Simultaneous estimation of population size changes and splits times using importance sampling

Marie Forest

Department of Statistics

University of Oxford
Christ Church College

A thesis submitted for the degree of

Doctor of Philosophy

Hillary 2014

Supervisors : Jonathan Marchini and Simon Myers

Simultaneous estimation of population size changes and splits times using importance sampling

Marie Forest, Christ Church College

Department of Statistics, University of Oxford

D.Phil. Thesis, Hillary 2014

Abstract

The genome is a treasure trove of information about the history of an individual, his population, and his species. For as long as genomic data have been available, methods have been developed to retrieve this information and learn about population history. Over the last decade, large international genomic projects (e.g. the HapMap Project and the 1000 Genomes Project) have offered access to high quality data collected from thousands of individuals from a vast number of populations. Freely available to all, these databases offer the possibility to develop new methods to uncover the history of the peopling of the world by modern humans. Due to the complexity of the problem and the large amount of available data, all developed methods either simplify the model with strong assumptions or use an approximation; they also dramatically down-sample their data by either using fewer individuals or only portions of the genome.

In this thesis, we present a novel method to jointly estimate the time of divergence of a pair of populations and their variable sizes, a previously unsolved problem. The method uses multiple regions of the genome with low recombination rate. For each region, we use an importance sampler to build a large

number of possible genealogies, and from those we estimate the likelihood function of parameters of interest. By modelling the population sizes as piecewise constant within fixed time intervals, we aim to capture population size variation through time. We show via simulation studies that the method performs well in many situations, even when the model assumptions are not totally met. We apply the method to five populations from the 1000 Genomes Project, obtaining estimates of split times between European groups and among Europe, Africa and Asia. We also infer shared and non-shared bottlenecks in out-of-Africa groups, expansions following population separations, and the sizes of ancestral populations further back in time.

In memory of my mother and my brother.

And to my father.

Acknowledgements

I would like first to thank my supervisors, Jonathan Marchini and Simon Myers, it has been a privilege to work with both of you.

I would like to thank my examiners Gil McVean and Richard Durbin for their helpful comments and suggestions.

I would like to acknowledge the financial support that I received from the Clarendon scholarship, the Natural Sciences and Engineering Research Council of Canada, Christ Church College and the Department of Statistics. Without their support, I would not have been able to pursue my doctorate.

I would like to thank: Claire C., Andy D., Elena F., Cathy F., Nancy F., Valentina I., Fabrice L., Melissa M., and Androniki M., for their help and support at different stages of my doctorate.

Thank you Didier, for your love and support, and for following me across the pond during this crazy adventure.

Thank you Papa, for your unconditional love and support. Thanks to you and Maman, I have always believed that everything was possible and achievable with enough effort.

Contents

Chapter 1

Introduction

For numerous years, researchers have tried to find the actual tree of life; how all species are related to each other. Near the bottom of this tree, we might be interested in how different human groups diverged from each other and at what time. It is important to develop methods specifically to answer those questions. Our aim is to study population structure using samples of genomic data from different populations. More precisely, we are interested in the time of divergence of closely related populations and how their sizes varied through time. The focus of this thesis is a new method of inference based on adapting the Stephens and Donnelly (64) importance sampler to model populations that split and change sizes through time.

In this chapter, we first introduce some concepts of genetics needed to understand this thesis. Second, we present the coalescent process used to model the genealogy of a sample of individuals. We explain how it can be used to simulate data and to make inference. Then we give an overview of some of the different methods that have been developed in the past to answer related questions. When we contemplate genomic data, we are observing the result of a genealogical process. Therefore, most methods estimate genealogical properties of the data and relate those to the parameters of interest. Some methods will directly build genealogies for the sample and others will simulate data using different parameters

to compare them with the real sample. The methods presented use the coalescent process to model the genealogy.

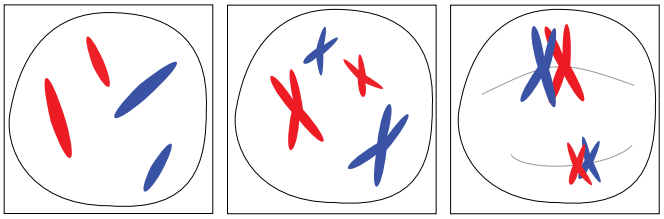
1.1 Genetic terminology

We will use human reproduction as an example to briefly explain some basic concepts and terms of genetics. This will be useful to understand this thesis. Every human being originated from only an unique cell that contained all the information needed to create him. This information is referred to as the human genome. The human genome is composed of 46 chromosomes; a chromosome is a long condensed coiled DNA molecule. Without going into details, we note that DNA is a double stranded helix linked by pairs of nucleobases, denoted G, C, T and A. The base G is always paired with the base C, and the base T with the base A. The sequence of these pairs control the sequence and expression of proteins, which determine the development and function of the body.

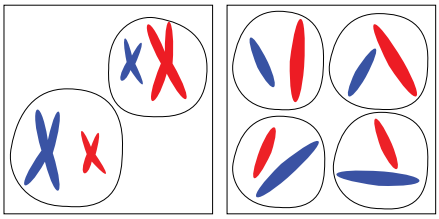
1.1.1 Human reproduction

In all sexual species, chromosomes come in pairs, each pair being formed by chromosomes from each parent. The chromosomes of a pair are called homologous. They are the same length and perform the same functions. Note that a cell composed of pairs of homologous chromosomes is called a diploid cell. In human beings, pairs of homologous chromosomes are numbered from 1 to 22, and the last pair of chromosomes is constituted of sex chromosomes X and Y.

A human sex cell (spermatozoid or ovule) is composed of only 23 chromosomes, one chromosome for each pair. We will now see in more detail the phenomenon of cell division called meiosis that creates the sex cells. This phenomenon is illustrated in Figure 1.1 where, for simplicity, we follow two pairs of chromosomes of different lengths. The red colour represents the chromosomes inherited from the mother, the blue, the ones inherited from



(a) Before the meiosis (b) Duplication (c) Pairing



(d) First division (e) Second division

Figure 1.1: Meiosis steps: we follow a cell containing two pairs of chromosomes, same length chromosomes are homologous.

the father. Chromosomes evolve in an independent manner in the interior of the cell (Figure 1.1(a)). Before the cell division, the chromosomes are replicated. This phenomenon, called replication, begins near the centre of a chromosome (more precisely at the centromere) and moves to the two ends in an independent manner. The chromosomes and their replicas remain joined, for now, at their centromere (Figure 1.1(b)). Then there is a pairing of homologous chromosomes (Figure 1.1(c)). This pairing is important because it is during this time that the phenomenon of crossover (or recombination) occurs, which allows greater genetic variety. Following the crossovers, the next phase is the first cell division, where homologous chromosomes are separated into two cells (Figure 1.1(d)). Finally, a second cell division takes place that separates the chromosome copies. After a meiosis, four daughter cells remain, each containing 23 chromosomes. These cells are what we call gametes and are haploid cells (Figure 1.1(e)).

During the pairing of homologous chromosomes, parts of the chromosomes are exchanged between members of the pair. The two resulting homologous chromosomes form

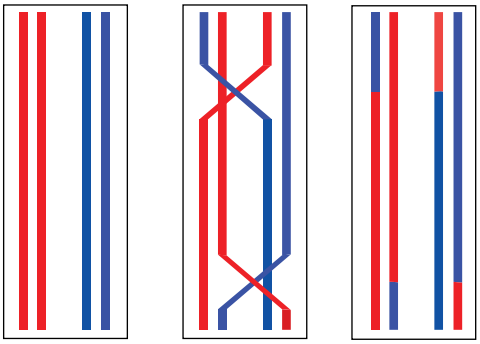


Figure 1.2: Stages of a recombination event. From left to right, the homologous chromosomes after replication (not connected by their centromeres), then two recombination events take place and finally the resulting chromosomes.

a mixture of paternal and maternal chromosomes. Figure 1.2 illustrates this phenomenon. The place where a crossover occurs is random. Throughout this thesis, we will use the term recombination when we make reference to crossovers. In summary, two phenomena of the meiosis create genetic variation: one is recombinations, the other is the random segregation of paternal and maternal chromosomes during the first division.

1.1.2 Genes, alleles and mutations

We have seen that all the information needed to make an individual is contained in his genome. We will now see how this information is expressed. While visually we can see a lot of differences between two human beings, any two individuals actually share about 99.9% of their genomes (7). After reflection, this number is not so far-fetched. One can easily imagine the large amount of information necessary to make a heart beat, to create blood vessels, make a brain function. Those activities essentially do not vary from one person to another. But differences do exist between individuals of the same species and appear on the genome.

Some of the differences between two individuals can be caused by environmental factors, but many features that define us depend on the information contained in our genome. In

genetics, these biological features are called phenotypic character (or trait, *e.g.* natural hair colour, blood group). Originally, the definition of a gene was the information on a chromosome influencing a trait. The reality is more complex: some characters are actually influenced by several genes. A gene is composed of a set of consecutive base pairs which together transmit information to perform a precise task. The exact location of a set of base pairs on a chromosome is called a locus (plural loci). The locus of a gene refers to the precise location and physical base pairs forming it. For the same gene the information supplied may differ: we do not have all the same hair colour, nor the same blood group. We call allele the different possibilities of information transmitted by a gene; the alleles are different versions of a same gene. Recall that chromosomes come in pairs, and so do the genes, and therefore two alleles determine a character (we speak of genes in the singular when in reality we reference a pair of genes).

During the meiosis and the duplication of chromosomes, errors can be introduced: a pair of nucleobases can be forgotten, wrongly reproduced or simply reproduced twice. We call these errors mutations. The mutations may be inconsequential, but they can also create a new allele for a gene. One consequence of a mutation can be, for example, to prevent the production of a certain protein. If one mutation is present, the "healthy" allele may produce enough protein to allow the body to function as usual. But if two mutant alleles are present, the protein cannot be produced and some cells cannot perform their tasks. The individual would then be suffering from a genetic disease.

To position the genes on chromosomes and on a genetic map, we need a measuring unit of position. In genetics, there are two units of measurement commonly used. The first is simple and uses pairs of nucleobases. This unit of measurement is physical, it counts the number of base pairs located between two genes and is denoted by *bp*. The second measuring unit is based on the work of Thomas Hunt Morgan and his team at the beginning of the twentieth century and is based on recombination. A recombination event between two genes is observed if they are not from the same parental chromosome,

that is to say, if there was an odd number of crossovers between these genes in the pairing of homologous chromosomes. If one recombination between two genes in 100 meiosis is observed, we say that there is a recombination rate of 1% between these genes. This is equivalent to a distance of 1 centiMorgan, denoted by cM, between those genes.

Usually genome data will be composed of genetic markers. These are genes or pieces of DNA sequences with known positions that show variation (different alleles) among and between populations. There are different types of genetic markers. One type is made of short DNA sequences where only a single base pair shows variation in the population. That is to say that for this short sequence of base pairs, all individuals in the population have exactly the same sequence with the exception of a precise base pair. This base pair can be use as a genetic marker, these are called single nucleotide polymorphisms (SNP). In general, there are two possible alleles per human SNP since the probability of mutation at a position of the genome is extremely low. The samples used in genetic data are composed of sequences of several SNP ordered on a chromosome. Figure 1.3 shows how, from a sequence of DNA, it is possible to extract the markers and to determine the distance between them.

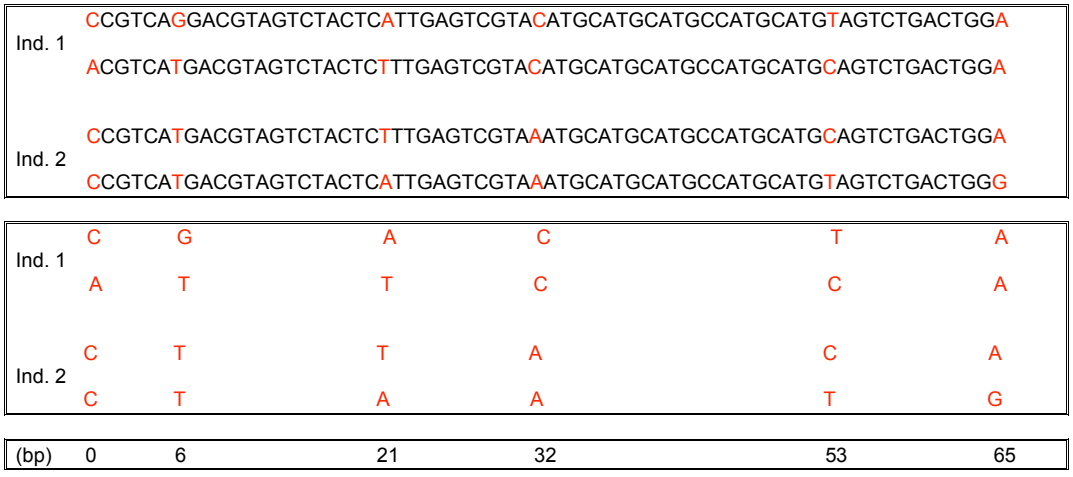


Figure 1.3: Sequences of markers: short DNA sequences formed of 65 bases pairs. The variations between those sequences are used as genetic markers. The distance between those markers is measure in bases pairs.

