the x axis (sub frequencies f_x). In addition one can clearly see a dark to light brown cluster at the upper third of the plot. Since the comparison is done by contrasting the constant model matrix by the bottleneck model matrix, the blue and brown clusters represent values that, compared to the constant size model, are bigger and smaller respectively in the bottleneck model. In panel C of figure 6.7 the same analysis is shown for a more ancient bottleneck ranging from 0.2 to 0.2025 time units ago. It is interesting to see, that the blue area slightly moved to the right, suggesting that the first column (F=1) is now affected in the opposite way (values >1). Further on, a similar brown cluster is seen but its location slightly moved downwards to higher F and f_x. As already mentioned, results here are only a subset of all the cases that were tested. For other bottleneck and expansion models similar effects could be observed. Interestingly the time differences between different events also seem to be reflected to a certain extent.

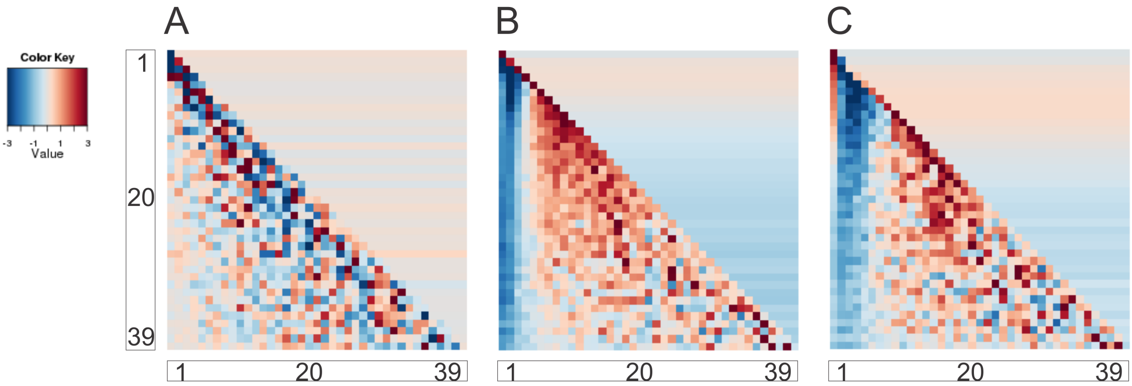


Figure 6.7: First tests of the sub-sfs method (the initial implementation of the 2 point spectrum) to analyze how sensitive this statistic is to past population size changes. A) shows the ratio between the sub-sfs entries of two independently simulated constant size models. B and C) show the same ratio between a constant size and a recent and ancient bottleneck event, respectively. Range of values is shown by the color key. Data was simulated for 40 haplotypes. Y-axis values represent frequencies F=1..39 and x-axis values represent frequencies f=1..39.

Since these initial results indicated that the pairwise frequency information seems sensitive to past population size changes to some degree, we analyzed whether it is possible to distinguish the two demographic histories shown in [101] (see methods for implementation details). Although being quite different, both demographic models result in the exact same site frequency spectrum (shown analytically by Myers et al.). After simulating two datasets, each based on one of the models, the SFS was calculated. As can be seen from figure 6.8, the SFS ratio between the constant and non-constant demographic histories, simulated for 20 diploid individuals, is rather stable around a value of 1.0. The 1% difference is due to the fact that the continuous N_e function was approximated as a discrete piecewise constant function with 5,000 data points (provided by Nick Patterson). This indeed indicated that the two models cannot be distinguished just by using the SFS alone. However, analyzing the difference between the (sub-sfs) pairwise frequency configurations revealed a subtle pattern. Figure 6.9 shows that especially the lower frequencies F show differences between the two models. Interestingly the range of values is between 0.79 and 1.37, which is clearly different from the previously tested constant size models. In conclusion there seems to be a slight tendency for the pairwise frequency configurations to be able to capture past population size events. However, as already mentioned, interpreting these results just based on the graphical output is rather difficult. Since the outcome of these calculations can be represented as pairs of allele frequencies, the values could easily be used as summary statistics and later be used in an ABC framework as previously done for the AF-IBD method. Though, this would imply being limited to a certain fixed number of model parameters, which is a useful and valid approach, but as I mentioned, I wanted to exploit the capabilities and flexibility of a variable number of model parameters.

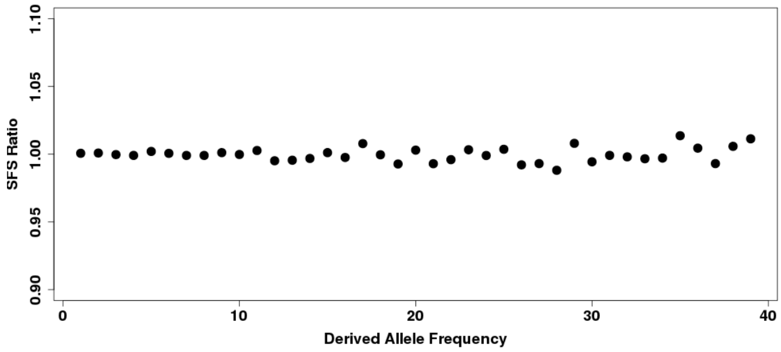


Figure 6.8: Shown is the ratio between the calculated site frequency spectra of the two simulated demographies from Myers et al. [101]. Both demographies, although clearly different from each other, are supposed to show the exact same SFS.

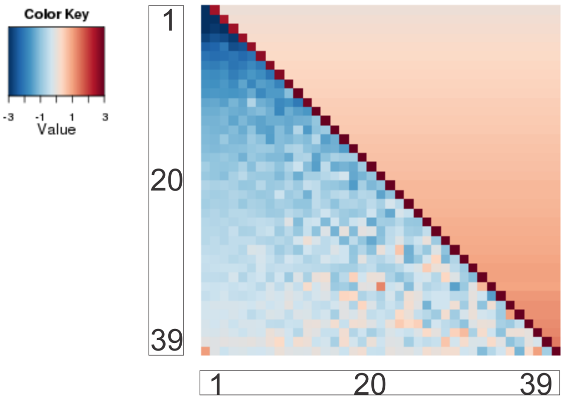


Figure 6.9: Similar setting as described in figure 6.7, but results are shown for the ratio between the constant and non-constant model, simulated from 5,000 data points (as given in Myers et al. [101]).

6.10.1 RJMCMC results

As already described, an rjMCMC algorithm implemented in the C language was used to investigate the potential of the 2 point spectrum method to distinguish different demographic scenarios and to infer underlying parameters of interest (again, now considering all possible pairs of sites). The first analyses were done by using the Monte Carlo method to calculate the approximated moments of pairs of

inter-coalescence times $E`[T_j T_k]$. I started by analyzing the simplest models, i.e. a population size constant over time. In the standard infinite sites model the frequency spectrum for a constant population size, where frequency j has a probability proportional to 1/j, is independent of the actual value of N_e. A change in population size rescales all branches in such a way that the frequency spectrum remains unaffected. As an example, the way that the SFS from constant model of N_e differs from the SFS of a model of $6 * N_e$ is in the total number of mutant sites (i.e. the branch lengths are stretched by a factor of 6) that we expect to see, not the relative frequencies of their allele counts. Since the current method is essentially conditioning on this, it is discarding that part of the data which is informative about the absolute value of N_e. The current method would result in the same likelihood for two constant size models that differ in the absolute value of N_e. Therefore, this method could later be used to infer $N_e(t)$ up to a scaling constant which could be chosen based on the total number of mutations in the data. One option could be Tajima‘s diversity estimator, which states that the diversity is an estimate of $\Theta = 4{*}N_e * \mu$. Supposed μ is known, N_e could then be inferred and used as the timescale in the coalescent model with $N_e(t)$ given relative to this timescale.
Taking the previously mentioned theoretical facts into account we expect to see a behavior of the chain that jumps between different models of constant population size only differing in the actual value of N_e. Therefore, inference results shown here are all scaled to a fixed present day population size of 1.0, which makes it easy to directly compare results from different MCMC runs with the true underlying simulated history. First, a data set of 10 and 20 haplotypes respectively were simulated with ms under a fixed value of N_e. Two independent chains were run for 500,000 steps, each analyzing one of the two constant size data sets. Interestingly, as previously expected, the chains jumped between different states of constant size, with the final results being relatively smooth and stable, indicating that no clear change in population size was detected. Results are not graphically shown, since the chain quickly jumped into a model with only one or two intervals. For this simple demographic model increasing the sample size does not significantly improve the accuracy of the final results. The results only rarely showed two or three intervals models, with the population sizes of the neighboring intervals being very similar, only differing by at most 0.2 units of N_e.

A commonly applied method when working with MCMC in general is to run multiple chains in parallel and to use the average over all chains to obtain a final result. Although MCMC, as compared to ABC, uses a directed approach to find the global maximum or minimum in a fitness landscape, the outcome of a single MCMC run can still be slightly biased due to the random nature of the algorithm. Therefore, averaging over several runs may increase the reliability and quality of the results (see section 6.6). For example, histories in figure 4.3 are shown as an average over 10 independent runs of PSMC and diCal. Therefore, if not explicitly stated, MCMC results represent the average over five independent Markov chains. Summarizing so far, the method is able to infer models of constant population size without introducing too many unnecessary dimensions, as was the case for the results of the first skyline-plot method. For the sake of clarity, a summary of the sets of chosen MCMC parameter values that turned out to be the most suitable for the inference of specific models will be given in section 6.11.

Gradually increasing the complexity of the underlying demographic model suggests to introduce a single size change event at a given time point. Going forward in time, this change can either be a reduction or an increase in N_e, with $N_{e\_ancestral} \lesseqgtr N_{e\_recent}$. Hence, two data sets in total were simulated with ms, namely a population increase and decrease, each for a sample size of five diploid individuals (n=10 haplotypes). The time of the event was set to be 0.05 time units ago, increasing or decreasing the population size by a factor of 2.5 and 4 respectively. Summarizing results from figure 6.10, the method clearly identifies the direction of the events and roughly infers the strength of the actual decrease or increase. Although the strength of 2.5 is almost exactly identified for the expansion model, the ancestral N_e for the decrease model is clearly underestimated. This may potentially be the cause of the nature of the coalescent. Looking backward in time, the relatively small population size until around 0.05 time units ago results in an increased rate of coalescent. As a consequence this forces a large number of lineages to coalesce, reducing the present day diversity of the population sample and potentially increasing the coalescent variance between periods of reduced and increased N_e. However, first results clearly show that the method can detect the direction of the size change event without introducing too much random noise. For addressing the sensitivity toward timings of events, the same basic demographic models were simulated but with different event times.

The height of a coalescent tree is $\sim$ 2 coalescent units ($4N_e$) in total [54]. That is, demographic changes much older that 2 coalescent units may be very difficult, if not impossible, to detect because of too few coalescent trees that reach sufficiently far back into the past. Hence, the chosen demographic events should be introduced in the range of 0-1 coalescent units in order to be sure that the actual event of interest is potentially still detectable. Figure 6.11 shows the results of shifting the demographic event back in time. The strengths of the events are the same as before (expansion 2.5 and decline 4). As can be seen from panel A and B, the inferred main size reduction steps follow the different times of demographic events. However, a similar noise as before can be observed, introducing more intervals than actually needed, dividing the single demographic event into several smaller ones. Panels C and D show the results for different times of expansion events. As already seen from the single expansion case, the inferred demographic functions closely follow the underlying simulated model without introducing too much noise. However, times

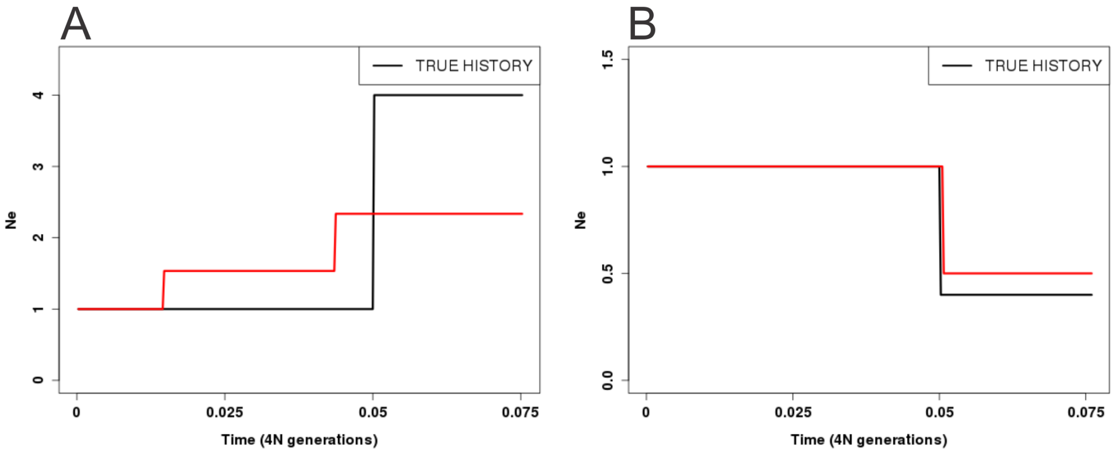


Figure 6.10: Single run rjMCMC results for simulated reduction and increase in size, 0.05 time units ago, for n=10 haplotypes.

are slightly underestimated, resulting in too recent events, but in particular the strengths and times of expansion can be inferred rather accurately.

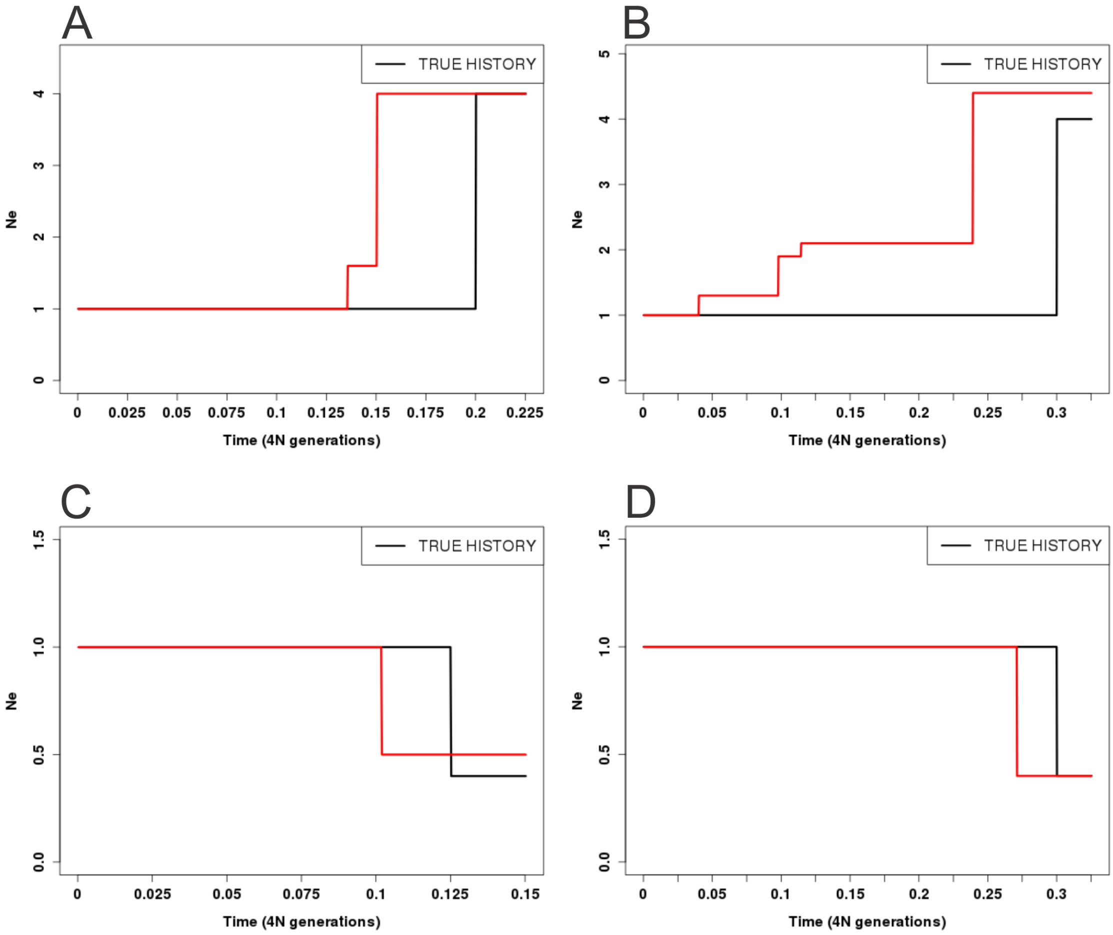


Figure 6.11: Single run rjMCMC results for simulated reduction and increase in size, for different time points, for n=10 haplotypes.

First results demonstrate the power of the 2 point spectrum method to infer past population size changes even if only using five diploid individuals.

As already explained and tested for the AF-IBD/S method, trying to distinguish between bottleneck models is a rather difficult task. The problem that arises is that short but strong bottlenecks may produce a similar SFS as a long but weaker bottleneck. The following tests are supposed to illustrate whether the current method is able to detect such temporal fluctuations in population size. The first bottleneck occurs from 0.075 to 0.175 in units of $4N_e$ generations. The ancestral N_e is reduced by a factor of 7.5 and then population size recovers again. This first model is a rather recent demographic event, so it is interesting to see how accurate the method can be for such time regions, since the accuracy for rather recent times is of special interest for this project. As figure 6.12A shows, the method is able to accurately infer the core part of the bottleneck. What can be seen is that (looking backward in time) the start of the bottleneck is accurately inferred and the strength of size reduction is inferred quite accurately for the core part of the event. Again, the recent population size is fixed to 1 and all subsequent sizes are scaled accordingly (due to the fact that the method is not able to infer absolute N_e values, but only on a relative scale). However, what can also be seen is that the method introduced some additional intervals that would not essentially be needed to explain the underlying model. The introduction of additional intervals is to some extent also due to the fact that this is shown as an average over five independent runs. However, when looking at the results from a single outcome with the best likelihood that was observed for this model it can be seen that the method is indeed able to fit the underlying model quite well (see figure 6.12B). Both results combined show that the way the likelihood function distinguishes between different demographic models, and how this information is used in combination with the rjMCMC approach seems to work properly and gives reasonable results.

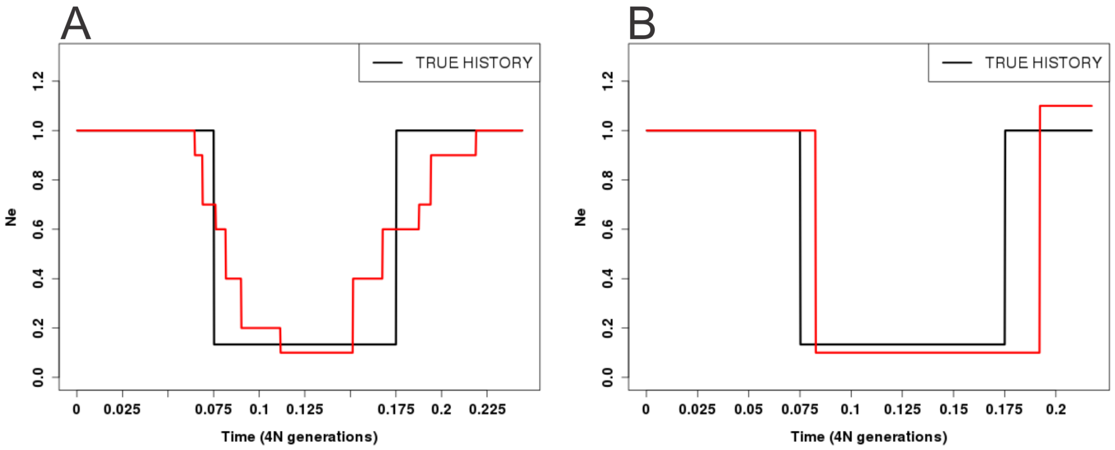


Figure 6.12: A) shows rjMCMC results for a simple three interval bottleneck model, reducing the population size (from 0.075 to 0.175 time units ago) by a factor of 7.5, simulated for n=10 haplotypes. As can be seen, the core part of the event is accurately inferred, although averaging over different runs introduces disturbing intervals. Trajectory is given as an average over five independent MCMC runs. B) shows rjMCMC single result for the same bottleneck model, but this time showing the result from a single run with the best likelihood that was observed for the analysis of this particular model. As can be seen, the inference is very accurate, with time and strength of the bottleneck closely following the underlying size trajectory.

Figure 6.13 shows the results from a bottleneck that was shifted further back in time with the same strength as before. The core part of the bottleneck is again inferred quite accurately. Additionally, the size before and after the event as well as the strength of the bottleneck are closely following the true underlying history. However, shifting such demographic events further back in time allows more mutation events to vanish the actual effects on the SFS and diversity recovers, which makes it more difficult to estimate population size changes, especially with SFS based statistics which are known to be more informative for recent time periods. Hence, the observed variance of the inferred timespan of the event is higher compared to the more recent bottleneck.

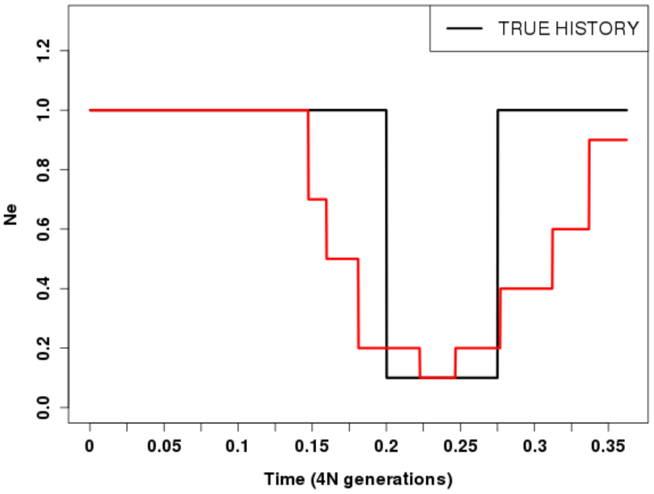


Figure 6.13: RJMCMC results for a simple three interval bottleneck model, reducing the population size (from 0.2 to 0.275 time units ago) by a factor of 7.5, simulated for n=10 haplotypes. As can be seen, the core part of the event is accurately inferred, although averaging over different runs introduces even more disturbing intervals than for the more recent bottleneck in figure 6.12. The inferred trajectory is given as an average over five independent MCMC runs.

So far the analyzed bottleneck models completely recover the N_e after the event. However, additional simulations were performed with the ancestral and recent population sizes being different from each other. These models are probably more realistic, especially in the case of human species that left the African continent and colonized the rest of the world. As can be seen in figure 6.14A, results averaged over five independent runs show a rather accurate trajectory with times and population sizes following the underlying demographic model.

Additionally, the same demographic model with the recent N_e being smaller than the ancient N_e was tested. Figure 6.14B shows the same model as before just with recent and ancient population sizes being swapped. As can be seen, the average over five independent runs gives slightly less accurate results as inference for the model in panel A. A similar trend could be observed when comparing expansion and decline models, with less final accuracy when the population size was decreasing toward more recent time periods.