1.2 Coalescent process

The coalescent process is widely used in population genetics to approximate the genealogy of a sample (33)(53)(23)(69). It is often referred to as the Kingman-coalescent in recognition of Kingman's work (36), but it was also independently developed by Hudson (32) and Tajima (66). The principal strength of the coalescent is that it is the limit process of a large variety of realistic genetic models of evolution. In this section, we first present the standard coalescent. Then, we summarise how we can adapt the coalescent to different genetic models, from recombination to population structure. Afterwards, we see how to use the coalescent to make inferences about population parameters and to simulate datasets under various models.

1.2.1 The standard coalescent

Following the usual approach, we will introduce the coalescent process as an approximation of the Wright-Fisher model of evolution, considering first discrete time and then continuous time. Finally, we will append mutations following the infinite alleles model and the infinite sites model.

1.2.1.1 Wright-Fisher model and the coalescent

First consider a clonal population. The Wright-Fisher model of evolution makes simple assumptions:

- The population size N, is constant over time,
- Generations are discrete and non-overlapping,
- Each individual has only one parent in the previous generation, so that the population consists of haploids.

Since the population size is constant, the total number of offspring in a generation must be equal to the population size, denoted by N. It can be seen as each offspring choosing randomly a parent from the previous generation. Therefore, the joint distribution of the number of offspring of all the individuals from a generation is a multinomial distribution $(N, 1/N, ..., 1/N)$.

Suppose now that we are interested in simulating the genealogical process under the Wright-Fisher model. We could either go forwards in time and select offspring from the multinomial distribution, or we could go backwards in time, from the present to the past, by sampling with replacement the parent of each individual. Figure 1.4(a) presents this genealogical process when $N = 10$ for ten generations. The ten dots at the bottom of the figure represent the individuals of the present generation, and as we go up in the figure, we go further back in time. Figure 1.4(b) highlights in blue the genealogy of the present generation, and Figure 1.4(c) keeps only the individuals that are part of that genealogy. We say that two lineages coalesce when they find a common ancestor. When all the individuals of a generation have found a common ancestor it means that we have found the most recent common ancestor (MRCA) of the population; in Figure 1.4(c) the only individual remaining in the most ancient generation is the MRCA.

Figure 1.4(c) contains more information than is needed since all the ancestors of each generation are kept. The generations of interest are only those in which we find an ancestor of more than one lineage; the others are redundant. Usually only a subset, of size $n << N$, of the population is sampled. Therefore, we are interested in modelling the genealogy of only a small sample of the population. Figure 1.5 represents the genealogy of a sample of size three from the population shown in Figure 1.4. In Figure 1.5(b), we have kept only the events that are related to the present-day sample.

We will now introduce the coalescent process used to approximate the simplified genealogy of a sample (as in Figure 1.5(b)). We are interested in describing the distribution of the waiting times before a coalescence event. First, we will concentrate on a specific pair

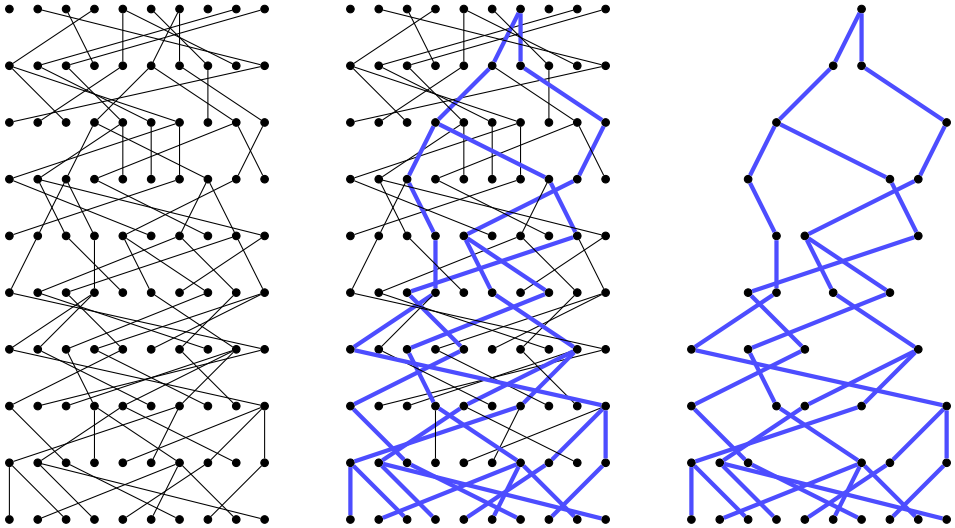


Figure 1.4: Example of a genealogy under the Wright-Fisher model

of individuals in the present generation and describe the waiting time before they find a common ancestor. At each generation, they will choose the same ancestor with probability $1/N$. The time —in number of generations— before they find a common ancestor is then geometrically distributed with parameter $1/N$. Let T_i be the time of the first coalescence event when there are i lineages in our sample. The probability that a fixed pair of lineages finds a common ancestor after j generations is then:

$$P(T_2 = j) = \left[1 - \frac{1}{N}\right]^{j-1} \frac{1}{N}.$$

When we are interested in k lineages, the probability that they remain distinct in the previous generation is:

$$\prod_{i=1}^{k-1} \frac{N-i}{N} = \prod_{i=1}^{k-1} \left(1 - \frac{i}{N}\right) = \left[1 - \binom{k}{2} \frac{1}{N}\right] + O\left(\frac{1}{N^2}\right). \tag{1.1}$$

This is because the first lineage can choose its parent from the N ancestors in the previous generation, the second lineage can only choose from the $N-1$ remaining ancestors, and so

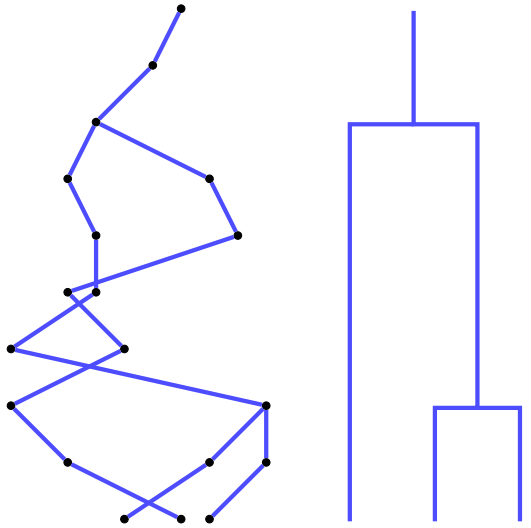


Figure 1.5: Genealogy of a sample of three individuals from the population represented by the most recent generation in Figure 1.4.

on. Therefore, ignoring second-order effects that tend toward 0 for large N, the probability that k lineages remain distinct for j generations is:

$$\left[1 - \binom{k}{2}\frac{1}{N}\right]^j.$$

The second-order effects in fact correspond to the probability that more than two lineages find a common ancestor in the same generation.

To obtain the description of the continuous coalescent process, we need to rescale time such that one unit of time is equivalent to N generations. The probability that the k lineages remain distinct for at least time t ($t = j/N$), where the time is in the new continuous scale, is:

$$P(T_k > t) = \left[1 - \binom{k}{2}\frac{1}{N} + O\left(\frac{1}{N^2}\right)\right]^{\lfloor Nt \rfloor} \to \exp\left\{-\binom{k}{2}t\right\},$$

as $N \to \infty$. The time before the first coalescence event when there are k lineages remaining follows an exponential distribution of rate $\binom{k}{2}$. Notice that the time depends on the number of lineages remaining, but it does not depend on the population size. Since each pair is equally likely to coalesce, the pair will be chosen uniformly at random. The property of discrete and non-overlapping generations of the Wright-Fisher model implies that the times between each coalescence event are independent.

1.2.1.2 The coalescent with mutation

Including mutation in the Wright-Fisher model is simple. Every time an individual has a child, a mutation will occur with probability μ. Going backwards in time, the probability that a lineage has experienced a mutation in the generation before is the same: μ. Again, in discrete time, the probability that the first mutation happens at the j^{th} generation back in time is $(1-\mu)^{j-1}\mu$.

Once more, we will rescale time such that one unit is equal to N generations. Since μ is the probability of observing a mutation in one generation, we need to rescale the mutation rate. Denote $\theta = 2N\mu$ to be the scaled mutation rate. Looking at one lineage only, the probability that the first mutation on this lineage happens at the j^{th} generation is:

$$P(T_M = j) = \left[1 - \frac{\theta}{2}\frac{1}{N}\right]^{j-1} \frac{\theta}{2}\frac{1}{N},$$

where T_M is the time before the mutation event. In continuous and rescaled time, the probability that the time before the first mutation event on a lineage is greater than t, where t is in the new rescaled time unit, is:

$$P(T_M > t) = \left[1 - \frac{\theta}{2}\frac{1}{N}\right]^{\lfloor Nt \rfloor} \to \exp\left\{-\frac{\theta}{2}t\right\},$$

as $N \to \infty$. Thus, mutations occur back in time as a Poisson process of rate $\theta/2$ per unit of scaled time on each edge of the coalescent tree.

In the coalescent, the time before a mutation event on a lineage is exponentially distributed and the mutation events on different lineages are independent of each other. The time before the first mutation event when we are looking at k lineages will then be exponentially distributed with the rate equal to the sum of all the independent rates, *i.e.* $\theta k/2$. The mutation events are independent of the coalescence events. The time before any event—coalescence or mutation—when there are k lineages, is exponentially distributed with rate

$$\binom{k}{2} + \frac{\theta k}{2} = \frac{k(k-1+\theta)}{2}.$$

The mutation events can also be viewed as the result of a Poisson process on the branch length of the topology of the genealogy. Therefore, the genealogical process can be viewed as separated from the mutation process.

Different models for the mutation process exist; we will present two. In the infinite alleles model, each new mutation event creates a new allele that has never been seen before. The sample is composed of a particular gene for each individual. Micro-satellite data can be modelled with the infinite alleles assumption. There is also the infinite sites model, for which the sample is composed of sequences and each new mutation happens at a new position on the sequence. Only one mutation event could have happened in the past at each position. This mutation model makes sense when SNP data are used. The Wright-Fisher model with mutation make also the assumption that every mutation are neutral, and therefore does not influence the probability of mating. This neutrality assumption means that there is no selection occurring in the sample.

1.2.2 Adaptions of the coalescent

In the previous section, we presented the basic coalescent process as an approximation of the Wright-Fisher model. Here, we will provide an overview of how it can be adapted to a variety of different genealogical models.

First, it can be shown that the coalescent process is an approximation of many other models and, in particular, the Moran model. The Moran model is characterised by overlapping generations, with each new generation created by choosing randomly one individual that will give birth and one that will die. This construction does not allow coalescence of more than two lineages. According to Hein *et al.* (23) this model is used mostly by theoreticians because of the simplicity of the associated calculations, but it seems to have less appeal to biologists.

Until now, the samples presented were composed of haploids, but the coalescent can also be adapted to diploids. In fact, usually a population of N diploids can simply be seen as a population of $2N$ haploids. When N is large, and the time is rescaled accordingly (one unit = $2N$), this approximation is correct.

When the populations sampled evolve significantly differently than the Wright-Fisher model, we need to rescale time according to the "effective" population size (N_e) to obtain a well defined coalescent process. Different definitions exist; in the coalescent framework, N_e is the population size of a Wright-Fisher population with the same properties as the population of interest (69). Nordborg (53) gives an example of this. Suppose that the individuals of a population of constant size still have on average one child (constant size) but the variance of the number of children is σ^2 (instead of 1 as in the Wright-Fisher model). Then a population evolving according to the Wright-Fisher model but of size N/σ^2 will have the same limiting properties. Therefore, to model the genealogy of this population we can use the coalescent process with the time rescaled such that one unit is equivalent to $N_e = N/\sigma^2$ generations.

We have assumed previously that the population size is constant. It is possible to extend the coalescent process to variable population size, but for this we must suppose that we know how time varies. If we can model the population size over time as a continuous function, it can be included in the coalescent process. The main idea is to change the scaling of the time as we go backwards in time as this will affect the coalescence rate;

when the population size is smaller, more coalescence events will happen, and vice versa. Nordborg (53) and Hein *et al.* (23) give more details about variable population size.

When using larger sequences from diploids we need to account for the possibility of recombination events. Hudson (32) was the first to present an adaptation of the coalescent to recombination, and Griffiths and Tavaré (19) introduced the ancestral recombination graph (ARG). When recombinations are taken into account, the genealogy is a graph instead of a tree, since a recombination event seen backwards in time introduces a new lineage instead of eliminating one as in a coalescence event. The genealogy can also be viewed as a collection of trees, one for each position/SNP along the sequences. The challenge with recombination is the infinite number of possible genealogies for a sample of sequences.