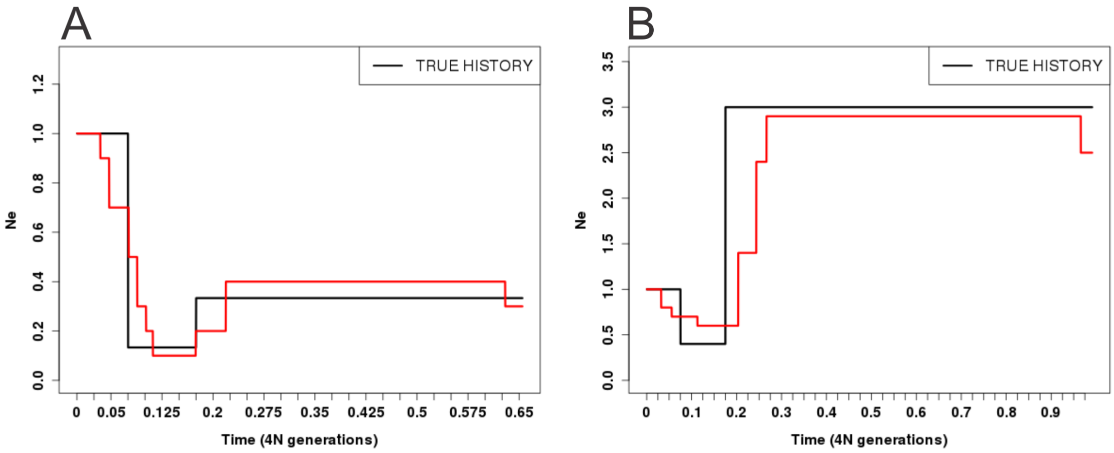


Figure 6.14: A) shows rjMCMC results for a simulated bottleneck, followed by an expansion 0.075 time units ago, simulated for n=10 haplotypes. B) shows rjMCMC results for a simulated bottleneck, followed by a weaker expansion 0.075 time units ago, simulated for n=10 haplotypes. The inferred trajectories are given as an average over five independent MCMC runs.

As a more numerical interpretation of the accuracy of the rjMCMC inference, table 6.1 gives the RMSE values calculated between a single simulated true size trajectory and the inferred final estimate (average over five independent runs). Focusing on the particular demographic models, accuracy clearly reduces as the time of the event is shifted further back in the past. Furthermore, as already mentioned, results from the 2-interval decline models show a rather big RMSE, since time of the event and ancestral N_e were not estimated very accurately. However, bottleneck results are still quite reasonable and even RMSE values are encouraging and reflect a quite precise inference.

In summary the presented outcomes suggest a rather accurate inference of past population size histories. The method is sensitive to timings of size changes as well as the actual relative ratio of population sizes between neighboring intervals.

Model	Number of intervals	Time of event	RMSE
Constant	1	-	0.041
Decline	2	0.05	1.21
Decline	2	0.2	1.36
Decline	2	0.3	1.62
Expansion	2	0.05	0.075
Expansion	2	0.125	0.183
Expansion	2	0.3	0.191
Bottleneck	3	0.0175 - 0.175	0.18
Bottleneck	3	0.2 - 0.275	0.32
Bottleneck-2	3	0.0175 - 0.175	0.14
Bottleneck-3	3	0.0175 - 0.175	0.54

Table 6.1: RMSE calculated between an underlying simulated demographic model and the final rjMCMC estimate for n=10 haplotypes. Inferred results are taken as the average over five independent MCMC runs. Bottleneck models -2 and -3 represent the cases from figures 6.14A and B, respectively.

6.11 Runtime observations and optimal parameter settings

In this section I will briefly explain which sets of parameters turned out to give reasonably accurate results. Accurate in this meaning is not a well defined term, but more like a combined interpretation of the averaged results and the RMSE values of the applied method. Tuning of associated parameters such as prior distributions, proposal variances, etc., is crucial to achieve efficient mixing and accurate results but can be very difficult. Therefore, testing the possible parameter combinations, data sets, and demographic models was a time consuming and computationally very costly process. The first step in the practical inference of parameters of interest is to obtain simulated data. All of the results I have shown are based on only ten simulated haplotypes. The decision not to start with a higher number of individuals and then gradually reducing the amount of data in order to subsequently approach the lower limit, was mostly driven by the fact that the computational complexity of the exact calculation of $\mathrm{E}[T_jT_k]$ is heavily dependent on the number of individuals and the number of intervals m of the demographic model. These facts set limits to the amount of data that were used to compare results from the approximated and exact calculation of second order moments of inter-coalescence times. However, even for this relatively small number of individuals I did observe a rather accurate parameter estimation. As one can see from figure 6.6, the difference between $\mathrm{E}[T_jT_k]$ and $E'[T_jT_k]$ is negligible when approximating over 10^5 inter-coalescence time tuples. This is also reflected in the final results of two independent runs, each with one of the possible methods to obtain the joint moments. The comparison was done by analyzing various demographic models with the same set of chain specific parameters. I could not observe a significant difference between the two methods, which is expected due to the fact that the calculated moments are statistically very similar (see figure 6.6). An example is shown in figure 6.15 where both methods were used to calculate the final likelihood (exact method in red, approximated method in light blue). As can be seen, no significant difference can be observed, which also holds for a variety of other tested demographic models and parameter combinations. However, the main difference between the two possible methods affects the calculation of the likelihoods. When calculating the likelihood for the same demographic model twice, the approximated method will result in two slightly different likelihoods, since the set of simulated inter-coalescence times is regenerated for each new calculation. The difference between the two calculated likelihoods is marginal but the chain applying the exact calculations was able to accept newly proposed steps slightly more often, which is probably caused by the fact that slight changes to the underlying model did generate a smaller likelihood change. Summarizing, the decision to use the exact method coupled with only five diploid individuals turned out to give reliable results for almost all of the tested demographic models. For every demographic model 2,000 independent replicates were simulated, each with the aforementioned number of $\sim$ 50 segregating sites. Hence, all inference results were based on $\sim$100,000 SNPs. Simulating more sites would potentially increase the power and accuracy of the approach, but since we are limited to recombination free regions, it is questionable how much data would practically be available depending on the species of interest.

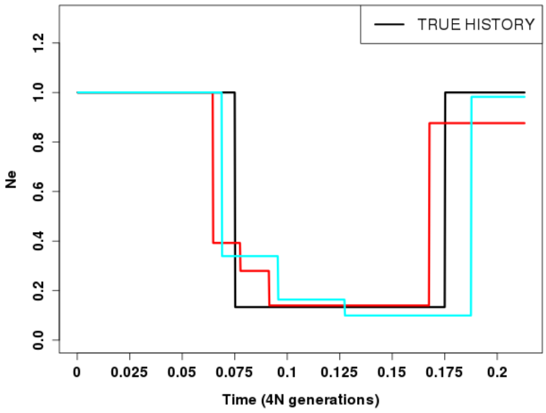


Figure 6.15: Single rjMCMC results for a bottleneck model, likelihood calculation based on the exact ($E[T_jT_k]$) or approximated ($E'[T_jT_k]$) moments calculation, simulated for n=10 haplotypes.

The prior and parameter settings that were adopted from [107] gave solid results throughout all the tested cases. However, the prior for a height change of an interval $N_e' = N_e \cdot exp(z)$ with z being uniformly distributed on [-$\frac{1}{2}$,$\frac{1}{2}$] turned out to be too strong. Choosing and interval of [-0.3,0.3] gave more stable results, without the N_e of neighboring intervals to be too different from each other. All other parameter values and prior distributions were found to work well in this setting. This is expected to some extent, since the setting in [107] is very similar to ours, mainly just differing by a different likelihood function and core statistics.

The Markov chain is started with an initial state which is a randomly generated demographic model. The number m of intervals is chosen as a random number between 1 and m_{max}=8 which was found to be a sufficient upper limit for the chain to efficiently explore the search space without introducing too much noise. Again, increasing the number of intervals affects the computational runtime of the calculation of joint moments of inter-coalescence times. However, for most of the analyses the chain was able to reduce the number of intervals to a number that closely fits the true simulated model, i.e. in the case of a simulated bottleneck model (m=3) final results mostly contained between three and five intervals. The initial demographic history spans a time between 0 and 1 in units of $4N_e$ generations and can be extended or shortened accordingly by applying the extension move types. The mutation rate for all simulations was chosen in a way to produce ~50 SNPs per replicate.

Every Markov chain was run for 600,000 iterations, whereas the initial 5,000 iterations are ignored for a burn-in period. The overall runtime of the approach heavily depends on the number of simulated haplotypes and the method to calculate the joint moments of inter-coalescence times. If using the exact method, runtime also depends on the number of N_e intervals m of the current demographic model. If instead using the approximation method, runtime depends on the number of simulated inter-coalescence time tuples x. Table 6.2 gives a general impression of the runtime of single steps based on different settings. As can be seen, the exact moments calculation is strongly dependent on the number of haplotypes and N_e intervals. At the time of this thesis the exact method did not yet benefit from a parallel implementation with multiple threads. This would be a crucial step towards a more practically feasible runtime for more than five individuals. The current implementation is, therefore, limited in the amount of data that can be analyzed.

Haplotypes(n)	Intervals(m)	Method	Tuples(x)	Threads	Time
10	2	approx.	500	0	0.1s
10	10	approx.	500	0	0.46s
10	2	approx.	500	5	0.05s
10	10	approx.	500	5	0.32s
40	2	approx.	500	0	1.5s
40	2	approx.	500	5	0.65s
40	10	approx.	500	5	0.96s
10	2	exact	-	0	0.06s
10	10	exact	-	0	1.1s
40	2	exact	-	0	25.5s
40	10	exact	-	0	52.5s

Table 6.2: MCMC runtime for different parameter combinations. n=number of haplotypes, m=number of N_e intervals, Method=method to calculate the joint moments E[] or $E'[]$, x=number of simulated inter-coalescence time tuples in thousands, Threads=number of threads running in parallel, Time=runtime for a single MCMC iteration in seconds (depending on the chosen method, a single steps comprises the simulation of x time tuples, the calculation of the joint moments and the calculation of the likelihood for the whole data set, given a newly proposed demographic model). The exact method does not benefit from parallel threads so far. Inference was done on a Intel Core 2 Quad (2.66Ghz), 64bit, 4Gb memory, Ubuntu 12.04, gcc version 4.6.3

6.12 Discussion

In the current chapter I introduced a new non-parametric approach to infer past population size changes as a function $\lambda(t)$ over time. With the advance of sequencing technology and computational facilities it is possible to process consistently growing amounts of genetic data in order to infer parameters of interest. With the cost for sequencing technology steadily decreasing it is possible to even use whole genome sequence data. Therefore, genome-wide non-parametric approaches become widely used and allow a parameter inference that is as accurate as was never achieved before. With the two most recent approaches like the skyline plot method family, *PSMC*, and its derivations of the pairwise sequentially Markovian coalescent framework (e.g. *diCal*), a significant step has been made in terms of obtaining an enormous amount of information from a rather small amount of actual sequence data. Although these methods are widely used, the need to develop further methods will always be indispensable. With the current 2 point spectrum project I used a commonly applied statistic (the site frequency spectrum) that is known to be sensitive to demographic changes. To the best of our knowledge this is the first study to investigate the behavior of a 2 locus site frequency spectrum on demographic inference. One of the potential advantages is the amount of information that can be obtained from a given set of segregating sites. Given a set of s segregating sites, a one locus SFS would summarize the frequency information in a vector of length s. In the case of the 2 locus SFS (the 2 point spectrum method) the data is summarized into a vector of length $\frac{s*(s-1)}{2}$, since the frequency information for every pair of sites is captured. Hence, the 2 point spectrum method is a rational function of not only first, but second or fourth order moments, suggesting that different histories with similar SFS will be easier to distinguish and that the method has improved power over the commonly known SFS. I mentioned fourth order moments, since a different approach to not only summarize the inter-coalescence times by second order moments $E[T_jT_k]$, but by fourth order moments $E[L^2T_jT_k]$, exists. The way this is calculated is as follows:

$$\begin{aligned} E[L^2T_jT_k] &= \sum_{h,l}^{N} hlE[T_hT_lT_jT_k] \\ E[T_hT_lT_jT_k] &= \frac{\sum_{i=1}^{x} v_j[i]v_k[i]v_h[i]v_l[i]}{x}. \end{aligned} \tag{6.17}$$

Using these fourth order moments should even increase the ability to distinguish between demographic models that have a similar one locus SFS or even quite similar $E[T_jT_k]$. However, at the time of this thesis the practical implementation of this approach still suffered from some computational limitations, especially when using more than only five diploid individuals and a high number of coalescent time tuples. Therefore, I was not able to intensively investigate the suggested approach and compare it to the already described algorithms. Hence, this is one of the potential improvements which could significantly increase the accuracy and usability of this method.

The current chapter clearly showed that the newly developed statistic is highly sensitive to past population size changes and can provide a significant amount of information for the inference of demographic parameters within the tested framework of *rjMCMC*. Summarizing, simple constant size, expansion, decline, and bottleneck models can be inferred with rather high accuracy. Even introducing unequal recent and ancient N_e can be inferred and the method still detects the relative shape of the size trajectories. With only 10 simulated haplotypes this method is able to provide accurate and reliable results. The low number of haplotypes is worth mentioning, since other SFS based methods usually require more than only five diploid individuals. Further increasing the sample size may improve the accuracy for later analyses. However, with the current setting 10 simulated haplotypes were the upper limit for the variety of different models and conditions that were tested. The direct comparison between the 2 point spectrum method and PSMC and diCal was done by simulating the exact same demographic history as in Sheehan et al. (history S2, [121]). Results from figure 6.16 are encouraging. The 2 point method is able to detect the time point and strength of the fist size change (at 0.1 in units of 2N generations), and also detects the second change and strength (at 0.4 in units of 2N generations) quite accurately. However, the time point of the third size change (simulated at 1 units of 2N) was not really detected anymore. These more ancient events seem to be undetectable for the current 2 point method. The method is thought to provide insights into the more recent time periods in order to complement already existing and well established methods. This goal is definitely achieved and further research should be done in order to fine tune and validate the results.

Interestingly, Jenkins and Mueller recently showed that the general triallelic frequency spectrum under demographic models with variable population size provides valuable information for distinguishing different demographic growth models from constant models [67]. They show that triallelic sites are more sensitive per site to the parameters of a population that has experienced historical growth, which suggests that they will have use when incorporated into demographic inference. However, their results turn out to be a good motivation for the current work, since triallelic sites are too rare to be of substantial use and are often filtered out due to sequencing errors. This in turn means that using pairwise frequency information, as in the case of the 2 point spectrum method, may provide similar information and acts as a more practical and attractive alternative.

Although I was able to show that with a relatively small number of preset parameters and assumptions the proposed method is able to accurately fit a variety of

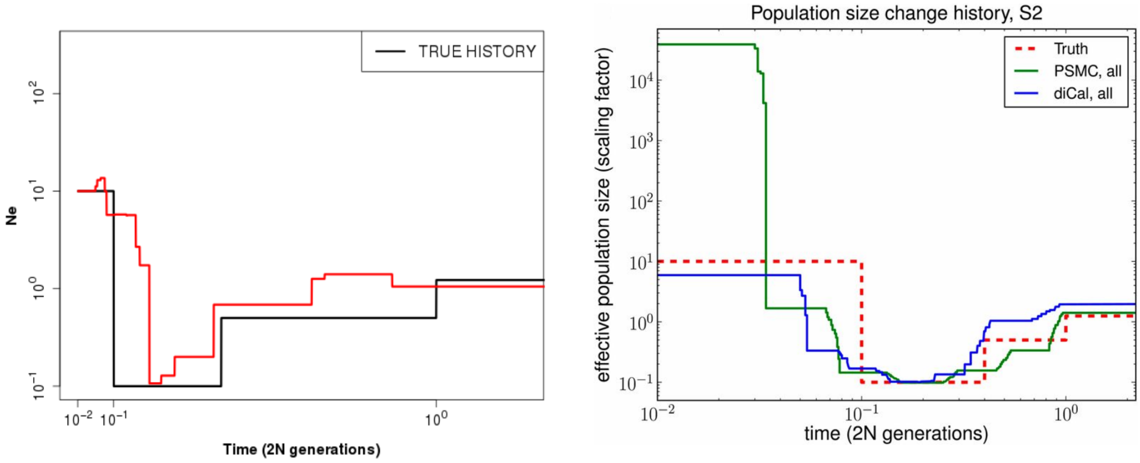


Figure 6.16: RJMCMC results for a demographic model S2, as shown in [121]. Shown is the result from five independent 2 point spectrum runs (left) and the results from PSMC and diCal (right, averaged over 10 runs). In both cases, the model was simulated for n=10 haplotypes. Time is given in 2N generations to better compare the two figures.

demographic models, I am fully aware that, at the time of this thesis, a range of further analyses still need to be performed in order to validate the initial results. When newly proposed and developed methods (algorithms) are introduced, it needs to be clear how they perform under a variety of realistic assumptions and backgrounds (as was investigated for the AF-IBD/IBS method). It still needs to be shown how the method performs when confounding factors (see section 4.1.3) are introduced to underlying simulated pseudo observed data sets. This is by far the most important step in order to more thoroughly understand the internal properties and qualities of this method. Secondly, applying the 2 point spectrum to publicly available data sets is important to show its potential value to empirical parameter inference. Altogether, initial results of the 2 point spectrum already suggest that this method might be of high interest, since even for small amounts of genetic data the accuracy for the tested cases is very encouraging and the potential use for population genetics is given.

So far, genomic data of interest need to provide a recombination map in order to identify regions of very low or no recombination activity at all. When valid regions are identified, they should not be longer than approximately 50 segregating sites in order to decrease the chance of hidden recombination events to affect the underlying genealogy. Therefore, an interesting modification and potential improvement of the method is including recombination rate variation into the method. As I was trying to make clear, the current implementation of the underlying algorithm does not take recombination into account for a specific reason. Recombination would not influence the SFS itself, but can for example have an effect on the estimation of summary statistics like Tajima's D (see [133]). In our case the 2 point spectrum captures the relative fractions of pairwise derived allele frequency information, but ignores information about the absolute number of segregating sites. The latter implies that different models of constant population size have a similar likelihood and further inference about the absolute N_e needs to be performed. On the other hand, the relative pairwise composition of derived allele frequencies is a summary of the data, describing the structure of the underlying coalescent genealogy. Thereby it captures demographic events over different time periods. If sites are obtained from more than one genealogy, the calculated 2 point spectrum is summarized over multiple coalescent trees, each with different branch lengths, subtree structures and TMRCAs. A similar problem occurs for the PSMC method. A hidden Markov model is applied to reconstruct the genome-wide TMRCA distribution across the autosomes, making use of the local densities of heterozygous sites. Segments that are separated by historical recombination events reflect regions of constant TMRCA. This idea could potentially be used to approach the problem of underlying recombination events in a similar way. The practical constraint of identifying regions of low recombination rate before running the inference approach could be avoided by incorporating the PSMC idea into our framework. Therefore, the underlying core algorithm could be used unmodified. A second possibility is to incorporate recombination into equations 6.1 and 6.2 and provide a fine scale map of local recombination rates along the genome. This approach would differ from the current implementation in that it still uses a recombination map but data would not need to be split in recombination free segments prior to the inference. Although the suggested ideas need a thorough revision and might turn out to be impracticable, they offer a high potential to eventually improve the practical use and reliability of the inferred demographic parameters.

Furthermore, a number of pending problems still need to be addressed. Depending on the randomly generated initial state of the chain (i.e. the initial demographic model), the chain sometimes gets stuck in local optima and does not proceed further. This is most likely not a problem of the underlying likelihood, but more a practical problem of the MCMC algorithm to efficiently explore the search space. Hence, further solutions and refinements need to be carried out in order to solve these problems.

Since this method is the first to introduce a two-locus site frequency spectrum and its potential applications to demographic parameter inference, one of the most interesting questions is to show the expected gain in accuracy and power compared to using the more commonly applied one-locus SFS. As shown in figure 6.2, the increase in potentially informative frequency values from the same data set may increase the power for gaining insights into the demographic history of populations or species. Since in the field of theoretical population genetics the SFS is of special interest, the currently proposed statistic and its differences to the previously applied methods should be of high value. The obtained insights and results of this method may assist to give a more thorough and deeper understanding of how demographic changes affect populations and how site frequency spectrum related statistics are influenced in particular.

Chapter 7

Discussion

"An expert is a person who has made all the mistakes that can be made in a very narrow field"

Niels Bohr (1885 - 1962)

In the current thesis I aimed to give a general introduction about the methodological parts of population genetics parameter inference. The coalescent process has been (and still remains to be) one of the most important mathematical models in theoretical population genetics. The past generations of scientists developed the core ideas and subsequently refined the basic principles to an enormously powerful conceptional framework that allows a flexible and easy calculation of a variety of features. With the introduction of the coalescent the simulation of evolutionary processes became possible and set the stage for a better and deeper understanding of how a variety of demographic and selective factors affect the transmission of genetic material from generation to generation. Additional to using the original principles of the Wright Fisher model, extensions like recombination, variable population size, migration, selection, etc., were quickly incorporated and soon enabled the investigation of complex demographic models. Being able to simulate genetic data based on previously defined parameters of interest not only refined the understanding of evolutionary processes, but made it possible to compare artificially generated genetic data to empirical data obtained from a population or species of interest. Over time such approaches have been consequently improved and their applications were enhanced from single statistics to genome-wide methods taking into account a variety of different data. Many of the first methods to study the demography of a population were designed to reject a null hypothesis, i.e. tried to answer whether or not a deviation from neutrality could be observed. Depending on the question of interest, neutrality could assume a constant population size over time, or that no selective events were acting on the individual genomes, etc. Further on, instead of just asking relatively simple questions, methods and algorithms were developed to gain specific details about parameters of interest, e.g. how strong potential selection was acting or when populations most likely split. The principle of comparing simulated to empirical data is the important core part of my work. The AF-IBS/IBD methods extensively compare observed (or pseudo observed) data to simulated data that were generated under the assumptions of certain demographic models. The framework that was used for the implementation is Approximate Bayesian Computation (ABC). I successfully showed that the two summary statistics we introduced capture enough information from the underlying data in order to assist in parameter inference based on past population size changes. The second method, the 2 point spectrum, only focuses on frequency information from segregating sites in order to calculate a likelihood for a data set given a demographic model. Its implementation is based on a different algorithm, namely MCMC. Both methods combined in a single study give an illustrative example of how the evolution of population genetics methods changes the focus from parametric to semi- or even non-parametric methods. In [109] the authors compare ABC and MCMC regarding the accuracy of parameter inference, concluding that both methods can give equally accurate results.

Results of the two methods that were proposed in the current thesis are encouraging and show that the core statistics are sensitive to aspects of past population size changes. Comparing the applicability of both methods is difficult, since they are based on different underlying principles. The AF-IBD/S method represents a parametric approach which has more limiting assumptions on the demographic model and parameter values under study. The 2 point spectrum method tries to put as few limitations and assumptions as possible on the actual model based inference process. However, both methods are of course based on the assumptions of the Wright-Fisher model and principles of the coalescent process. Although chapter 5 provided a comprehensive framework for parameter inference that includes AF-IBD/S as its core statistics, the main conclusion is the usability of AF-IBD/S as summary statistics. They can not only be used in the setting I proposed here but could be incorporated into already existing frameworks and studies to potentially contribute valuable insights into the demographic history of populations of interest. Another significant difference is the amounts of data that are required for each of the methods. AF-IBD/S was tested for ∼40 haplotypes (20 diploid individuals), and using large parts of all autosomal chromosomes. I simulated a variety of model misspecifications and started to account for confounding factors and biases (e.g. recombination rate variation, phasing errors, ascertainment bias). Depending on the computational equipment the runtime of the entire pipeline strongly depends on the number of demographic models that are to be tested and compared. In the case of the 1-, 3- and 5-parameter models, the entire analysis was running for ∼1.5 days, where the simulations for the SD-ratio grid actually consumed most of the time.

The 2 point spectrum method uses the frequency information of pairs of segregating sites and incorporates this information into a likelihood based rjMCMC framework. In terms of user-friendliness this method is probably more convenient to apply, since the goal was to allow for a high flexibility and efficiency. The underlying principles of the rjMCMC approach enable direct conclusions about the demographic model that was inferred. Again, the 2 point spectrum method can in principle not only be used in the framework that was proposed in the current work. The statistic can also be incorporated into already existing frameworks and studies although we think that the method we presented here (likelihood calculation, inference with rjMCMC, etc.) is already quite efficient. In terms of the amounts of data the 2 point spectrum method was tested for much less data (number of individuals and number of segregating sites) than the AF-IBD/S method. However, as I mentioned before, several things still need to be tested and validated for the final evaluation of the 2 point spectrum method. This includes a more thorough investigation of model misspecifications, ascertainment bias, hidden recombination events, phasing errors, etc.