Finally, we will present models that include multiple populations. First, there is the island model, for which we assume that a long, long time ago, a sub-population split from an original population and, since then, those populations have exchanged migrants. The assumptions are that the split occurred so long ago that there is no need to model that split and that the two populations have since attained equilibrium, which is often unrealistic. Figure 1.6(a) illustrates this model. In this model, a coalescence event can only happen when two lineages belong to the same population. Usually, the population sizes are assumed constant but different. A choice needs to be made for the rescaled time, and the coalescence rates need to be adapted. The migration rates by generation can be equal or allowed to differ. Different island models have been presented with more than two populations, sometimes geographically organised (49).

The island model allows us to model the genealogy of samples from different populations, but doesn't allow inference about the time of the split. For this, we need to include the split in the model. We examine the simplest case of population structure: two isolated populations that originate from the same ancestral population. This model is often referred as the isolation model, with the assumption that no migration occurred between the populations of constant size. Figure 1.6(b) illustrates this model.

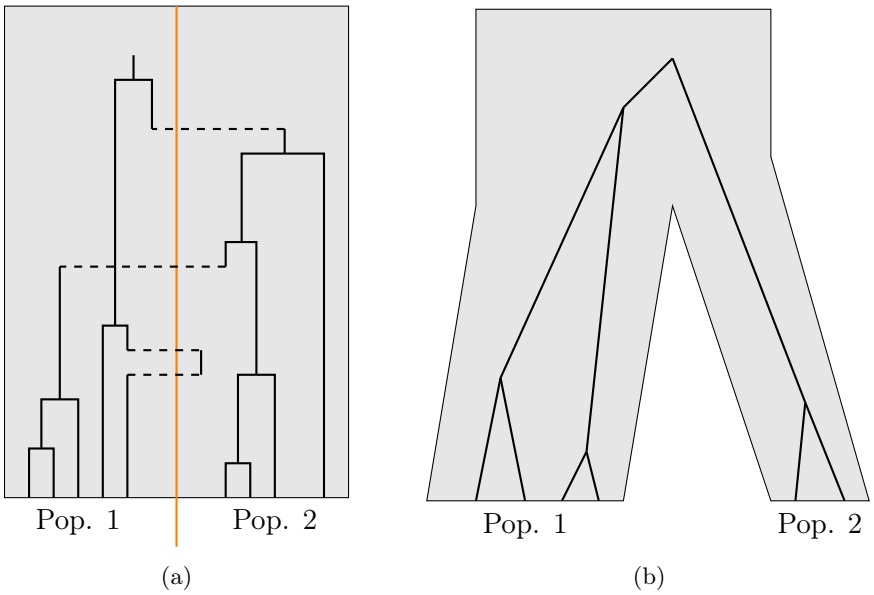


Figure 1.6: (a) example of the island model and (b) of the isolation model . The orange line in Figure (a) illustrates the separation between the two populations. Migration events are represented by the dotted lines. The horizontal length of the grey area represent the population sizes. In Figure (b) the split is represented by the separation of the grey region into two grey regions.

As with the coalescent, the genealogy is shaped by coalescence and mutation events. Going backwards in time, lineages of different populations cannot coalesce before their populations join to form the ancestral population. Before the time of divergence, denoted here by T, the events happening in one population are independent of the events happening in the other, and can be modelled independently with the coalescent process. After T, all lineages from the populations are united in one ancestral population and are then evolving according to the standard coalescent.

Formally, the time is measured in a way such that one unit is equivalent to N generations; the mutations follow the infinite sites model with a rate $\theta = 2Nu$, where u is the mutation rate per generation. Let N_i denote the effective and constant size of population i. Before time T, the time until an event in population i (t_{ik}) is exponentially distributed with the rate $\lambda_k = \binom{k}{2} N/N_i + k\theta/2$ when there are k lineages in population i. With prob-

ability $(\binom{k}{2}N/N_i)/\lambda_k$, the event is a coalescence, or else the event is a mutation. After T, the lineages remaining in the different populations are assembled together, and the time before the next event is exponentially distributed with the same rate λ_k where N is the effective size of the ancestral population, and k the number of lineages remaining. The ratio N/N_i comes from the fact that time is measured in units of N generations but the probability of a coalescence of two lineages in the next generation is $1/N_i$ in population i (before rescaling the time).

We have presented some of the many possible adaptations of the coalescent process. There are yet more, for example it can be adapted to model selection. Recently, different approximations of the coalescent have been proposed; for example, an adaption using a hidden Markov model and the sequential coalescent (41)(47)(74).

1.2.3 Inference and data simulation based on the coalescent

We have introduced the coalescent process and some of its extensions to more realistic genetic models. In this section, we will introduce some of the uses of the process. First, we will see how it can be used to simulate datasets under the basic Wright-Fisher model with mutation, and under a split model. Afterwards, we will explain how we can make inferences about parameters using a clever recursion proposed by Ethier and Griffiths (13). And finally we will see how we can estimate the likelihood using importance sampling.

1.2.3.1 Simulating data under the coalescent

To investigate properties of samples under various models, or to test a new method, being able to simulate data using the coalescent is useful. Luckily, it is rather simple to simulate sequence data under the coalescent.

Using the infinite sites model for the mutation, suppose we want to simulate a sample of n sequences from a population with a constant size N over time and a scaled mutation rate of θ. Time is rescaled such that one unit is equivalent to N generations. We will first

simulate the genealogy of this sample, going backwards in time, and then follow the tree of the genealogy forward to end up with our sample of sequences.

Remember that coalescence events and mutation events are independent, and when we have k lineages remaining the time before a coalescence or a mutation follows an exponential distribution of rate $\binom{k}{2}$ or $k\theta/2$, respectively. The time for the next event—coalescence or mutation—will then be exponentially distributed with rate $k(k-1+\theta)/2$, and the event will be a coalescence with probability $(k-1)/(k-1+\theta)$, otherwise a mutation.

An algorithm for simulation under this model is :

$k \leftarrow n$

while $k > 1$ **do**

Simulate a time T_k for the next event where $T_k \sim \text{Exp}(\binom{k}{2} + k\theta/2)$

With probability $(k-1)/(k-1+\theta)$ it is a coalescence, otherwise it is a mutation.

if The event is a coalescence **then**

Choose a pair of lineages to coalesce.

$k \leftarrow k-1$

else

Choose a lineage and add a mutation to it.

end if

end while

To obtain the sample of sequences, we just need to go from the top of the tree to its leaves, one at a time. When a mutation happens on a lineage this means that the sequences at its leaves will have it. This is represented by Figure 1.7. For example, starting from the top and going down towards the sequence 2, we encounter the orange and blue mutations only and so consequently sequence 2 has those mutations. The positions of the mutations on the sequences are not relevant since no recombination can occur. But, if more than one sequence shares a mutation, it makes sense to put this mutation at the same place on all the sequences sharing it, as in Figure 1.7.

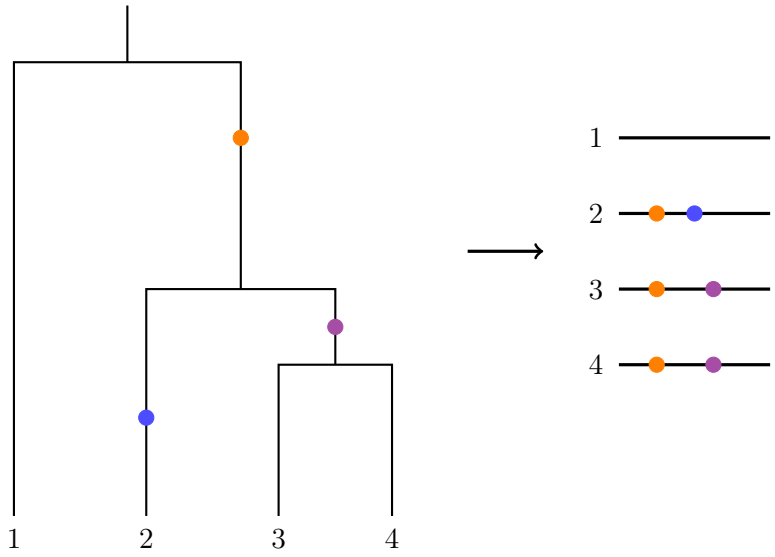


Figure 1.7: Simulation of a dataset under the coalescent. First, a mutation happens on lineage 2, then a coalescence of lineages 3 and 4, etc. After the genealogy is built, we can deduce the four sequences in the sample by looking at which mutations are shared.

The sequence of events of the genealogy is independent of the branch length (*i.e.* the time of the events). Therefore, it is possible to simulate only this sequence of events ("jump chain"), and if needed we can always simulate the time of each event later. To simulate a dataset under a split model with no migration, we need to specify the time of divergence of the two populations T, and we also need to simulate the time of each event. If the time is rescaled according to the ancestral population size N_a, we need to know the ratio N_a/N_i for the two descendant populations ($i \in \{1, 2\}$) in order to adapt the coalescence rate in those populations. Since no migration occurs, the genealogies in each descendant population before T will be independent of each other, and can be simulated separately.

The idea is to run the previous algorithm three times, but for the descendant populations the stopping point will be when the sum of the time of each event is greater than T. Once the algorithm has been run on each subpopulation, we run it on all the remaining lineages in both subpopulations combined.

1.2.3.2 Inference under the coalescent

The coalescent process is an interesting tool to use for parameter inference. The aim is to look at all the genealogies consistent with the sampled sequences because these genealogies hold information about the parameters of interest. For very small samples, it is possible to enumerate all the possible genealogies (without specifying the times of events) agreeing with the sample, Ethier and Griffiths (13) proposed a clever recursion function to do this. However, as the sample size increases, the space of possible genealogies explodes, though it is still finite. This is where the approach of importance sampling becomes very useful. We present the method of Griffiths and Tavaré (20) and its improvement by Stephens and Donnelly (64).

Both approaches are based on Ethier and Griffiths's recursion. We present this for a sample of labelled sequences and then derive a recursion for unlabelled sequences. Labelled sequences are all considered distinct from one another; unlabelled sequences, however, are grouped by type. Two sequences are of the same type if they share the same alleles at each SNP. Before presenting the recursion function, it is necessary to introduce some notation. The sequences are stored by type in a matrix $\mathbf{M}$ in which the entry m_{ij} is the allele of the i^{th} type of sequence at the j^{th} SNP. Therefore, each row of $\mathbf{M}$ represents a type of sequence and each column a SNP. We store the number of sequences of each type in a vector $\mathbf{n}$, where n_i is the number of sequences of type i. The total number of sequences is denoted by $n = \sum_k n_k$.

On the pair ($\mathbf{M}$, $\mathbf{n}$) we apply two types of operations that modify them to represent either a coalescence event or a mutation event back in time. Let $\mathcal{C}_i$ represents a coalescence of two sequences of type i. Therefore, $\mathcal{C}_i(\mathbf{M}, \mathbf{n})$ will return the same matrix $\mathbf{M}$ but with one subtracted from n_i. We represent a mutation event on a sequence of type i that becomes a sequence of type j by $\mathcal{M}_i^j$. The only difference between the sequences of type i and type j is therefore that i has one mutation fewer. The result of $\mathcal{M}_i^j(\mathbf{M}, \mathbf{n})$ will depend on sequence i. If a sequence of type j already exists in $\mathbf{M}$, then to obtain the resulting

matrix we need to remove the i^{th} row of $\mathbf{M}$ as well as the i^{th} element of $\mathbf{n}$ and then add one to the j^{th} element. If there is no sequence of type j in $\mathbf{M}$, then only the element of $\mathbf{M}$ where the mutation happened changes to an ancestral allele.

We need to establish when we can use those operations, *i.e.* when coalescence events and mutation events are possible. Only sequences that are of the same type can coalesce. Therefore n_i must be greater than one to observe a coalescence of two sequences of type i ($\mathcal{C}_i$). Since we are using the infinite sites model for the mutation, going backwards in time we can remove a mutation only when one type of sequence carries it and when there is only one sequence of that type. That is, to apply $\mathcal{M}_i^j$ at the s^{th} SNP, n_i must be equal to one and for all k, m_{ks} must be equal to zero, except of course when $k = i$.