The site frequency spectrum was extensively used in this thesis, at first as a summary statistic combined with the extent of identity by state/decent, and secondly as the core statistic for the calculation of a likelihood for a pairwise frequency configuration. The demographic sensitivity of SFS based statistics has been intensively investigated in numerous studies (e.g. [113, 133]). As already mentioned, population structure can massively skew the SFS and influence the results of parameter inference. The more recent approaches, be it methods based on IBD tracts, skyline plot methods, methods using ROH or LD, or derivatives of PSMC, are definitely affected by multiple confounding factors as well. However, the impact of these factors on the mentioned methods still needs to be studied and investigated which requires a lot of caution when interpreting their results.

Due to the constantly improving technological possibilities, inference based on genetic data alone is becoming more reliable and robust. However, this has not always been true. Until recently, statements about past populations were inferential in nature and depended on information from other fields such as archeology, palaeontology, or linguistics. This is due to the fact that inference is mostly based on underlying models. As I hope is clear after reading the current thesis, such models more or less depend on limitations, assumptions, and rules that need to be made in order to simulate and reproduce evolution. Because of that, different models are artificially forced to appear similar to some extent just because they rely on the same underlying limitations. This in turn equalizes their outputs, resulting in multiple hypotheses that can not clearly be refused or accepted. With the use of larger genomic regions (i.e. even genome-wide data) information from a large number of independent loci, distributed over multiple chromosomes, can be used. Furthermore, the increase in computational speed and power enables the implementation of more model parameters, which might increase the reliability of inferred parameters, weakening the importance of additional evidence from different data analyses.

Chapter 8

Appendix

8.1 AF-IBD AF-IBS appendix

8.1.1 Part I

It is known that the intra-allelic LD pattern contains information about the allele age in absolute time scale (in generations), independent of demographic assumptions . This can be illustrated by examining the decay of the ancestral haplotype following the introduction of a new variant. The ancestral segment around the variant becomes shorter as the ends recombine with other haplotypes in succeeding generations. Assuming the recombination distribution is Poisson along the genetic distance, it is easy to show that the expected length of the remaining ancestral segment measured in Morgans is simply the inverse of the age of the variant τ_1 in number of generations . On the other hand, allele frequency provides age information on the coalescent time scale, as a variant of frequency j can occur only after (back in time) the j lineages coalesced into one. Griffiths and Tavaré derived an approximation of the age distribution given the allele frequency:

$$P(\tau_1 \leq \tau) = E[(1-p)^{n(\tau)-1}] \tag{8.1}$$

where p is the allele frequency, $n(\tau)$ is the number of lineages that are ancestral to the sample at τ. τ_1 is the allele age . The time in coalescent scale can be written as $\int_0^\tau \frac{dx}{2N(x)}$, which is a function of population size N(τ) over the absolute time τ in generations. The formula above suggests that each particular allele frequency j (for a given sample size n) represents a time range on the coalescent scale, with lower (higher) allele frequencies denoting the more recent (ancient) time ranges. When the allele age in absolute time scale, measured by intra-allelic LD, is contrasted against the allele frequency, it actually reflects the coalescence rate of a coalescent range given by equation 8.1. The coalescence rate is again determined by the population size trajectory N(τ). Since each allele frequency represents a different coalescent range, the intra-allelic LD measurements conditioned on allele frequency will reveal population sizes of different parts of the entire coalescent process, and therefore may have the resolution to trace even small changes on the population size trajectory

8.1.2 Part II

Coalescent trees of n nodes can be repeatedly sampled from a given demographic model, and mutations that define subtrees can then be super-imposed onto the root edges leading to j nodes. AF-IBD can then be calculated over a sample of these mutations according to equation 5.2. Since AF-IBD is calculated on genome scale data, we assume the number of mutations approaches infinity. Given this, the above step of creating mutations can also be omitted. Instead, since the frequency of mutations happening on an edge is proportional to the length of the edge, we calculate AF-IBD by weighting on the length of the root edges. The Monte Carlo sampling can be described via the following equation:

$$AF - IBD_{n,j} = \frac{1}{(\mu + \rho)} \sum_{C} I(j, C) \sum_{J} \frac{\omega_J}{T_J} \tag{8.2}$$

Here C indicates instances of coalescent trees, I(j,C) is 1 if there are one or more cases of sub-trees of j lineages given C, and 0 otherwise. J denotes cases of sub-trees of j lineages and T_J the total length of J. The weight term ω_J is proportional to the length of the root edge (figure 5.1A) and ω_J sums to 1. Briefly, the simulation is done by first sampling a large number of coalescent trees C, followed by detecting all sub-trees J of j lineages. $AF - IBD_{n,j}$ is then calculated over all J's as shown in equation 4.

8.1.3 Part III

Noting that AF-IBS and AF-IBD are strongly related, and their ratios are relatively robust against changes in demographic parameters, we constructed a ratio grid on which AF-IBD can be efficiently converted to the corresponding AF-IBS. The simulated AF-IBS can then be compared to the empirical AF-IBS to determine the demographic parameters that give the best fit. The grid is defined as follows: Assume the prior of any parameter is distributed as ~U[a,b]. The grid should then be designed such that its boundaries coincide with a and b (i.e. for any additional gridpoint x, $a < x < b$). The remaining points x can then be uniformly spaced between a and b, whereas a tradeoff between the amount of pre-calculated grid points (i.e. accuracy) and the amount of time needed to generate the grid has to be found. For the 1 parameter constant size model we calculated the ratio for 21 parameter values, namely 1,000, 3,000...41,000 for steps of every 2,000 (i.e. grid points are equally spaced on the prior range, which is ~U[1000, 41000]). For the 3 parameter model, we chose 4 values covering the prior range of each parameter dimension. This gives a total of $4^3 = 64$ different combinations of parameter sets. For the 5 parameter model, we similarly chose 4 values for each. For any arbitrary parameter sets within the parameter space that are not represented on the grid, the elements of the ratio vector were imputed by assuming a local linear dependency of it on the parameter dimensions. For the 1 parameter model, assume the parameter variable (i.e. N_e) is represented by l_1, $l_2 ... l_k$, and the corresponding ratio elements are a_1, $a_2 ... a_k$ respectively. Assume the parameter point x, for which the ratio is to be imputed, occurs between l_j and l_{j+1}, then the ratio is estimated as ratio(x) = $a_j + (l_j + 1 - l_j) * (x - l_j)/(a_{j+1} - a_j)$. See figure 8.4 for a graphical representation. For multi-dimensional parameter space, we use a simple approximate method to impute the ratio elements. Using the 3 dimensional space as an example, we denote the 3 dimensional grid points as $l_{i,j,k} = (l_1, l_2, l_3)$ where I, j and k are the indexes of the fixed parameter values on each parameter dimension, and its corresponding ratio element is $a_{i,j,k}$. We assume that an arbitrary parameter point x=(x_1, x_2, x_3) occurs within the cubic space defined by the two diagonal grid points $l_{i,j,k}$ and $l_{i+1,j+1,k+1}$ where $l_{i+1,j+1,k+1} = (l'_1, l'_2, l'_3)$ and its ratio element is $a_{i+1,j+1,k+1}$. We obtain a ratio estimation on each dimension as $ratio(d)(x) = a_{i,j,k} + (l'_d - l_d) * (x_d - l_d)/(a_{i+1,j+1,k+1} - a_{i,j,k})$ where d is the dimension index 1, 2, or 3. In the end the ratio is estimated as the average of ratio(d)(x) over all the three dimensions. This method similarly extends to the model of 5 parameters.

8.1.4 Part IV

ABC settings that generally apply for all analyses in the current paper are described in this section. We implemented the ABC approach as described previously . The rejection-regression algorithm basically involves fitting a local-linear regression of all simulated parameter values to simulated summary statistics. Furthermore, the observed summary statistics are then substituted into a regression equation. All parameters were transformed with log tan before the actual regression analysis . Distances between observed and simulated summary statistics were calculated as Euclidean distances. Out of $1 * 10^6$ simulated coalescent trees, the parameter combinations with the smallest Euclidean distances were kept with an acceptance rate of 1%. Throughout all estimations, parameter values were drawn from predefined uniform prior distributions. Priors are shown in Tables 5.2 and 5.3.

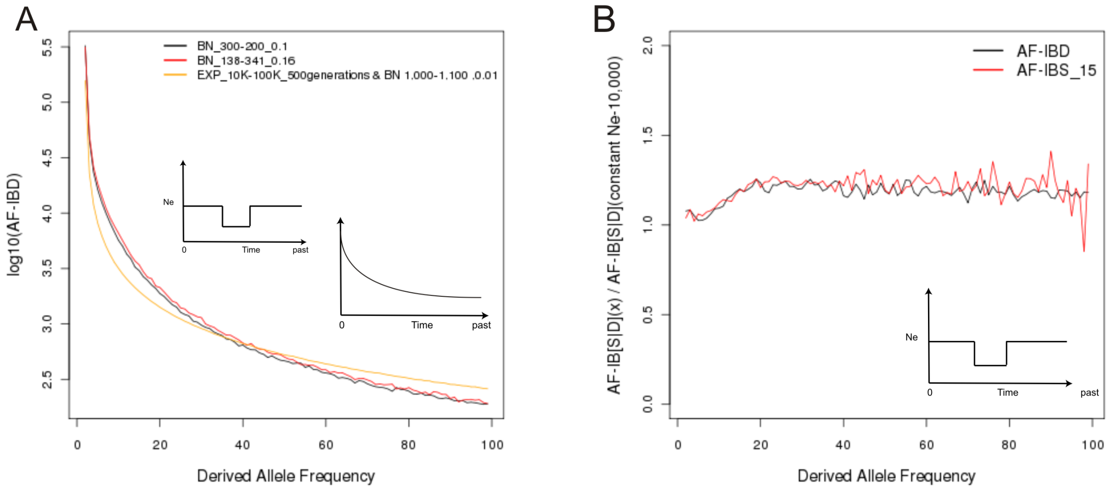


Figure 8.1: AF-IBD curve for bottleneck scenarios and AF-IBD, AF-IBS for additional bottleneck scenarios. A) AF-IBD curves calculated from three different demographic models. (Black) A population that has undergone a single bottleneck occurring from 300 to 200 generations ago, reducing the population size of 10,000 to 1,000. After the bottleneck, the population recovered to the original size of 10,000. (Red) The same setting as in A), but with the bottleneck occurring from 341 to 138 generations ago, reducing the size by a factor of 1.6%. (Orange) The complex model described in figure 5.2E (red). B) Ratios calculated between a model of a single bottleneck occurring from 1,010 to 1,000 generations ago, reducing the population size by N_e*0.01 (recovering the initial size after the bottleneck) and a model of constant population size of N_e=10,000. Ratios are given for AF-IBD and AF-IBS