The idea of the recursion proposed by Ethier and Griffiths is to define the probability of a pair ($\mathbf{M}$, $\mathbf{n}$) as the summation over all possible events one step back in time conditioned on those events. That is, if we let $\mathcal{E}$ denote a possible event, then:

$$p(\mathbf{M}, \mathbf{n}) = \sum_{\mathcal{E}} p(\mathbf{M}, \mathbf{n}|\mathcal{E})p(\mathcal{E}). \tag{1.2}$$

The conditional probability $p(\mathbf{M}, \mathbf{n}|\mathcal{E})$, is equivalent to the probability of the remaining sequences after the event $p(\mathcal{E}(\mathbf{M}, \mathbf{n}))$. This probability can also be rewritten as a recursion using Equation 1.2. The possible events can be partitioned into two classes: mutation events and coalescence events. Let A be the set of all types of sequences for which a mutation can occur. Then we can rewrite Equation 1.2 as:

$$\begin{aligned} p(\mathbf{M}, \mathbf{n}) =& \frac{n-1}{n-1+\theta} \sum_{i:n_i>1} \frac{n_i(n_i-1)}{n(n-1)} p(\mathcal{C}_i(\mathbf{M}, \mathbf{n})) \\ &+ \frac{\theta}{n-1+\theta} \sum_{i:i\in A} \frac{1}{n} p(\mathcal{M}_i^j(\mathbf{M}, \mathbf{n})). \end{aligned} \tag{1.3}$$

The probabilities of observing a coalescence event and a mutation event are given by the terms that precede the summation symbols. Knowing that the event is a coalescence, the

probability that it involves a particular pair will be $1/\binom{n}{2}$, since there are $\binom{n}{2}$ possible coalescence events in total. We are interested in coalescence of sequences of type i, and because there are $\binom{n_i}{2}$ of those pairs the probability of observing a coalescence of sequences of type i is given by the term after the first summation symbol in Equation 1.3. However, conditioned on a mutation event, the probability that this event happens to any of the sequences is $1/n$, since all sequences are equally likely to mutate. This recursion can be used to define the likelihood function of a sample of sequences.

The probability of a sample can be viewed as a sum over all the possible trees, or genealogies, that could have generated it. A tree can be partitioned into two parts: its history, *i.e.* the chronological order of the events (mutation and coalescence events) forming this tree (denoted here by $\mathcal{H}$ for history); and the branch lengths, *i.e.* the times between events (denoted here by $\mathcal{T}$). We are using the times between events instead of the times of the events because we know that they are exponentially distributed. If D represents the sample of sequences, then:

$$p(D) = \sum_{\mathcal{H}} \int_{\mathcal{T}} p(D|\mathcal{H}, \mathcal{T}) p(\mathcal{H}, \mathcal{T}) d\mathcal{T} \tag{1.4}$$

represents the likelihood function of D. The probability $p(D|\mathcal{H}, \mathcal{T})$ does not depend on the times of the events $\mathcal{T}$, but only on the events themselves and will be equal to one if the genealogy $\mathcal{H}$ could have generated the sample D or else it is equal to zero. The joint density $p(\mathcal{H}, \mathcal{T})$ can be rewritten as $f(\mathcal{T}|\mathcal{H})p(\mathcal{H})$ as the times of the next event do not influence the probability of the next event, but the times do depend on the number of lineages remaining, and therefore on the genealogy. Equation 1.4 can be rewritten as:

$$p(D) = \sum_{\mathcal{H}} p(D|\mathcal{H}) p(\mathcal{H}) \int_{\mathcal{T}} f(\mathcal{T}|\mathcal{H}) d\mathcal{T}. \tag{1.5}$$

Since there is no restriction on the times of the events and $f(\mathcal{T}|\mathcal{H})$ is a density function, the integral is equal to one. Thence, the likelihood of a sample D is equal to $\sum_{\mathcal{H}} p(D|\mathcal{H})p(\mathcal{H})$.

A history, or genealogy, $\mathcal{H}$ can be described as a sequence of events. A summation over all possible histories can also be viewed as a summation over the possible events one step back, which is similar to Equation 1.2. Therefore, the likelihood can now be expressed using the recursion function 1.3 of Ethier and Griffiths.

To be able to write down the likelihood we need to modify our notation. We need to describe the sequences remaining after an event. Let $(\mathbf{M}_k, \mathbf{n}_k)$ represent the sequences remaining in the genealogy after the k^{th} event (mutation or coalescence). Therefore, the sample of sequences will be represented by $(\mathbf{M}_0, \mathbf{n}_0)$, and the likelihood function is simply $p(\mathbf{M}_0, \mathbf{n}_0)$ (and calculated using Equation 1.3).

Importance sampling As mentioned before, the likelihood function rapidly gets cumbersome to evaluate explicitly. One solution is the use of importance sampling to obtain an estimate of the likelihood function. We have previously seen that the likelihood function can be written as:

$$L(\theta) = \sum_{\mathcal{H}} P(D|\mathcal{H}, \theta) P_\theta(\mathcal{H}), \tag{1.6}$$

where the summation is taken over all the possible genealogies of a sample of sequences of size n. Remember that D is the sample of sequences and $\mathcal{H}$ represents a genealogy. The probability $P(D|\mathcal{H}, \theta)$ will be equal to one if $\mathcal{H}$ is consistent with the sample, otherwise it will be equal to zero, and $P_\theta(\mathcal{H})$ is the distribution of those genealogies. We saw in the last subsection that we can simulate genealogies under this distribution; therefore, we can use the likelihood estimate:

$$L(\theta) \approx \frac{1}{M} \sum_{i=1}^{M} P(D|\mathcal{H}^{(i)}, \theta), \tag{1.7}$$

where $\mathcal{H}^{(i)}$ is a genealogy simulated according to $P_\theta(\mathcal{H})$. The main problem with this approximation is that nearly all of the genealogies simulated would have $P(D|\mathcal{H}^{(i)}, \theta) = 0$. Therefore, this approach would be useless for the estimation of the likelihood in practice

because when we simulate a genealogy backward in time using P_θ, the only information we have is the sample size. We do not use the data to simulate a genealogy. The sequences in the sample are defined only after the genealogy is built (see Figure 1.7). It is important to understand how Equation 1.3 differs when the likelihood is not defined as a sum over all the possible genealogies, but rather decomposed into steps (where in each step we are summing over all the possible events that could have happened).

Therefore, we need to find a way to simulate histories that agree with the observed data and use those simulations to estimate the likelihood and do parameter inference. Importance sampling allows us to do just this and is based on the observation that:

$$\begin{aligned} L(\theta) &= \int P(D|\mathcal{H}) \frac{P(\mathcal{H})}{Q(\mathcal{H})} Q(\mathcal{H}) d\mathcal{H} \\ &\approx \frac{1}{M} \sum_{i=1}^{M} P(D|\mathcal{H}^{(i)}) \frac{P(\mathcal{H}^{(i)})}{Q(\mathcal{H}^{(i)})}, \end{aligned}$$

where $Q(\cdot)$, is any distribution on genealogies with support $\{\mathcal{H} : P(D|\mathcal{H}) = 1\}$. We need to find a good distribution $Q(\cdot)$, called the proposal distribution, to build genealogies that are always consistent with the sample of sequences.

Griffiths and Tavaré (20) suggested a proposal distribution proportional to the recursion of Ethier and Griffiths (13). In 2000, Stephens and Donnelly (64) presented a new importance sampling (IS) algorithm for full-likelihood based inference that assumed no recombination events. They showed that their method gave much lower variance in likelihood estimates than the method of Griffiths and Tavaré. We describe their importance sampling algorithm for infinite sites data. Their proposal distribution was based on the exact distribution of the parent-independent mutation model, a special case of the infinite alleles model.

For infinite sites data, we can easily enumerate all the events possible back in time from $(\mathbf{M}_k, \mathbf{n}_k)$. A coalescence event can occur only if at least one element of $\mathbf{n}_k$ is greater than one, that is if there are at least two sequences of the same type. A mutation event

can occur on a sequence of type i only if the i^{th} element of $\mathbf{n}_k$ is equal to one and if it is the only sequence carrying a certain derived allele. Therefore, we deduce that a sequence in $(\mathbf{M}_k, \mathbf{n}_k)$ cannot satisfy both of these conditions simultaneously.

Stephens and Donnelly suggested a simple algorithm: first randomly choose a sequence that meets one of these conditions and then perform the unique event involving this sequence. Based on this algorithm, we can deduce the proposal distribution by decomposing $Q(\mathcal{H})$ as $\prod_{i=0}^{\tau-1} q((\mathbf{M}_{i+1}, \mathbf{n}_{i+1})|(\mathbf{M}_i, \mathbf{n}_i))$ since we are building genealogies as we go backwards in time. Denoting the number of sequences satisfying one of the conditions in $(\mathbf{M}_i, \mathbf{n}_i)$ by n^* , we obtain the following proposal distribution to build a genealogy step by step:

$$q((\mathbf{M}_{i+1}, \mathbf{n}_{i+1})|(\mathbf{M}_i, \mathbf{n}_i)) = \begin{cases} \frac{n_\alpha}{n^*} & \text{if coalescence of 2 sequences of type } \alpha \\ \\ \frac{1}{n^*} & \text{if mutation.} \end{cases} \tag{1.8}$$

This algorithm specifies that each of the n^* sequences has the same probability to be picked, $(1/n^*)$. The probability to perform one $\mathcal{C}_\alpha$ is then $\sum_{i=1}^{n_\alpha} \frac{1}{n^*}$. This proposal distribution does not involve any unknown parameters when sampling a possible history (for example the mutation rate θ), and it also does not simulate the times of each event.

Labelled or unlabelled sequences So far we have presented the likelihood function of a labelled sample (often called ordered). However, there is in fact no need to keep labels on the sequences because it does not add meaningful information and usually the unlabelled version of the likelihood is used. In the context of a split model, we will see in the next chapter that using unlabelled sequences slightly complicates the estimation of the likelihood. Nevertheless, we think it is interesting to look at the two versions.

First, let the probability of a labelled sample be renamed $p^o(\mathbf{M}_0, \mathbf{n}_0)$ and use $p(\mathbf{M}_0, \mathbf{n}_0)$ for the probability of an unlabelled sample. Then to obtain the likelihood of an unlabelled

sample we simply need to multiply the likelihood of a labelled sample by the number of possible distinct ways we can label this sample. Therefore,

$$p(\mathbf{M}_0, \mathbf{n}_0) = \frac{n!}{n_1! n_2! \cdots n_d!} p^o(\mathbf{M}_0, \mathbf{n}_0)$$

since there are $n!$ ways to permute n sequences and $n_i!$ ways to permute identical sequences of type i, where d is the number of different types of sequences. The coefficient in front of p^o represents the number of all possible permutations of the labels of the sequences in a sample. Including the number of all possible permutations in Equation 1.3, the likelihood of an unlabelled dataset is:

$$\begin{aligned} p(\mathbf{M}_0, \mathbf{n}_0) = & \frac{n-1}{n-1+\theta} \sum_{i:n_i>1} \frac{(n_i-1)}{(n-1)} p(\mathcal{C}_i(\mathbf{M}_0, \mathbf{n}_0)) \\ & + \frac{\theta}{n-1+\theta} \sum_{i:i\in A} \frac{n_j+1}{n} p(\mathcal{M}_i^j(\mathbf{M}_0, \mathbf{n}_0)). \end{aligned} \tag{1.9}$$

Note that the added term does not depend on the parameter of interest. Therefore, using labelled or unlabelled sequences should not affect parameter inference. We mention also that the proposal distribution of Stephens and Donnelly (1.8) does not need to change when using unlabelled sequences.

In this likelihood, the probabilities that the next event is either a coalescence or a mutation is still written before the summation symbols. However, the new probabilities after the summations can now be described by a forwards in time argument; the standard coalescent can be worked forwards in time, from the past to the present, using a diffusion process (see for example (64)). Starting from the lineage of the MRCA, a coalescence event is seen as a split, creating a new identical lineage.

Thinking forwards in time, and knowing that a coalescence event is observed, the probability that it involves a sequence of type i is simply $(n_i - 1)/(n - 1)$ (*i.e.* the probability that one of the $n_i - 1$ sequences remaining after the event has split to form

an additional lineage, since we know that a coalescence has occurred and that one of the $n-1$ lineages remaining splits after the event). Similarly, knowing that a mutation event has happened to one of the n sequences, the probability of a certain mutation depends on the number of sequences of the same type as the one created after the mutation (when going backwards). If currently there are n_j sequences of type j and a mutation event from a sequence of type i creates one more sequence of type j ($\mathcal{M}_i^j$), then any of the $n_j + 1$ sequences could have undergone a mutation event in the future, giving the probability $(n_j + 1)/n$. It is important to note that this forwards–in–time description is only valid when the process has attained equilibrium, and therefore it is not valid when considering a population split model.

In this section, we have described 1) how we can use the coalescent process to simulate data, and 2) the basis for inference on the parameters of the model in the context of the standard coalescent with the infinite sites mutation model. We will use this in the next chapter to suggest a proposal distribution for an isolation split model.

1.3 Survey of existing methods

Large genomic projects are now offering access to high quality data from a vast number of individuals. It is a challenge to develop methods that can use all of this data. Advances in computing have helped greatly, but the massive amount of available data is still practically challenging to analyse. Different methods have been proposed to scale to modern massive data sets. We will concentrate our survey on coalescent based methods that try to understand population structure within a species, focussing particularly on methods that estimate the time of population divergences or that estimate variable population sizes.

Every method uses some simplifications of the data analysed or to the model used. We can identify three categories of methods. First, we have the methods that use simulations of the coalescent process to estimate the likelihood of the data. The idea behind these

methods is to build possible genealogies for the data and to use these to obtain an estimate of the likelihood of the parameters of interest. Usually, these methods will use less data: for example, a region of the genome or multiple independent short regions. This can allow the simplification of the coalescent process by not modelling recombination.

The second category consists of methods that use an approximation of the coalescent process to simplify the simulation of possible genealogies. Those methods usually use a more complex model: for example, the model can include recombination. The amount of data used will again vary: it can be the whole genome of one individual or multiple large regions for multiple individuals.

Finally, the methods in the last category use different statistics to summarise the data. They then compare the summarised data to summary statistics of simulated data. Therefore, these methods do not require that the simulated genealogies match the data exactly, but only the statistics used. In this survey, we will cover chronologically the main methods in each category and explain their characteristics and strengths.

1.3.1 Methods based on the coalescent

In the preceding section, we have seen that the number of possible genealogies of a sample explodes. Even for simple models, the exact likelihood gets rapidly impossible to evaluate directly, and if recombination events are allowed in the model, the number of possible genealogies becomes infinite. Therefore, the methods that use the coalescent directly to estimate parameters of interest will usually estimate the likelihood via Monte Carlo methods, building only a subset of the possible genealogies of a sample. Some methods will use the exact likelihood, but only a few individuals to simplify the possible genealogies. Nevertheless, all the methods presented in this category ignore recombination events in their model and most of them will use multiple independent loci.

Nielsen method (1998)

In 1998, Nielsen (50) adapted the importance sampling method of Griffiths and Tavaré (20) to a model of two population splitting, also called an isolation model since it does not allow migration between the sampled populations. The method directly estimates the time of divergence (T) and uses the infinite sites model of mutation with no recombination. The genealogies are built backward in time conditionally on the time of divergence. The times of the events need to be simulated, because while the time of the next event is more recent than the time of the split, no coalescence events are allowed between pairs of lineages of different populations. The first time the next event is more ancient than the time of divergence, the remaining lineages in both descendant populations are grouped together and can then be involved in a coalescence event. The likelihood function can then be estimated using importance sampling for different values of T. Nielsen also proposed an adaptation for multiple populations. If there are r populations, then the genealogies are built conditionally on all the $r-1$ different times of divergence. By trying different values for the times of divergence, the most likely phylogeny can be found.

In his paper, Nielsen did not use simulated data to validate his method, but it was used on two small datasets: one composed of two African populations, and the other from three different turtle sub-species. The datasets were small: around fifty sequences of ten SNPs. All population sizes were assumed to be equal, reducing the number of parameters to estimate. There is no program implementing the method available, and the method was not extended, as it was probably too computationally intensive at the time.

IM, IMa, IMa2

Afterwards, Nielsen worked with others on the development of a method called IM for the Isolation with Migration model. The method was introduced in 2001 by Nielsen and Wakeley (52) and extended and improved in 2004 and 2007 (renamed IMa) by Hey and Nielsen (27) and (28). The method estimates the time of divergence, the migration rates

between the two populations and the constant population sizes. In IM the population sizes can also be modelled as an exponential function, but not in IMa. It uses phased data, either sequences of SNPs or micro-satellites, and can use multiple loci. The method assumed no recombination within a locus and free recombination between loci. The parameters are estimated using a Markov chain Monte Carlo (MCMC) method to sample genealogies instead of importance sampling.

The idea behind MCMC is to create a Markov chain for which the equilibrium distribution will be equivalent to the sampling distribution of interest. The state space of the Markov chain is the set of parameters of interest. In the long run, and when the chain is at equilibrium, the amount of time the chain spends in each state should be proportional to the posterior distribution of the parameters. The Metropolis-Hasting algorithm allows the proposal of the next state of the chain and the step is accepted or rejected based on the probabilities of the current and proposed states. Consecutive steps of the chain are correlated, but sampling equidistant steps at large enough intervals allows for approximately uncorrelated sampling, as long as the chain has reached equilibrium.

In a bayesian setting we wish to estimate $f(\Theta|D) = cf(D|\Theta)f(\Theta)$, where c is a normalising constant, Θ is the set of all the parameters and D is the data. We have seen previously that $f(D|\Theta)$ is intractable unless we integrate over all the genealogies. In the method IM a genealogy does not include mutation events, but does include migration events; to avoid confusion we will represent a genealogy by G. Therefore we have:

$$f(\Theta|D) = cf(\Theta)\int_G f(D|G,\Theta)f(G|\Theta)dG. \tag{1.10}$$

IM uses the Metropolis-Hasting method to integrate over the genealogies and to obtain an estimate of the posterior distribution (or likelihood function). At each step, a small update of the current genealogy or of one of the parameters is proposed. Updates of the genealogy consist of picking a branch at random, and attaching it somewhere else on the

tree, and resampling migration events on this branch. The number of parameters to update is quite large, and can make it hard to assess if the chain has reached stationarity. In 2007, Hey and Nielsen improved their method by analytically integrating out the population size and migration parameters, thereby needing only to sample the genealogy and the time of divergence via MCMC.

Both IM and IMa are adaptable to different mutation models (infinite-sites, stepwise). Though no recombination is allowed within a locus, the method can make use of multiple unlinked loci. In 2004, IM was tested on simulated datasets of twenty sequences of five loci (each composed of multiple SNPs). The method seemed reliable and better results were expected for larger datasets. It is also possible to test nested models using the likelihood ratio test, for example to test the hypothesis of no migration between the descendant populations. Recently, Hey (26) extended the method, now called IMa2, to include the possibility of analysing multiple populations. Some drawbacks of the improvement are the need to know *a priori* the phylogeny of the populations and the difficulty to determine whether the chain has converged and is well mixed. Both softwares are available and have been tested and used in different studies((25)(17)(8)(65)(58)).

The extended bayesian skyline plot

In 2008, Heled and Drummond (24) presented the extended bayesian skyline plot (EBSP) which estimates the variable size of a population. It is the latest of the skyline-plot methods (see (29) for a review of these) and the first to allow the use of multiple non-recombinant loci. The method estimates the size of a population as a smooth function. It models only one population and, in particular, it does not infer time of divergence of multiple populations. The first versions of the skyline plot assumed that the genealogy was known and from it inferred the population size as piecewise constant within time intervals determined by the coalescent. The first version estimated the population size for each interval between coalescence events and was therefore noisy. A second version of the skyline plot used a

concatenation of consecutive coalescence time intervals to obtain a smoother estimation of population sizes.

The EBSP uses an MCMC algorithm to find the limits of the time intervals (spanning multiple coalescence event) and for each intervals modelled the population size as a linear function conditionally on the population size of the previous interval. The genealogy of each locus is estimated at different steps of an MCMC algorithm and they are used to obtain the posterior distribution of the population sizes. The method is implemented in the software package BEAST (12) and can use any of the methods included in the software to sample the genealogies. It can make use of different types of loci and can use a non constant mutation rate. The authors tested their method with simulated and real data. They used no more than 32 loci and 10 individuals. The advantage of using more than one locus was clearly demonstrated, as it allows the more ancient estimation of the population size in the presence of a strong bottleneck. This was not possible with the previous version.

Wang and Hey (2010)

The methods presented so far have chosen to use fewer loci but more individuals. IM and its variations are able to use around hundreds of loci for hundreds of individuals. Wang and Hey (72) in 2010 decided to take a different route and they presented a method that is able to use thousands of loci but only pairs of haplotypes for each locus.

The method uses an isolation with migration model without recombination and the infinite sites mutation model. For three populations, it estimates the three population sizes (assuming they are constant), the two migration rates and the times of divergence. For each locus, two haplotypes are sampled; they can either be from the same population or from different populations. Using only two haplotypes per locus means that the genealogy is of a simple form with only one coalescence event. Therefore, the method can calculate precisely the likelihood by numerical integration over all the possible genealogies.

A simulation study showed that the method works well if many loci are used, otherwise the estimates are biased. Another requirement is that a good repartition of the origin of the sampled haplotypes is observed. In other words, the data must have loci with haplotypes sampled from the same population (from both populations) and some loci with one haplotype per population. The simulated data consisted of ten thousand loci of 1kb each. The method was also applied to two species of Drosophila, and the results obtained were similar to a previous study. It was noted that a small migration rate between the two species was detected, and that ignoring migration leads to biased estimates of the ancestral population size. The authors argued that using few individuals at many loci will give better results for ancient times of divergence.

G-PhoCS

In 2011, Gronau *et al.* presented the method G-PhoCS (21) (Generalised Phylogenetic Coalescent Sampler) for the inference of the times of divergence, the effective population sizes and the migration rates of multiple populations. It uses one individual per population genotyped at tens of thousands of loci. The method builds on the work of Rannala and Yang (60) and included the possibility of adding bands of migrations between populations.

The method was applied to five human populations and assumed that the phylogeny is known. The authors used whole-genome sequences of six individuals and created their own pipeline for genotype inference. They used around 37,000 loci of length 1kb and chose them to be able to ignore recombination events within loci. To estimate the posterior distribution of the parameters, possible genealogies are built for each locus using an MCMC algorithm. The method integrates over all possible phases when computing the probability of a genealogy. Interesting results of human divergence were obtain with confidence intervals narrower than previous estimates. We will discuss those results in the last chapter of this thesis.

1.3.2 Methods based on an approximation of the coalescent

The previous methods did not take into account recombination or use a larger amount of data since it becomes computationally intractable. However, different strategies have been developed to get around these issues. One possibility is to find a suitable approximation of the coalescent process that is easier to compute but keeps some important features. Two really important approximations have been presented nearly ten years ago. Both have been used greatly in statistical genetics modelling and have been applied to estimate demographic parameters. We will first give a short description of those two approximations, then we will present some of the methods that have been developed to estimate the time of divergence or variable population sizes using those approximations.

The main idea shared by both approximations is to see the coalescent with recombination as a stochastic process along the sequence. Wiuf and Hein (74) first presented a description of the sequentially coalescent process with recombination. The idea is to start with a genealogy for the first position and then move along the sequences letting the genealogy change with recombination events. Unfortunately, the sequential coalescent is not Markovian as one needs to remember all the previous genealogies to evaluate the probability for next position on the sequences.

In 2003, Li and Stephens (39) proposed the product of approximate conditionals model (PAC) that approximate the probability of observing a new haplotype conditional on those already sampled. The probability of a sample of haplotypes $h_1, \ldots, h_n$ conditional on some parameters Θ can be seen as:

$$Pr(h_1, \ldots, h_n|\Theta) = Pr(h_1|\Theta)Pr(h_2|h_1;\Theta)\ldots Pr(h_n|h_1, \ldots, h_{n-1};\Theta). \tag{1.11}$$

Those conditional probabilities are intractable, therefore different approximations have been proposed over the years (64)(15). Li and Stephens proposed an approximation based on seeing the newly sampled haplotype as an imperfect mosaic of the previous haplo-

types. The idea is to start at the first position of the sequence and randomly choose from which previous haplotypes the new one will be copying. Error in the copying process is allowed and represents mutation events. Then moving along the sequence, the haplotype from which we are copying can change to reflect recombination events. A hidden Markov model along the sequence can be used to model this copying process, where the hidden state is the identity of the haplotype that is copied at this position. The probability $Pr(h_i|h_1, \ldots, h_{i-1}; \Theta)$ can then be approximated by the hidden Markov model using the standard forward algorithm. One of the main drawbacks of the likelihood function based on the PAC is its dependence on the order of the haplotypes. Therefore, the authors suggested to use the mean over a number of permuted orders of haplotypes.

Another important approximation of the coalescent is the sequentially Markovian coalescent presented by McVean and Cardin (47). This approximation is really close to the coalescent with recombination. An ancestral recombination graph (ARG) is built along the sequence by allowing recombinations that change the topology of the tree at some positions. McVean and Cardin showed that this approximation is equivalent to a coalescence events of lineages that do not share regions of ancestral material. Remember that a recombination event introduces non-ancestral material on the two remaining lineages, going back in time. When a sequence undergoes recombination, one of the resulting lineages will carry its material on the left side of the recombination point, while the other lineage will carry the remainder of the sequence on the right side of the recombination. We refer to the original material as ancestral. Non-ancestral material can coalesce with any sequences without restrictions.

Davison, Pritchard and Coop (2009)

Davison *et al.* (9) presented in 2009 the first method allowing recombination in a population splitting model (not using summary statistics). The method uses the isolation model, does not include migration and uses the infinite sites mutation model. For simplicity the three

population sizes are assumed to be equal. The approximation of the likelihood is based on the product of approximated conditional likelihoods as developed by Li and Stephens in 2003 (39). In the method of Davison *et al.*, the hidden states include the level at which the copying is made. There are two different levels, either in the ancestral population or in the descendant populations. The likelihood is estimated conditionally on the time of divergence and is evaluated over a grid of possible values.

The method was developed for two different types of data. First, for unlinked data, like a set of dispersed SNPs or micro-satellites, for which there is not need to model recombination. This is similar to using independent loci, except that only one SNP or micro-satellite is used. The method was tested using one hundred datasets of twenty sequences (ten per population) with sixty SNPs. The method performs well with unlinked data but, as mentioned by the authors, this type of data offers only incomplete information about the genealogical process.