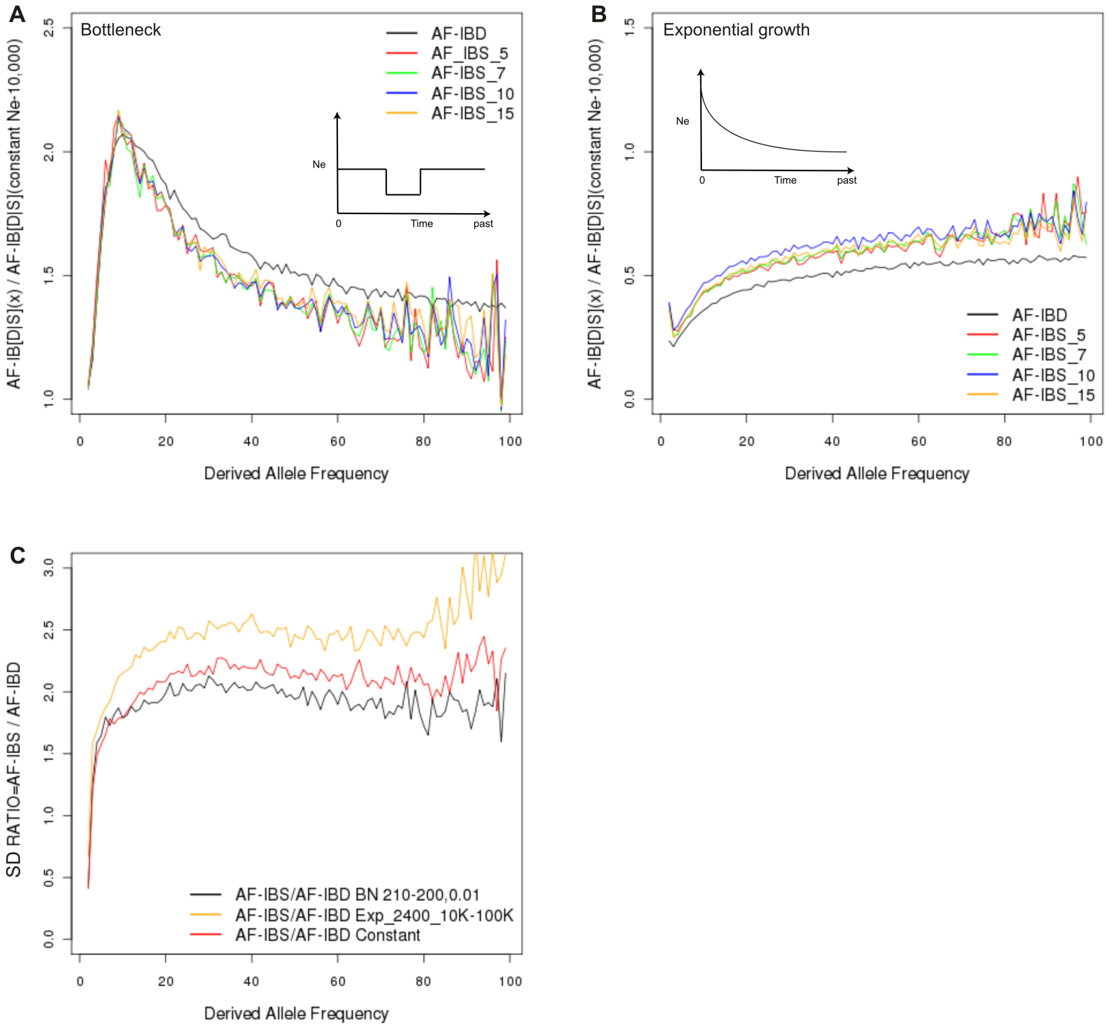


Figure 8.2: Comparison between AF-IBD and AF-IBS for several demographic models. Shown are comparisons between AF-IBD calculated from simulated coalescent genealogy data and AF-IBS calculated from simulated sequence or polymorphism data. For each comparison, data were simulated with the same underlying demographic parameter values (see subfigure cartoons). A) Ratios calculated between a model of a single bottleneck occurring from 210 to 200 generations ago, reducing the population size of 10,000 by a factor of 0.01, and a model of constant population size of N_e=10,000. Ratios are given for the three different statistics (AF-IBS and AF-IBD). For AF-IBS, the statistic was calculated for four different artificially ascertained data sets (see Materials and Methods). AF-IBS_5 represents a stronger ascertainment bias than AF-IBS_15. B) Same ratios as in (A), but calculated between a model of exponential growth (starting 2,400 generations ago, ancient N_e=10,000, present N_e=100,000) and a constant population size of N_e=10,000. C) Ratio coefficients between AF-IBS_5 and AF-IBD, calculated for 3 different demographic models. As can be seen, all 3 statistics show similar results.

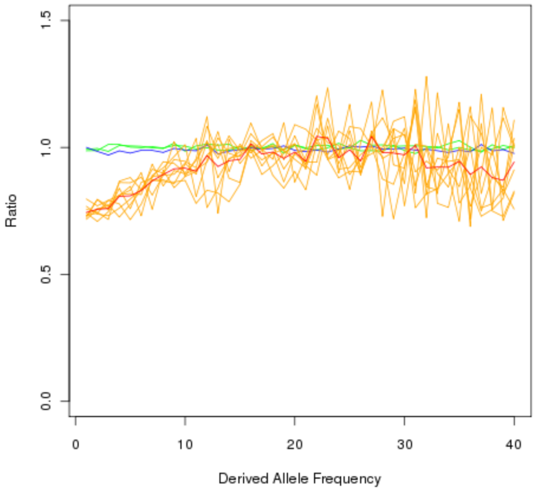


Figure 8.3: Shown are results of the effects of hidden population structure. AF-IBS from an ancestral population was contrasted to AF-IBS of a daughter population (see Methods). Ratio without migration shown in blue, ratios with migration (0.1% & 0.5% per generation) shown in green. Also shown are results from the effects of phase reconstruction errors on AF-IBS. In orange, 10 ratios for AF-IBS before and after phasing are shown (see Methods) for a wide variety of demographic models. Shown in red is the average ratio we used for the correction of our empirical data.

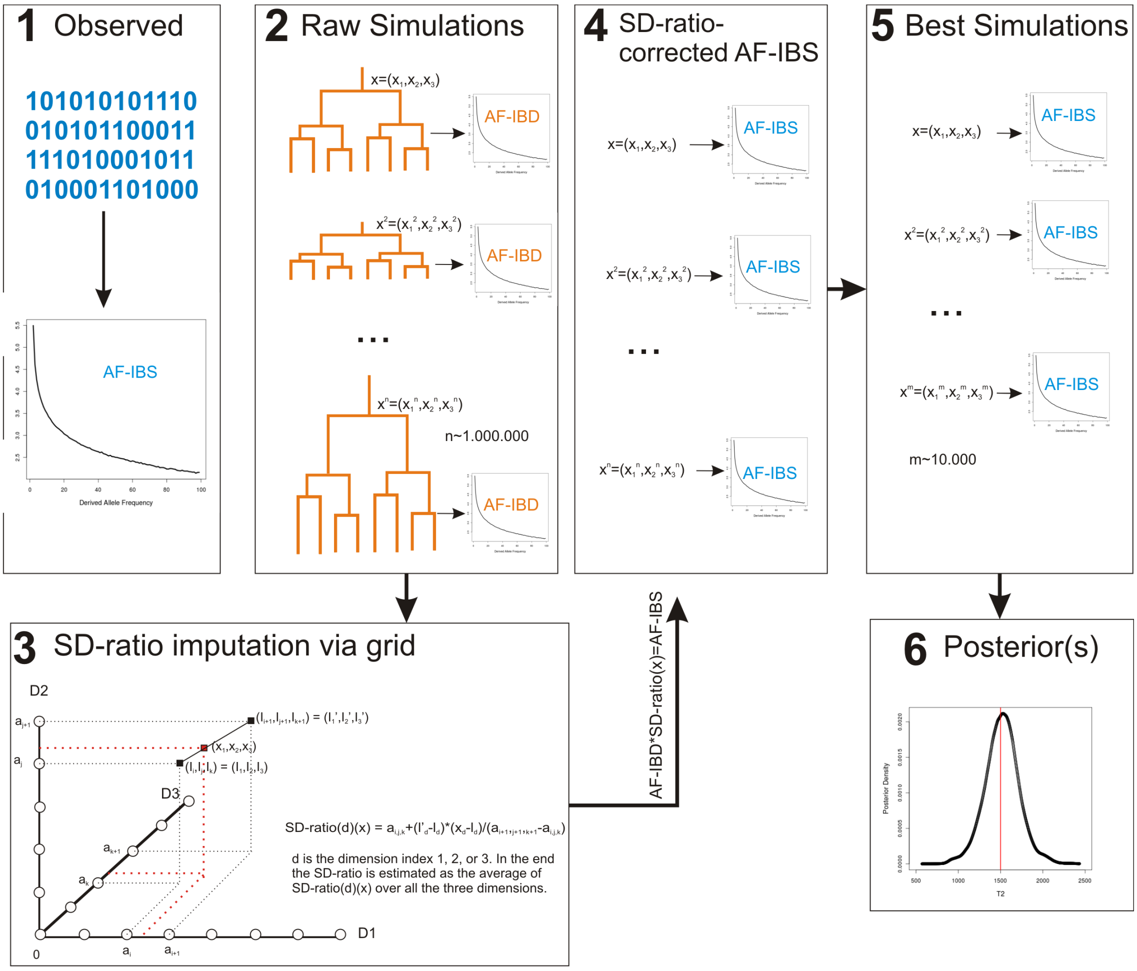

1 Observed
101010101110
010101100011
111010001011
010001101000
AF-IBS
Derived Allele Frequency
2 Raw Simulations
x=(x1,x2,x3)
AF-IBD
x2=(x12,x22,x32)
...
xn=(x1n,x2n,x3n)
n~1.000.000
3 SD-ratio imputation via grid
D2
D3
D1
0
aj+1
aj
ak+1
ak
ai
ai+1
(Ii+1,Ij+1,Ik+1) = (I1',I2',I3')
(x1,x2,x3)
(Ii,Ij,Ik) = (I1,I2,I3)
SD-ratio(d)(x) = ai,j,k+(I'd-Id)*(xd-Id)/(ai+1,j+1,k+1-ai,j,k)
d is the dimension index 1, 2, or 3. In the end the SD-ratio is estimated as the average of SD-ratio(d)(x) over all the three dimensions.
AF-IBD*SD-ratio(x)=AF-IBS
4 SD-ratio-corrected AF-IBS
5 Best Simulations
xm=(x1m,x2m,x3m)
m~10.000
6 Posterior(s)
Posterior Density
T2
500
1000
1500
2000
2500

Figure 8.4: AF-IBS Parameter Inference Scheme.Shown is the general AF-IBS ABC scheme. Step 1) AF-IBS is calculated from any observed phased SNP data. Step 2) According to a specific demographic model and predefined prior distributions, coalescent trees are simulated and AF-IBD is calculated for each data set. Step 3) Based on a predefined grid of SD-ratio coefficients, a new ratio is imputed for each parameter vector x. Step 4) Each AF-IBD from Step 2 is corrected by the imputed ratio, resulting in AF-IBS for each data set. Step 5) After the rejection step, only the simulations that best fit the observed AF-IBS from Step 1 are kept and used to generate posterior distributions for each parameter of interest.