The second type of data considered is what the authors call loosely linked data, which consists of loci for which recombinations event are allowed. The method was tested using the same datasets, but now seeing those SNPs as dependent. The recombination rate was estimated by the method. The estimates of the recombination rate are biased and this effect also biases the estimates of the time of divergence. The authors tried to correct for the bias, but mentioned also that the correction probably needs to be tailored to the dataset used. They also give some ideas of how migration events could be included in their method.

coalHMMs

Mailund *et al.* (41) developed in 2011 a new method using a coalescent Hidden Markov model to estimate the time of divergence of two populations. Their method assumes the sequentially Markovian coalescent and allows for recombination but not for migration. Only one haplotype per population is used to build the genealogies for pairs of adjacent

nucleotides. The method uses two Markov chains: one that models the distribution of the coalescence times using a discrete space Markov model and another to model the genealogy of adjacent nucleotides back in time as a continuous time Markov model with finite states. Looking only at the genealogy of two adjacent positions on two haplotypes, there are fifteen possible states. The two adjacent positions can be linked only if they are on the same haplotype — coalescence and recombination events can make or break those links. The two same positions can find a shared ancestor and then there is only one copy left of this position in the genealogy. Then a simplification is done by dividing the continuous time into fixed intervals. As mentioned by the authors: "the first model is used as the hidden Markov model when estimating parameters, while the second is used to compute the transition probabilities of the first".

The method has been tested using one hundred simulated datasets of 500 kb. The results were biased and the authors found that to obtain less biased estimates, a greater number of time intervals is needed; fifteen intervals improved largely the bias compared to five intervals. The estimates of the recombination rate have always a large negative bias. From all the parameters estimated, the time of divergence was the one with the least bias. The method was also applied to whole genome sequences of two sub-species of orang-utan. All the parameters were estimated for regions of 1Mb, then box-plots of the results per chromosomes were used to obtain the final estimates. The results were consistent with recent results of another study. The method is limited in the number of sequences it can uses : for three sequences two hundred and three states are needed and with four sequences more than four thousand.

The method was extended in 2012 by Mailund *et al.* (42) to include migration events. The model assume that migration followed the split for a certain time, allowing for a more gradual split. A new parameter for the time at which the migration events stop is inferred. Therefore, going back in time, under this model we have a clean split with no migration between the two populations, then at time T_2 migrations are allowed between

the populations, and at time T_1 the populations merge to form the ancestral population. The state space gets much larger, and a computer algorithm is used to generate the state space and the rate matrix. The method was used to infer split times of different species of Great Apes using multiple regions of 10Mb. Their results showed evidence of a clear split between bonobo and common chimpanzee, while the split between the gorilla and orang-utan species seems to have been more gradual with gene flow occurring over several hundred thousand years.

PSMC

In 2011, Li and Durbin (38) presented a method to infer variable sizes of a population through time. It has some similarities with the method of Mailund *et al.*, as it looks at the distribution of the coalescence times along the genome using the sequentially Markovian coalescent. But instead of looking at two haplotypes from two different populations, it uses the genome of one individual, and is therefore named the pairwise sequentially Markovian coalescent (PSMC). The distribution of the coalescence time is then directly related to the coalescence rate and therefore to the population size.

The observations in the model are the hetero- or homozygosity of the positions and the hidden states are the discretised times (epochs) to the most recent common ancestor of each pair of alleles of an individual. Transitions between the hidden states occur through recombination events. Therefore, there is no assumption about how the population size has changed in the past, since it can vary freely in between epochs. An EM algorithm is used to find the estimates of the parameters, *i.e.* the scaled mutation and recombination rates and the piece-wise constant population sizes.

Through simulations using one hundred sequences of 30 Mb, the method was shown to perform well from around 20,000 years ago to 3 millions years ago. For more recent or older times, there are not enough coalescence events happening to get good estimates. The method seems to also have some difficulties to infer some of the drastic changes in

population sizes, but overall it seems to get a good sense of how the population size had varied through time. The estimate can be rescaled to real time given a certain mutation rate and generation time.

The method was applied to real human genomes from different populations, analysing one individual at a time. The sequences are divided in 100bp bins and if one position inside a bin is heterozygous then the bin is labelled as a heterozygote. Also, if more than 90 bases were uncalled then the bin is labelled as missing. The method gives interesting results about how the size of different human populations has changed, and we will discuss those results in the last chapter of this thesis. The authors tried also to infer the time of divergence of different pairs of populations by observing the time when their population size histories diverged. A program implementing the method is available and has been used to infer population sizes in various studies.

diCoal

We explained previously the estimation of the probability of observing a new haplotype conditionally on some already sampled haplotypes of Li and Stephens (also referred as the conditional sampling distribution (CSD)). The approximation was inspired by the coalescent process but was not directly based on it. Paul, Song and Steinrücken have been working on an approximation of the CSD based on the coalescent and the diffusion process (56)(57). It has been adapted to an island model (63) and used in a method named diCoal to estimate the variable size of a population using multiple genomes (62).

The idea behind this approximation is to start by assuming that we know the real genealogy of the sequences already sampled and then ask what is the probability for a new sequence to join the genealogy. It is possible to describe the probabilities for a small number of sequences but the number of states grows exponentially with the number of loci (here a locus represent a SNP or a micro-satellite) and, of course, the real genealogy is not known. The idea of the authors is to use a trunk genealogy for the sampled haplotypes, which means

that we let all the lineages sampled go to infinity without coalescence, recombination or migration events. The approximation also uses the sequentially Markovian coalescent to estimate the CSD of a new sequence.

The method diCal uses an HMM and assumes that the mutation and recombination rates are given. As with PSMC the time needs to be discretised, and for each epoch the population size is inferred. The population size is modelled as piecewise constant and an EM algorithm is used to obtain the estimates per epoch. The method was used on simulated and real data, with around tens of individuals for one region of a few Mb. It gets better estimates than PSMC for more recent times since it can use more than one individual, but the results are more biased for older times.

1.3.3 Methods based on summaries of the data

Another class of methods are based on only summaries of the data. The idea is to reduce the dimensionality of the dataset into statistics that keep features of interest. The main reason to summarise the data is computational efficiency, which in turn allows the analysis of more data.

Wakeley and Hey

In 1997, Wakeley and Hey (70) presented a method to estimate the time of divergence of two populations and their effective population sizes, including the ancestral population size. The method is based on the categorisation of polymorphic sites into four groups. A site, or position, is considered polymorphic if two possible alleles are present among the sampled haplotypes. The four groups are: those that are polymorphic in both populations (S_s); those that are polymorphic in population i but not in population j (two groups S_{x1} and S_{x2}); and finally those that are fixed differences (all individuals in one population have the derived allele that is absent in the other population S_f). The total number of polymorphic sites in the sample S, is then equal to $S_{x1} + S_{x2} + S_f + S_s$. All these quantities

contain information about the population size and time of divergence. A divergence that occurred more recently will create more shared polymorphism compared to a more ancient time of divergence. Large population sizes and more ancient divergence times will create long terminal branches that will create fixed differences.

The isolation model (no migration) and the infinite sites mutation model are assumed. Using the Wright-Fisher model, the authors found the expected number of sites in each category using the parameters of interest. The four parameters are then estimated by solving numerically the systems of equations composed of the four expectations. Recombination events do not affect the expectation of the four statistics and will even lower the variance of the estimates. Therefore, there is no constraint on the data used regarding recombination. In fact, better results will be obtained if multiple independent loci are used, since it is only in the presence of recombination that sites in all four categories will be observed. One genealogy, represented by a tree, cannot included both fixed difference and shared polymorphism. By using multiple loci, different genealogies and categories of sites are observed.

The method was tested on simulated datasets composed of twenty sequences of ten loci (formed of multiple SNPs). Results showed that the data need to contain multiple independent loci to obtain good estimates. Also, the estimates of the ancestral population size are more biased and the time of divergence will be underestimated when the divergence is more ancient.

MIMAR

More recently in 2007, Becquet and Przeworski (5) developed a method called MIMAR that uses MCMC to estimate the parameters of an isolation with migration model allowing recombination. To facilitate computation the method uses the same summaries of the data used by Wakeley and Hey. The infinite sites mutation model is assumed as well as symmetric migration rates, and the data used are composed of multiple loci. For each

locus, a number of genealogies (ancestral recombination graphs) are built conditionally on the current parameter values. The method computes the sum of the lengths of the branches that could give rise to the different categories of polymorphism to evaluate the probability of the observed values of the four statistics conditionally on a genealogy.

MIMAR was tested and compared to IM using simulated data. They first tested the methods using an isolation model. The simulated datasets were composed of 20 haplotypes per population using either 20 or one hundred non recombining loci. Both methods performed well, with MIMAR giving slightly more accurate results on datasets with twenty loci and IM giving better results for the datasets with one hundred loci. IM was two to three times faster than MIMAR. Then the ability of MIMAR to detect gene flow was tested using 20 simulated data sets, each composed of 40 recombining loci, using both a model with migration and one without. Using the data sets without migration, the migration rate estimates were biased but close to 0. For the datasets with migration, MIMAR detected migration for 17 out of 20 data sets. The method was also applied to ape data.

In 2009, Becquet and Przeworski presented a paper (6) in which they tested the robustness of MIMAR and IM to model misspecification. They used simulated datasets under different models of divergence, with gene flow between the two descendant populations happening at different times. They found that both MIMAR and IM gave biased estimates of the time of divergence and ancestral population size.

PopABC

PopABC (40) is a method and software developed in 2009 by Lopes, Balding and Beaumont that uses an approximate Bayesian computation method (ABC) to estimate the parameters of an isolation with migration model. The idea behind ABC methods is to: 1) sample the parameters from the prior distribution, 2) simulate data using the sampled parameters, 3) summarise the data using a set of summary statistics, 4) repeat steps 1 through 3 for a large, fixed number of iterations. Finally, keep all the points in your parameter space that

have summary statistics close to the ones observed in your data. From those points you can obtain the posterior distribution of your parameters and therefore your estimates.

The coalescent simulator in popABC (described in (4)) can use the step-wise and the infinite-sites mutation models. It can be applied to multiple loci (recombinant or not). In theory the method can use multiple populations, but in practice the number of statistics used gets too large for more than three populations. It is also possible to compare different models to assess the presence of migration or to estimate the most likely topology (for more than two populations) by using an estimate of the Bayesian posterior probability.

The summary statistics used depend on the mutation model. For DNA sequences there are nine different statistics that are computed for each population as well as for all pairs of populations. A simulation study done by Beaumont *et al.* (4) showed that the method gave reliable results for a two population isolation model but showed also its limitations for more complex models (*e.g.* more than two populations with migrations included).

∂a∂i

In 2009, Gutenkunst *et al.* (22) presented a method named ∂a∂i for Diffusion Approximation for Demographic Inference. The method estimates the parameters of a model that can include multiple populations, migration events and exponential growth of the effective population sizes. It does so by comparing the observed joint allele frequency spectrum (AFS) to an estimate of the expected frequency spectrum under the model of interest. The AFS consists of the joint distribution of allele frequencies across binary variants. For a two populations model, the joint AFS is an $n_1 \times n_2$ matrix where entry (i, j) will be the number of variants for which the derived allele is observed exactly i times in population 1 and j times in population 2, and where n_i is the number of sequences sampled in population i.

The expected frequency spectrum of the fixed model is numerically computed using a diffusion approximation for a grid of possible parameter values. The method assumes that all the variants are independent and uses the diffusion approximation to the one-locus,

two-allele Wright-Fisher process. Composite likelihood is used to obtain an estimate of the likelihood over all the variants and confidence intervals are obtained via bootstrapping. In practice, the method can be applied to models with up to three descendant populations. The method was applied to real human data using a 5Mb region in 68 individuals from four populations. The results obtained will be discussed and compared in the last chapter of this thesis.

Tellier 2011

In 2011, Tellier *et al.* (68) presented a new method that builds on the work of Wakeley and Hey (70) and Becquet and Przeworski (5)(MIMAR). The method assumed a two populations isolation with migration model with data composed of multiple recombinant loci. The authors decided to try different ways to subdivide the summary statistics used by Wakeley and Hey and MIMAR. Their motivation was to use data from different organisms with higher intra-locus recombination rates and lower amounts of available data : they gave the example of two recently diverged species of wild tomatoes with only 7 to 13 genes available. In this situation a method that supposes no recombination is not adequate, and the summary statistics of Wakeley and Hey have limited power to distinguish models with migration.

They divided the class of shared polymorphism into singletons and doubletons, arguing that recent migrations will increase the number of shared singletons and doubletons. They tried four different partitions of the variants, with the coarsest including seven statistics and the finest containing twenty-three. To estimate the parameters of the model, the authors tried two strategies : a maximum likelihood method (similar to MIMAR) and a composite likelihood method assuming independence between variants.