8.2 Exact calculation of E[$T_j T_k$]

$$E(T_k) = \sum_{j=k}^{n} (-1)^{j+k} \alpha_{n,j,k'} \int_0^\infty t g_j(t) dt, 2 \leq k \leq n, \tag{8.3}$$

where

$$\begin{aligned} \alpha_{n,j,k} &= \frac{(2j-1)n!(n-1)!(k+j-2)!}{(j-k)!k!(k-1)!(n-j)!(n+j-1)!}, \\ g_j(t) &= \frac{\binom{j}{2}}{\lambda(t)} exp(-\binom{j}{2} \int_0^t \frac{1}{\lambda(u)} du), \end{aligned} \tag{8.4}$$

with g_2(t) being the density of $(T_2)_2$ with the second index 2 indicating the sample size n=2.

$$g_{j,i}(t,t') = \frac{\binom{j}{2}\binom{i}{2}}{\lambda(t)\lambda(t+t')} exp\left\{-\binom{j}{2}\int_0^t \frac{1}{\lambda(u)} du\right\} exp\left\{-\binom{i}{2}\int_t^{t+t'} \frac{1}{\lambda(u)} du\right\} \tag{8.5}$$

Nomenclature

N_e	Effective Population Size
A priori	In the first place
ABC	Approximate Bayesian Computation
AF-IBD	Allele Frequency - Identity by Descent
AF-IBS	Allele Frequency - Identity by State
bot	Bottleneck
bp	Base Pair
CEPH-HGDP	Centre d'Etude du Polymorphisme Humain - Human Genome Diversity Panel
DAF	Derived Allele Frequency
diCal	Demographic Inference using Composite Approximate Likelihood, Software
DNA	Deoxyribonucleic Acid
EHH	Expected Haplotype Homozygosity
gen	Generation
IBD	Identity by Descent
IBS	Identity by State
kbp	10^3 Base pairs
Kya	Thousand (kilo) Years Ago

LD	Linkage Disequilibrium
MAE	Mean Absolute Error
Mb	10^6 Base pairs
MCMC	Markov Chain Monte Carlo
MH-MCMC	Metropolis Hastings - Markov Chain Monte Carlo
MRCA	Most Recent Common Ancestor
ms	Hudson's Coalescent Simulation Software *make samples*
mtDNA	Mitochondrial DNA
PSMC	Pairwise Sequential Markovian Coalescent, Software
rjMCMC	Reversible Jump Markov Chain Monte Carlo
RMSE	Root Mean Square Error
SFS	Site Frequency Spectrum
SNP	Single Nucleotide Polymorphism
TMRCA	Time to Most Recent Common Ancestor

List of Figures

2.1 The basic principle behind the coalescent 10

3.1 Basic ms output 23
3.2 Wright-Fisher basic sampling 25
3.3 Coalescent time scales 27
3.4 Example genealogy 29
3.5 Comparison between a constant size and bottleneck model 32
3.6 Comparison between an exponential growth and constant size model . 33
3.7 Wright-Fisher with recombination 34
3.8 Simple ancestral recombination graph 35

4.1 Skyline plot methods 42
4.2 Illustration of the PSMC method 43
4.3 Comparison between PSMC and diCal 44

5.1 Illustration of the model and the statistics AF-IBD and AF-IBS . . . 50
5.2 AF-IBD calculated for several simulated demographic models 68
5.3 ABC estimation results for pseudo-observed AF-IBS 69
5.4 AF-IBS calculated for several CEPH-HGDP populations 70

6.1 Demographic model from Myers et al. 72
6.2 Illustration of the 2 point spectrum method 75
6.3 Four topological cases under the infinite sites model 76
6.4 Illustration of inter-coalescence times 79
6.5 Possible coalescent trees for a simple bottleneck model 81
6.6 Difference between exact and approximated joint moments 82
6.7 Results from the sub-SFS method 93
6.8 SFS ratio between Myers et al. demographies 94
6.9 Sub-SFS result for Myers et al. demographies 94
6.10 RJMCMC results for reduction and expansion 97
6.11 RJMCMC results for reduction and expansion at different time points 98

6.12 RJMCMC results for a simple bottleneck model 99
6.13 RJMCMC results for a more ancient bottleneck model 100
6.14 RJMCMC results for reduction and expansion at different time points 101
6.15 Inference differences between exact and approximated joint moments 103
6.16 Comparison between PSMS, diCal and the 2 point method 108

8.1 AF-IBD/S curves for bottleneck scenarios 118
8.2 Comparison between AF-IBD and AF-IBS 119
8.3 Model misspecification 120
8.4 AF-IBS parameter inference scheme 121

List of Tables

3.1 Mutation principle of the Coalescent 28

5.1 AF-IBD/S main principles 51
5.2 Measures of accuracy for AF-IBD-ABC parameter estimation 53
5.3 Measures of accuracy for AF-IBS-ABC parameter estimation 54
5.4 Power of AF-IBS-ABC to recover the true model 60
5.5 ABC estimation results for empirical data for Yoruba and French . . 62
5.6 African demographic parameters compared among different studies . 65
5.7 European demographic parameters compared among different studies 66

6.1 2 point spectrum accuracy 102
6.2 MCMC runtimes 105

Bibliography

[1] A. M. Adams and R. R. Hudson. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics*, 168(3):1699–1712, Nov 2004.

[2] S.H. Ambrose. Late pleistocene human population bottlenecks, volcanic winter, and differentiation of modern humans. *Journal of Human Evolution*, 34:623–651, 1998.

[3] Q.D. Atkinson, R.D. Gray, and A.J. Drummond. mtdna variation predicts population size in humans and reveals a major southern asian chapter in human prehistory. *Molecular Biology and Evolution*, 25(2):468–474, 2008.

[4] A. Auton and G. McVean. Recombination rate estimation in the presence of hotspots. *Genome Research*, 17:1219–1227, 2007.

[5] D.J. Balding, M. Bishop, and C. Cannings. *Handbook of Statistical Genetics*, volume 2. Wiley, 2003.

[6] M.A. Beaumont. Detecting population expansion and decline using microsatellites. *Genetics*, 153(4):2013–2029, 1999.

[7] M.A. Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160, 2003.

[8] M.A. Beaumont, Z. Wenyang, and D.J. Balding. Approximate bayesian computation in population genetics. *Genetics*, 162:2025–2035, 2002.

[9] G. Bertorelle, A. Benazzo, and S. Mona. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Molecular Ecology*, 19(13):2609–2625, Jul 2010.

[10] S. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455, 1998.

[11] B.L. Browning and S.R. Browning. A fast, powerful method for detecting identity by descent. *American journal of human genetics*, 88:173–182, 2011.

[12] S. R. Browning and B. L. Browning. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.*, 12(10):703–714, Oct 2011.

[13] U.S. Census Bureau. Website. Available online at `http://www.census.gov/main/www/popclock.html`; Last visited on December 05th 2013.

[14] A. Carvajal-Rodríguez. Simulation of genomes: A review. *Current Genomics*, 9:155–159, 2008.

[15] Y. L. Chan, C. N. Anderson, and E. A. Hadly. Bayesian estimation of the timing and severity of a population bottleneck from ancient dna. *PLoS Genet.*, 2:0451–0460, 2006.

[16] N.H. Chapman and E.A. Thompson. A model for the length of tracts of identity by descent in finite random mating populations. *Theoretical Population Biology*, 64(2):141 – 150, 2003.

[17] G.K. Chen, P. Marjoram, and J.D. Wall. Fast and flexible simulation of dna sequence data. *Genome Research*, 19:136–142, 2009.

[18] L. Chikhi, V.C. Sousa, P. Luisi, B. Goossens, and M.A. Beaumont. The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics*, 186(3):983–995, 2010.

[19] A. G. Clark, M. J. Hubisz, C. D. Bustamante, S. H. Williamson, and R. Nielsen. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research*, 15(11):1496–1502, Nov 2005.

[20] P. Congdon. *Bayesian Statistical Modelling.* Wiley, 2001.

[21] D. F. Conrad, M. Jakobsson, G. Coop, X. Wen, J. D. Wall, N. A. Rosenberg, and J. K. Pritchard. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat. Genet.*, 38(11):1251–1260, Nov 2006.

[22] Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[23] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, Oct 2005.

[24] J. M. Cornuet and G. Luikart. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics*, 144:2001–2014, 1996.

[25] C. Darwin. *On the Origin of Species*, volume 1. John Murray, London, 1959.

[26] F. Depaulis, S. Mousset, and M. Veuille. Power of neutrality tests to detect bottlenecks and hitchhiking. *J Mol Evol*, 57:190–200, 2003.

[27] F. Depaulis and M. Veuille. Power of neutrality tests to detect bottlenecks and hitchhiking. *Molecular Biology and Evolution*, 15:1788–1790, 1998.

[28] A. J. Drummond, A. Rambaut, B. Shapiro, and O. G. Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22(5):1185–1192, 2005.

[29] L. Excoffier. Human demographic history: refining the recent african origin model. *Current Opinion in Genetics & Development*, 12:675–682, 2002.

[30] L. Excoffier, A. Estoup, and J. M. Cornuet. Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics*, 169:1727–1738, 2005.

[31] L. Excoffier and G. Heckel. Computer programs for population genetics data analysis: a survival guide. *Nat. Rev. Genet.*, 7(10):745–758, Oct 2006.

[32] N. J. Fagundes, N. Ray, M. Beaumont, S. Neuenschwander, F. M. Salzano, S. L. Bonatto, and L. Excoffier. Statistical evaluation of alternative models of human evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 104(45):17614–17619, Nov 2007.

[33] J. C. Fay and C. I. Wu. Hitchhiking under positive Darwinian selection. *Genetics*, 155(3):1405–1413, Jul 2000.

[34] P. Fearnhead. Consistency of estimators of the population-scaled recombination rate. *Theoretical Population Biology*, 64(1):67 – 79, 2003.

[35] P. Fralick, D.W. Davis, and S.A. Kissin. The age of the gunflint formation, ontario, canada: single zircon u pb age determinations from reworked volcanic ash. *Canadian Journal of Earth Sciences*, 39(7):1085 – 1091, 2002.

[36] O. François, M. G. B. Blum, M. Jakobsson, and N. A. Rosenberg. Demographic history of european populations of arabidopsis thaliana. *PLoS Genet.*, 4:e1000075, 05 2008.

[37] Y. X. Fu. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, 147(2):915–925, Oct 1997.

[38] Y.X. Fu and W.H. Li. Statistical tests of neutrality of mutations. *Genetics*, 133:693–709, 1993.

[39] N. Galtier, F. Depaulis, and N. H. Barton. Detecting bottlenecks and selective sweeps from dna sequence polymorphism. *Philosophical Transactions of the Royal Society of London B*, 155:981–987, 2000.

[40] L. M. Gattepaille, M. Jakobsson, and M. G. Blum. Inferring population size changes with sequence and snp data: lessons from human bottlenecks. *Heredity*, pages 409–419, 5 2013.

[41] Human genome project. Website. Available online at `http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml`; Last visited on November 25th 2008.

[42] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in practice.* Chapmann & Hall, 1997.

[43] Chikhi L. Goldstein, D.B. Human migrations and population structure: What we know and why it matters. *Annu. Rev. Genomics Hum. Genet.*, 3:129–152, 2002.

[44] S. Gravel, B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, A. G. Clark, F. Yu, R. A. Gibbs, The 1000 Genomes Project, and C. D. Bustamante. Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U.S.A.*, 108(29):11983–11988, Jul 2011.

[45] P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[46] R.E. Green, J. Krause, and A.W. et al. Briggs. A draft sequence of the neandertal genome. *Science*, 328:710–722, 2010.

[47] R.C. Griffiths and S. Tavaré. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London B*, 344:403–410, 1994.

[48] The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409:928–933, 2001.

[49] A. K. Gupta. Origin of agriculture and domestication of plants and animals linked to early holocene climate amelioration. *Current Science*, 87:54–59, 2004.

[50] A. Gupta Hinch, A. Tandon, and N. et al. Patterson. The landscape of recombination in african americans. *Nature*, 476:170–175, 2011.

[51] R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.*, 5(10):e1000695, Oct 2009.

[52] H.C. Harpending, A.R. Sherry, S.T. Rogers, and M. Stoneking. The great human expansion. *Current Anthropology*, 34(4):483–496, 1993.

[53] K. Harris and R. Nielsen. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet.*, 9(6):e1003521, 06 2013.

[54] J. Hein, M. H. Schierup, and C. Wiuf. *Gene Genealogies, Variation and Evolution - A Primer in Coalescent Theory.* Oxford University Press, 2005.

[55] J. Heled and A. Drummond.

[56] B. M. Henn, L. L. Cavalli-Sforza, and M. W. Feldman. The great human expansion. *Proc. Natl. Acad. Sci. U.S.A.*, 109(44):17758–17764, Oct 2012.

[57] W. G. Hill. Estimation of effective population size from data on linkage disequilibrium. *Genetics Research*, 38:209–216, 1981.

[58] S. Y. W. Ho and B. Shapiro. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Molecular Ecology Resources*, 11(3):423–434, 2011.

[59] S. Hoban, G. Bertorelle, and O. E. Gaggiotti. Computer simulations: tools for population and evolutionary genetics. *Nat. Rev. Genet.*, 13(2):110–122, Feb 2011.

[60] A. Hobolth and C. Wiuf. The genealogy, site frequency spectrum and ages of two nested mutant alleles. *Theoretical Population Biology*, 75(4):260 – 265, 2009.

[61] R.R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23:183–201, 1983.

[62] R.R. Hudson. Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics*, 18:337–338, 2002.