Their different methods are compared to MIMAR, popABC and $\delta a \delta i$ in simulation studies. Multiple data sets for different parameter values, with either seven loci or 100 loci, were simulated. The results showed that the new method gave more accurate estimates

of the time of divergence and migration rates and was also faster, but for recent times of divergence all methods overestimate T and the migration rates.

1.3.4 Summary

The characteristics of the methods presented in this survey and our method are summarised in Table 1.1. The first column includes the reference of the methods, then the next six columns present the key features included in the model: 1)T, the time of divergence; 2)*Mig*, whether migrations are allowed; 3)*Rec*, whether recombination is allowed; 4) *Var pop size*, whether variable population sizes (through time) are estimated; and finally 6) if a software implementation is available. The last two columns give a short description of the type of data used and of the method itself. Note that *Exp.* means that the population sizes are modeled using the exponential distribution.

At first, methods based on the coalescent using importance sampling or MCMC had many drawbacks. The computation times were huge even for small datasets and, since amount of data available was limited, taking into account recombination was important to use as much of the data as possible. This has motivated the development of methods using summaries of the data. But these methods, although relatively fast, do not make use of all the information in the data.

Since the arrival of larger genomic projects, the development of new methods has taken a different direction by trying to approximate the coalescent process to include recombination events while using more of the information present in the data. Moreover, computers have become much faster, which enables the use of more complex models. But the amount of data is still too large to analyse exactly, and some simplifications are needed for all methods: choices need to be made.

We present a new method to estimate the time of divergence of two populations and their variable population sizes. From Table 1.1 and to our knowledge, we can see that no other method is able to jointly estimate both the time of divergence and the variable

population sizes. We have decided to use importance sampling to build possible genealogies and from those genealogies we estimate the population sizes, modelled as piecewise constant. We will be using only cold regions of the genome to infer genealogies for a sample of multiple individuals. Cold regions have low recombination rates and, therefore, we can assume that no recombination occurred within these regions. This enable us to ignore recombination events when building the genealogies. By using more individuals we aim to have enough information for more recent splits and population sizes. We have chosen to adapt the importance sampler of Stephens and Donnelly (64). This sampler does not allow recombination events and it has been proven to give better results than the importance sampling method of Griffiths and Tavaré (20).

The next chapter will present the importance sampler of Stephens and Donnelly, and our adaptation to a population split model. The validation of the method and the results of simulation studies will also be presented. In the third chapter, we present the extension of the model to variable population sizes and the results of simulation studies to test for robustness to model misspecification are presented in the fourth chapter. Finally, the last chapter presents the results we obtained when using real data from the 1,000 Genomes project.

Method	Model includes						Type of data	Description
	T	Mig.	Rec.	Var pop size	> 2 pop.	Software		
Nielsen(50)	✓				✓		1 locus, sequence (10 SNPs, $\approx$ 45 ind.)	IS (GT) Full Like.
IM(27), IMa(28), IMa2(26)	✓	✓		Exp. IM	IMa2	✓	seq., mutli. loci, microsat., multi. ind.	MCMC Full Like.
Ext. Bay. Skyline Plot(24)				✓		BEAST (12)	seq. of DNA bases, multi loci, multi ind.	MCMC, but needs to simulate the genealogies before.
Wang and Hey(72)	✓	✓					seq., multi loci, 2 ind. (10,000 loci of 1kb)	Full Like. Numerical Integration
G-PhoCS(21)	✓	✓			✓	✓	seq. of DNA bases, 37,574 loci of 1kb 6 ind.	Full like. MCMC
Davison(9)	✓		✓				seq. of SNPs, multi loci, multi. ind. (10/pop.)	Approx. like. PAC
coalHMM(41)	✓		✓				whole genome, 2,689 $\times$ 1Mb 2 ind.	Coal. approx. seq. Markov, HMM
PSMC (38)			✓	✓		✓	diploid whole genome. 1 ind.	HMM, coal. approx. seq. Markov
diCoal(62)			✓	✓		✓	one few Mb region, tens ind.	coal. approx.
Wakeley & Hey(70)	✓		✓				SNPs, multi. loci, multi ind.	Sum. Stats
MIMAR(5)	✓	✓	✓			✓	polymorphism data, multi. loci, multi. ind.	Sum. Stats MCMC
PopABC(40)	✓	✓	✓		✓	✓	seq. of SNPs, microsat, multi loci, multi ind.	Sum. Stats ABC
∂a∂i(22)	✓	✓	✓	Exp.	✓	✓	many SNPs, many ind. (20/pop)	Diffusion approx. of expected SFS of multi. pop.
Tellier *et al.*(68)	✓	✓	✓				SNPs, multi. loci, multi. ind.	Sum. Stats based on the JSFS
Our method	✓			✓			seq. multi. loci, multi. ind.	Full Like. Imp. Samp.

Table 1.1: A comparison of methods studying populations histories.

Chapter 2

Novel method for analysis of a population split model

In this chapter, we present a novel method for the estimation of the parameters of a split model with two populations. Simulations are then used to assess the performance of the method. Finally, we demonstrate how this method can be used efficiently on larger datasets.

2.1 Importance sampling for a population split model

We present here our adaptation of the IS scheme of Stephens and Donnelly (64) for a split model with two descendant populations and without migration. The different population sizes are held constant across time. We will first state the general idea of our method, present an algorithm to simulate a genealogy and deduce the proposal distribution. Afterwards, we will explain the probability of a genealogy under this model and combine everything together to write down the likelihood function.

2.1.1 The proposal distribution

Our adaptation is based on the observation that until the time of divergence T, the genealogies of the two sampled populations remain independent. Therefore we can build a genealogy in three steps; 1) we start in one of the descendant populations, and we perform events, going backwards in time, until either we reach the time T or we come to a point where no event is possible, 2) we repeat this process in the other descendant population, and 3) with the remaining lineages at time T, we build the rest of the genealogy back in time. We use the proposal distribution of Equation 1.8 to build the genealogy in the ancestral population, and we will use a similar proposal distribution for the descendant populations. The only difference being that it uses event times, and is more restrictive on the possible events. We will focus on the construction of the genealogy before T for one of the descendant populations.

Consider the genealogy of only the sequences of one population before T; contrary to Stephens and Donnelly, we need to simulate the time of each event because we need to know whether they occur before or after T. In the standard coalescent, we know that the times between each event are independent and exponentially distributed with rate $\lambda_k = \binom{k}{2} + k\theta/2$ when there are k lineages. We state that since the histories of the two populations are independent before T, the times of their events are also independent. Therefore, rescaling the time such that one unit is equivalent to N generations, the time between each event in population p, before T, is exponentially distributed with rate $\lambda_k = \binom{k}{2} N/N_p + k\theta/2$, where k is the number of lineages in population p at this time and N_p is its population size. Here follows the algorithm for the tree building in one population, before, and conditional on, the time of divergence T between the two populations:

$k \leftarrow n$

$i \leftarrow 1$

$t \leftarrow 0$

while $t < T$ **do**

if $n^* = 0$ **then**

Break

end if

Simulate a time t_i for the next event from an $\text{Exp}\Big(\binom{k}{2}\frac{N}{N_p} + \frac{k\theta}{2}\Big)$ distribution.

$t \leftarrow t + t_i$

if $t > T$ **then**

Break

else

Choose an event to perform from the uniformly-distributed n^* possible ones.

(update n^* and k according to the chosen event)

$i \leftarrow i + 1$

end if

end while

where n is the number of sequences in the original sample of this sub population and n^* is the number of lineages that can be involved in an event back in time.

From this algorithm, we deduced the proposal density for population p before time T. First define $\mathbf{D}_{i,t}$ to be the set of sequences present (in population p) after the i^{th} event ($\mathbf{D}_{i,t} = (\mathbf{M}_i, \mathbf{n}_i)$) and at time t. The sampled sequences in population p are then $\mathbf{D}_{0,0}$. Note that if $\mathbf{D}_{i,t} = \mathbf{D}_{j,t'}$ and $t \neq t'$, then no event occurred between time t and t' and $i = j$. Define by T_i a random variable from an exponential distribution with parameter $\binom{k}{2}N/N_p + k\theta/2$ to be the time between the $(i-1)^{th}$ and i^{th} event, and let t_i be an outcome of T_i. Denote by f the density function of T_i.

The proposal distribution to build genealogies is then the function:

$$q(\mathbf{D}_{j,t'}|\mathbf{D}_{i,t}) = \begin{cases} f(t_j) \cdot \frac{n_\alpha}{n^*} & \text{if coalescence } \alpha \text{ and } t' < T \ (j = i+1) \\ f(t_j) \cdot \frac{1}{n^*} & \text{if mutation and } t' < T \ (j = i+1) \\ pr(T_{i+1} \geq T - \sum_{k=1}^{i} t_k) & \text{if } t' \geq T \ (j = i) \\ 1 & \text{if } n^* = 0 \ (j = i) \end{cases}$$

where n_α is the number of sequences of type α in $\mathbf{D}_{i,t}$. As we can see, when an event happens, the proposal distribution is the probability of the time of the event multiplied by the probability of the event. In the end there are only two possibilities: 1) the time for the next event ends up being bigger than T, or 2) no is event possible. If the time of the next event makes the total time bigger than the time of divergence, then the proposal distribution is the probability to observe that. Finally, if there is no event possible then the probability of observing those sequences up until the time of divergence will be one (going backward in time).

The possible events when building a genealogy are more restricted than in the Stephens and Donnelly sampler. The condition for a coalescence event is the same; it is only restricted to sequences belonging to the same population. However, the condition for a mutation event is somehow more complicated because we model mutations according to the infinite sites model, meaning that only one mutation event can occur at each site. A sequence can mutate at a position back in time only if: 1) this sequence is the only one of this type in $\mathbf{D}_{i,t}$, and 2) there is no other sequence that shares the same mutation in its own population, or in the other population. Therefore, there are two differences between the proposal distribution of Stephens and Donnelly and our proposal distribution. 1) the need to simulate the time of each event, and 2) the added restrictions on the possible events.

2.1.2 The probabilities of an event and the likelihood

In the last section, our focus was on how to build a genealogy and the proposal distribution $Q(\cdot)$. Now, we will see how we need to define the probability of a genealogy $P(\cdot)$ for labelled and unlabelled sequences, and how to combine everything to obtain the likelihood function.

The probability of a genealogy is based on Equation 1.3 if labelled sequences are used, or on Equation 1.9 if unlabelled sequences are used. For the section of the genealogy before T —going backwards in time— the histories of the subpopulations are independent and we need to consider the times of the events. We present a step-by-step calculation of the probability of a genealogy in a subpopulation before T for labelled sequences. The probability of the next event, going backwards in time is :

$$p(\mathbf{D}_{j,t'}|\mathbf{D}_{i,t}) = \begin{cases} f(t_j) \cdot \frac{n_\alpha(n_\alpha-1)}{n(n-1)} \cdot \frac{(n-1)N/N_p}{(n-1)N/N_p+\theta} & \text{for a coalescence of 2 sequences} \\ & \alpha \ (j=i+1) \\ f(t_j) \cdot \frac{1}{n} \cdot \frac{\theta}{(n-1)N/N_p+\theta} & \text{for a mutation resulting in} \\ & \text{a sequence } \alpha \ (j=i+1) \\ pr(T_{i+1} \geq T - \sum_{k=1}^{i} t_k) & \text{if } t' \geq T \ (j=i) \\ \exp\left\{-\left(T - \sum_{k=1}^{i} t_k\right) \cdot \lambda_k\right\} & \text{if } n^* = 0 \ (j=i) \end{cases}$$

where t' is greater than t, N_p is the effective population size of the subpopulation and $\lambda_k = \left(\binom{k}{2} N/N_j + k\theta/2\right)$. When an event occurs, the probability is the probability of the time multiplied by the probability of the event, as defined previously in the recursion of Equation 1.3. Note that, when it is not possible to perform any more events (i.e.: $n^* = 0$), we know with probability one the set of sequences present at time T. But, when using the true distribution of the genealogy $P(\cdot)$, mutation and coalescence events are always

possible because it is not based on the observed data. Therefore, for this part of the tree, we need the probability that no events occurred during this time interval.

In an importance sampling scheme, the goal is to base the inference of the parameters of interest on the likelihood function. Remember that we can write the likelihood as :

$$\begin{aligned} L\left(\theta, T, \frac{N}{N_1}, \frac{N}{N_2}\right) &= \int P(\mathbf{D}_0|\mathcal{H}) \frac{P(\mathcal{H})}{Q(\mathcal{H})} Q(\mathcal{H}) d\mathcal{H} \\ &\approx \frac{1}{M} \sum_{i=1}^{M} P(\mathbf{D}_0|\mathcal{H}^{(i)}) \frac{P(\mathcal{H}^{(i)})}{Q(\mathcal{H}^{(i)})} \end{aligned} \tag{2.1}$$

where $\mathbf{D}_i = (\mathbf{M}_i, \mathbf{n}_i)$ and $\mathcal{H}^{(i)}$ are sampled from $Q(\cdot)$, the proposal distribution, which implies that $P(\mathbf{D}_0|\mathcal{H}^{(i)})$ is always equal to 1.