[63] R.R. Hudson, K. Bailey, D. Skarecky, J. Kwiatowski, and F.J. Ayala. Evidence for positive selection in the superoxide dismutase (sod) region of drosophila melanogaster. *Genetics*, 136:1329–1340, 1994.

[64] H. Innan and M. Nordborg. The extent of linkage disequilibrium and haplotype sharing around a polymorphic site. *Genetics*, 165(1):437–444, 2003.

[65] H. Jeffreys. Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31:203–222, 4 1935.

[66] P. Jenkins. The joint frequency spectrum of two completely linked segregating sites with variable population size. *Unpublished work*, August 2013.

[67] P. Jenkins, J. Mueller, and Y. S. Song. General triallelic frequency spectrum under demographic models with variable population size. *Genetics*, 196(1):295–311, 2014.

[68] P. Jenkins and Y. S. Song. The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele. *Theor Popul Biol*, 80(2):158–173, Sep 2011.

[69] M.A. Jobling, M.E. Hurles, and C. Tyler-Smith. *Human evolutionary genetics - Origins, peoples & disease.* Garland Science, 2004.

[70] T.H. Jorgensen, B. Degn, and A.G. et al. Wang. Linkage disequilibrium and demographic history of the isolated population of the faroe islands. *Eur J Hum Genet*, 10(6):381–387, 2002.

[71] T.H. Jukes and C.R. Cantor. Evolution of protein molecules. *Mammalian protein metabolism. Academic Press*, pages 21–132, 1969.

[72] A. Keinan, J. C. Mullikin, N. Patterson, and D. Reich. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat. Genet.*, 39(10):1251–1255, Oct 2007.

[73] J. K. Kelly. A test of neutrality based on interlocus associations. *Genetics*, 146:1197–1206, 1997.

[74] M. Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61:893–903, 1969.

[75] M. Kimura and J. Crow. The number of alleles that can be maintained in a finite population. *Genetics*, 49:725–738, 1964.

[76] M. Kimura and G. H. Weiss. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49:561–576, 1964.

[77] J.F.C. Kingman. The coalescent. *Stochastic Processes and Their Applications*, 13:235–248, 1982.

[78] J.F.C. Kingman. Exchangeability and the evolution of large populations. *Exchangeability in Probability and Statistics*, pages 97–112, 1982.

[79] J.F.C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19A:27–43, 1982.

[80] J.F.C. Kingman. Origins of the coalescent: 1974-1982. *Genetics*, 156:1461–1463, 2000.

[81] M. Kirin, R. McQuillan, C. S. Franklin, H. Campbell, P. M. McKeigue, and J. F. Wilson. Genomic runs of homozygosity record population history and consanguinity. *PLoS ONE*, 5(11):e13996, 11 2010.

[82] A. Kong, D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, S. T. Palsson, M. L. Frigge, T. E. Thorgeirsson, J. R. Gulcher, and K. Stefansson. A high-resolution recombination map of the human genome. *Nat. Genet.*, 31(3):241–247, Jul 2002.

[83] M. K. Kuhner, P. Beerli, J. Yamato, and J. Felsenstein. Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics*, 156:439–447, 2000.

[84] M. K. Kuhner, J. Yamato, and J. Felsenstein. Estimating effective population size and mutation rate from sequence data using metropolis-hastings sampling. *Genetics*, 140(4):1421–30, 1995.

[85] M. K. Kuhner, J. Yamato, and J. Felsenstein. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, 149:429–434, 1998.

[86] J. Lachance and S.A. Tishkoff. Snp ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *BioEssays*, 35(9):780–786, 2013.

[87] H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475:493–496, 7 2011.

[88] H. Li and W. Stephan. Inferring the demographic history and rate of adaptive substitution in drosophila. *PLoS Genet.*, 2:e166, 10 2006.

[89] J. Z. Li, D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, and R. M. Myers. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866):1100–1104, Feb 2008.

[90] B. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80:220–239, 1988.

[91] K. E. Lohmueller, C. D. Bustamante, and A. G. Clark. Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. *Genetics*, 182(1):217–231, May 2009.

[92] K. E. Lohmueller, A. R. Indap, S. Schmidt, A. R. Boyko, R. D. Hernandez, M. J. Hubisz, J. J. Sninsky, T. J. White, S. R. Sunyaev, R. Nielsen, A. G. Clark, and C. D. Bustamante. Proportionally more deleterious genetic variation in european than in african populations. *Nature*, 451:994–997, 2008.

[93] D. Lopez Herraez, M. Bauchet, K. Tang, C. Theunert, I. Pugach, J. Li, M. R. Nandineni, A. Gross, M. Scholz, and M. Stoneking. Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. *PLoS ONE*, 4(11):e7888, 2009.

[94] G. Luikart, F.W. Allendorf, J. M. Cornuet, and W. B. Sherwin. Distortion of allele frequency distributions provides a test for recent population bottlenecks. *The Journal of heredity*, 89:238–247, 1998.

[95] P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain monte carlo without likelihoods. *Proc. Natl. Acad. Sci. U.S.A.*, 100:15324–15328, 2003.

[96] G. T. Marth, E. Czabarka, J. Murvai, and S.T. Sherry. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, 166(1):351–372, 2004.

[97] M. S. McPeek and A. Strahs. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *American Journal Of Human Genetics*, 65:858–875, 1999.

[98] G.A.T McVean and N. J. Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1387–1393, 2005.

[99] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[100] V. N. Minin, E. W. Bloomquist, and M. A. Suchard. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution*, 25(7):1459–1471, 2008.

[101] S. Myers, C. Fefferman, and N. Patterson. Can one learn history from the allelic spectrum? *Theoretical Population Biology*, 73:342–348, 2008.

[102] M Nei. *Molecular Evolutionary Genetics.* Columbia University Press, New York, 1987.

[103] M. Nei, T. Maruyama, and R. Chakraborty. The bottleneck effect and genetic variability in populations. *Evolution*, 29:1–10, 1975.

[104] R. Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154:931–942, 2000.

[105] R. Nielsen and J. Signorovitch. Correcting for ascertainment biases when analyzing snp data: applications to the estimation of linkage disequilibrium. *Theoretical Population Biology*, 63:245–255, 2003.

[106] M. Nordborg and S. Tavaré. Linkage disequilibrium: what history has to tell us. *Trends in genetics*, 18:83–90, 2002.

[107] R. Opgen-Rhein, L. Fahrmeir, and K. Strimmer. Inference of demographic history from genealogical trees using reversible jump markov chain monte carlo. *BMC Evolutionary Biology*, 5, 2005.

[108] B. Peng, C.I. Amos, and M. Kimmel. Forward-time simulations of human populations with complex diseases. *PLoS Genet.*, 3(3):0407–0420, 2007.

[109] V. Plagnol and S. Tavaré. Approximate bayesian computation and mcmc. *Monte Carlo and Quasi-Monte Carlo Methods*, 2002.

[110] A. Polanski, A. Bobrowski, and M. Kimmel. A note on distributions of times to coalescence, under time-dependent population size. *Theoretical Population Biology*, 63(1):33 – 40, 2003.

[111] S. Ptak and M. Przeworski. Evidence for population growth in humans is confounded by fine-scale population structure. *Trends Genetics*, 18:550–563, 2002.

[112] O. G. Pybus, A. Rambaut, and P. H. Harvey. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, 155(3):1429–1437, 2000.

[113] A. Ramírez-Soriano, S. E. Ramos-Onsins, J. Rozas, F. Calafell, and A. Navarro. Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics*, 179:555–567, 2008.

[114] S. E. Ramos-Onsins and J. Rozas. Statistical properties of new neutrality tests against population growth. *Molecular Biology and Evolution*, 19:2092–2100, 2002.

[115] A. R. Rogers, A. E. Fraley, M. J. Bamshad, W. S. Watkins, and L. B. Jorde. Mitochondrial mismatch analysis is insensitive to the mutational process. *Molecular Biology and Evolution*, 13(7):895–902, 1996.

[116] N. A. Rosenberg and M. Nordborg. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature*, 3:380–390, 2002.

[117] P.C. Sabeti, D.E. Reich, J.M. Higgins, and H.Z.P. et al. Levine. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419:832–837, 2002.

[118] A. Sano and H. Tachida. Gene genealogies and properties of test statistics of neutrality under population growth. *Genetics*, 169:1687–1697, 2005.

[119] S. F. Schaffner, C. Foo, S. Gabriel, D. Reich, M. J. Daly, and D. Altshuler. Calibrating a coalescent simulation of human genome sequence variation. *Genome Research*, 15:1576–1583, 2005.

[120] P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, 78(4):629–644, 2006.

[121] S. Sheehan, K. Harris, and Y. S. Song. Estimating variable effective population sizes from multiple genomes: A sequentially markov conditional sampling distribution approach. *Genetics*, 194(3):647–662, 2013.

[122] M. Slatkin. Simulating genealogies of selected alleles in a population of variable size. *Genetic Research*, 78:49–57, 2001.

[123] M. Slatkin. A bayesian method for jointly estimating allele age and selection intensity. *Genetic Research*, 90(1):129–137, 2008.

[124] M. Slatkin and G. Bertorelle. The use of intraallelic variability for testing neutrality and estimating population growth rate. *Genetics*, 158(2):865–874, 2001.

[125] K. Strimmer and O. G. Pybus. Exploring the demographic history of dna sequences using the generalized skyline plot. *Molecular Biology and Evolution*, 18(12):2298–2305, 2001.

[126] C. Strohbeck. Average number of nucleotide differences in a sample from a single subpopulation: A test for population subdivision. *Genetics*, 117:149–154, 1987.

[127] J.A. Sved. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology*, 2(2):125 – 141, 1971.

[128] A.C. Syvänen. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Review Genetics*, 2(12):930–942, 2001.

[129] F. Tajima. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123:585–595, 1989.

[130] S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, Feb 1997.

[131] C. Theunert. Using the coalescent to infer human demographic history from genomic data, 2009.

[132] C. Theunert, K. Tang, M. Lachmann, S. Hu, and M. Stoneking. Inferring the history of population size change from genome-wide snp data. *Molecular Biology and Evolution*, 29(12):3653–3667, 2012.

[133] K. Thornton. Recombination and the properties of Tajima's D in the context of approximate-likelihood calculation. *Genetics*, 171(4):2143–2148, 2005.

[134] K. Thornton and P. Andolfatto. Approximate bayesian inference reveals evidence for a recent, severe bottleneck in a netherlands population of drosophila melanogaster. *Genetics*, 172:1607–1619, 2006.

[135] B. F. Voight, A. Adams, L. A. Frisse, Y. Qian, R. R. Hudson, and A. Di Rienzo. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc. Natl. Acad. Sci. U.S.A.*, 102:18508–18513, 2005.

[136] D. Živković and T. Wiehe. Second-order moments of segregating sites under variable population size. *Genetics*, 180(1):341–357, 2008.

[137] J. Wakeley, R. Nielsen, S. N. Liu-Cordero, and K. Ardlie. The discovery of single nucleotide polymorphisms-and inference about human demographic history. *American Journal of Human Genetics*, 69:1332–1347, 2001.

[138] J. D. Wall. Recombination and the power of statistical tests of neutrality. *Genetical Research*, 74:65–79, 1999.

[139] J. D. Wall, K. E. Lohmueller, and V. Plagnol. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Molecular Biology and Evolution*, 26(8):1823–1827, Aug 2009.

[140] D. Wegmann, C. Leuenberger, and L. Excoffier. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, 182(4):1207–1218, Aug 2009.

[141] S. Wright. Evolution in mendelian populations. *Genetics*, 16:97–126, 1931.

[142] A. Yu, C. Zhao, Y. Fan, W. Jang, A. J. Mungall, P. Deloukas, A. Olsen, N. A. Doggett, N. Ghebranious, K. W. Broman, and J. L. Weber. Comparison of human genetic and sequence-based physical maps. *Nature*, 409(6822):951–953, Feb 2001.

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialen oder erbrachten Dienstleistungen als solche gekennzeichnet.

Ort und Datum

Unterschrift