In the likelihood used by Stephens and Donnelly, $P(\mathcal{H}) = \prod_{i=0}^{\tau-1} p(\mathbf{D}_{i+1}|\mathbf{D}_i)$ and $Q(\mathcal{H}) = \prod_{i=0}^{\tau-1} q(\mathbf{D}_{i+1}|\mathbf{D}_i)$. In our adaptation, $Q(\mathcal{H})$ is the product of all the $q(\mathbf{D}_{j,t'}|\mathbf{D}_{i,t})$ in the two descendant populations multiplied by the product of $q(\mathbf{D}_{i+1}|\mathbf{D}_i)$ for the ancestral population. The calculation of $P(\mathcal{H})$ is analogous, except when using unlabelled sequences. First, we define the step-by-step probability of a genealogy in one descendant population, before T as:

$$p(\mathbf{D}_{j,t'}|\mathbf{D}_{i,t}) = \begin{cases} f(t_j) \cdot \frac{n_\alpha - 1}{n-1} \cdot \frac{(n-1)N/N_p}{(n-1)N/N_p + \theta} & \text{for a coalescence of 2 sequences } \alpha \\ & (j = i+1) \\ f(t_j) \cdot \frac{n_\alpha + 1}{n} \cdot \frac{\theta}{(n-1)N/N_p + \theta} & \text{for a mutation resulting in} \\ & \text{a sequence } \alpha \ (j = i+1) \\ pr(T_{i+1} \geq T - \sum_{k=1}^{i} t_k) & \text{if } t' \geq T \ (j = i) \\ \exp\left\{ -\left(T - \sum_{k=1}^{i} t_k\right) \cdot \lambda_k \right\} & \text{if } n^* = 0 \ (j = i) \text{ where} \\ & \lambda_k = \left(\binom{k}{2} N/N_j + k\theta/2 \right), \end{cases}$$

again, where t' is greater than t, and N_p is the effective population size of the descendant population. The difference in the calculation of $P(\mathcal{H})$ is that we need to add an additional term to it. Indeed, if the sequences had been labelled, no additional term would have been needed. Hence, the lineages going in to each subpopulation at the time of the split would have been clearly established. In an unlabelled setting, we only know which types of sequences go in to each subpopulation, and there is more than one way to divide those sequences and obtain the same partition. To acquire the additional term, we will first count all the ways we could label the sequences before T in both descendant populations, then unlabelled them after T when going backwards in time. We will then multiply the likelihood by the number of ways we can label the sequences before T and divide it by the number of ways we can label them after T. Denote by n_i the number of lineages in population i just before T, and by n_{ij} the number of sequences of type j in population i just before T. There are $\binom{n_p}{n_{p1},\cdots,n_{pk}}$ different ways to label all the lineages in population p before T, and $\binom{n_1+n_2}{n_{11}+n_{21},\cdots,n_{1k}+n_{2k}}$ ways to label the lineages after T. Therefore, the additional term is:

$$\frac{\binom{n_1}{n_{11},\cdots,n_{1k}}\binom{n_2}{n_{21},\cdots,n_{2k}}}{\binom{n_1+n_2}{n_{11}+n_{21},\cdots,n_{1k}+n_{2k}}}. \tag{2.2}$$

Practically, it is useful to look at how the ratio $\frac{p(H_{j,t'}|H_{i,t})}{q(H_{j,t'}|H_{i,t})}$ can be simplified in the likelihood function. We obtain the following ratio for the estimation of the likelihood for a set of unlabelled sequences :

$$\frac{p(\mathbf{D}_{j,t'}|\mathbf{D}_{i,t})}{q(\mathbf{D}_{j,t'}|\mathbf{D}_{i,t})} = \begin{cases} \frac{n_\alpha - 1}{n_\alpha} \cdot \frac{N/N_p}{(n-1)N/N_p + \theta} \cdot n^* & \text{for a coalescence of 2 sequences} \\ & \alpha \ (j = i+1) \\ \frac{n_\alpha + 1}{n} \cdot \frac{\theta}{(n-1)N/N_p + \theta} \cdot n^* & \text{for a mutation resulting in} \\ & \text{a sequence } \alpha \ (j = i+1) \\ 1 & \text{if } t' \geq T \ (j = i) \\ \exp\left\{ -\left(T - \sum_{k=1}^{i} t_k \right) \cdot \lambda_k \right\} & \text{if } n^* = 0 \ (j = i) \text{ where} \\ & \lambda_k = \left(\binom{k}{2} N/N_p + k\theta/2 \right) \end{cases} \tag{2.3}$$

We now have all the pieces to build genealogies according to the proposal density and to estimate the likelihood function. Remember that, practically, a genealogy is built in three steps and we evaluate the likelihood as we build it. We start in one of the descendant populations, and we perform events, going backwards in time, until either we reach the time T or we come to a point where no event is possible. During this process, we need to verify that we do not remove mutations that are shared with the other subpopulation. We then repeat the process in the other descendant population. Afterwards, we can evaluate the additional term corresponding to Equation 2.2. The last step is to build the genealogy in the ancestral population using the remaining lineages.

2.1.3 Different parametrisation

Unlike Stephen and Donnelly's importance sampler, our proposal distribution contains many unknown parameters. The definitions of those parameters depend on the rescaled time. First, we have the time of divergence of the two populations T, then there is the scaled

mutation rate θ, and finally we indirectly have the effective size of each population. All five parameters depend on how we rescale the time. Because they are all interdependent, only four parameters can be estimated. Usually, the one that is left out is the one used for rescaling the time.

We can rescale the time such that one unit is equivalent to N_a generations (where N_a is the effective size of the ancestral population). Therefore, we will have estimates of the ratio N_a/N_i for the two subpopulations. We choose to use this parametrisation in our analysis; the results are presented in the next section.

Another possible parametrisation would be to rescale the time according to the mutation rate. Hey and Nielsen (27) used this parametrisation for the method IM. The time is rescaled such that 1 unit is equal to μ generations, where μ is the mutation rate per sequence per generation. Under this rescaling of time, we can deduce the coalescence rate from the probability to still observe k lineages after j generations:

$$\begin{aligned} P(T_k > j) &= \left[1 - \binom{k}{2}\frac{1}{N} + O\left(\frac{1}{N^2}\right)\right]^j \\ &= \left[1 - \binom{k}{2}\frac{1}{N}\frac{\mu}{\mu} + O\left(\frac{1}{N^2}\right)\right]^{\frac{j\mu}{\mu}} \\ &= \left[1 - \frac{k(k-1)}{\theta}\mu + O\left(\frac{1}{N^2}\right)\right]^{\frac{t^*}{\mu}} \\ &\to \exp\left\{-\frac{k(k-1)}{\theta}t^*\right\}, \end{aligned}$$

as $\mu \to 0$ and $t^* = j\mu$. In this parametrisation, there is no need to have the ratio of population sizes in the coalescence rates. The difference is that we have three different rescaled mutation rates, one for each population, θ_1, θ_2 (for the two descendant populations), and θ_a for the ancestral population.

We do not see any evidence that one parametrisation is better than the other. With the exception that if either the population size or the mutation rate can be considered as

known, then it might be advantageous to rescale the time using this parameter. We can also fix the mutation rate μ, since we have a relative consensus of its estimated value to a factor of two (61). This will allow us to then change our parametrisation using a general predetermined value of N to rescale the time, and use three different parameters for the coalescent rate (population ratio sizes), one for each population (including the ancestral population). This eliminates the need to estimate θ and will give the same parameter estimates regardless of the value of N used. This way of parametrisation will be used in the next chapter when we will introduce a method to estimate the population ratio sizes per epoch.

In this section, we have detailed the basis of the coalescent theory needed to understand our proposed method. We have also explained in detail our proposal distribution and how we plan to use it to do parameter inference. The next section will present the tests we have done to validate our proposal distribution, and different simulation studies to understand the behaviour of our method.

2.2 Simulations to assess model performance

Anyone developing a new method of inference needs to thoroughly assess the capacity of that method. We present here the results of our analysis. We have implemented the method in a program called *CEPHi* (Coalescent based Estimation of Populations History) using parts of an already-existing program called *pyArg* developed by Didier Amyot and available on the website Lauchpad(2). The program is coded in Python and Go; the data are managed in Python and the trees are built using Go.

For simplicity, we initially make the assumption that all the population sizes are equal, *i.e.* $N_1 = N_2 = N_a$ where N_a is the ancestral population size. This can be seen as an ancestral population split in two, with the descendant populations both experiencing rapid expansions up to their ancestral population size. This assumption means that we do not

need to estimate the ratio of population sizes; we only need to estimate $\theta = 2N_a\mu$ and $T = N_a t$, the rescaled mutation rate and time of divergence, respectively. When building trees, we need to know the value of θ and T in order to simulate the times of each event and to know whether each event is before or after T. Therefore, we estimate the likelihood function for a grid of values of θ and T and then obtain an estimation of the likelihood surface.

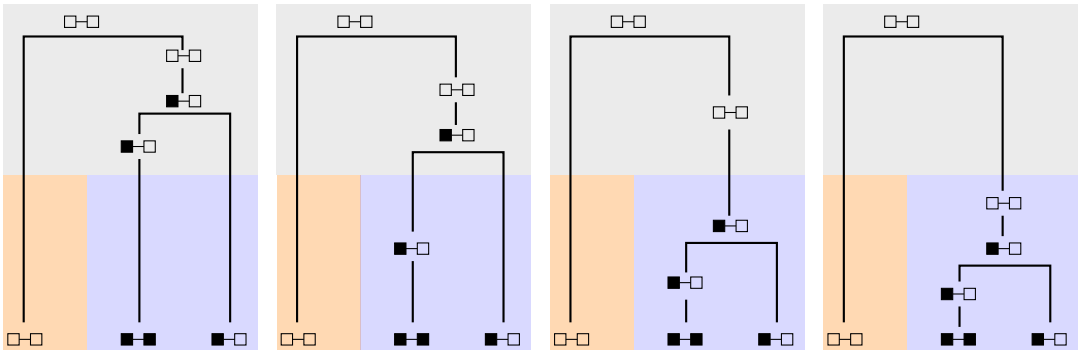
We start by doing a test to validate our proposal distribution and program simultaneously. Afterwards, we present the results of our method for a large simulation study and conclude with a more realistic simulation study.

2.2.1 Validation of the proposal distribution

We will validate our method in two different ways. First, when we have a small sample, we evaluate the likelihood function exactly, which we then compare with our importance sampling-based estimates. Further validation leads us to fix the value of T to zero in our estimation and we then compare the resulting likelihood of θ with the likelihood obtained with the software *Genetree* developed by Griffiths and based on his work with Tavaré (19).

2.2.1.1 Comparison with the real likelihood

The first validation test consists of estimating the likelihood function for a small sample for which it is possible to analytically evaluate the likelihood. The sample is composed of three sequences, each with two segregating sites. The first sequence is composed of two ancestral alleles and is from population 1, and the other two sequences are sampled from the second population, one with two derived alleles, and one with one ancestral allele and one derived allele. Considering the time of the event relative to the time of divergence T, there are four possible genealogies but only one possible sequence of events. Figure 2.1 presents these genealogies, labelled τ_1 to τ_4. The coloured regions represent the genealogy of each subpopulation, and the grey the ancestral population.



(a) τ_1 (b) τ_2 (c) τ_3 (d) τ_4

Figure 2.1: The four genealogies possible for the dataset. An empty square represents an ancestral allele and a black square the derived allele. The two different colours represent the two populations in which the sequences evolve back in time. The grey region corresponds to the shared ancestral population.

The likelihood function is the sum of the four conditional likelihood functions $p(D \mid \tau_i)$, where D is the sample of sequences. Conditional on a genealogy, the likelihood of T and θ is easily calculated. When an event happens after T, the time of this event starting from the previous event can be anywhere between 0 and infinity and is exponentially distributed. Therefore, when we integrated over all the possible times for this event we obtain the value one, leaving only the probability of the event in the likelihood. This is usually the reason why the times are not simulated when building trees (19) (64). However, when the event occurs before T, the time interval does not go to infinity, but is truncated at T. Thus we need to simulate event times, and when we evaluate the real likelihood function we need to integrate over the possible times.

The conditional likelihood functions, including the additional term explained in the previous section (see Equation 2.2), are:

$$\begin{aligned} p(D \mid \tau_1) &= \exp\left\{\frac{-\theta T}{2}\right\} \times \exp\left\{-(1+\theta)T\right\} \times \frac{\binom{1}{1}\binom{2}{1,1}}{\binom{3}{1,1,1}} \times \frac{\theta}{2+\theta} \cdot \frac{2}{3} \times \frac{2}{2+\theta} \cdot \frac{1}{2} \\ &\quad \times \frac{\theta}{1+\theta} \times \frac{1}{1+\theta} \\ &= \frac{2\theta^2}{9(1+\theta)^2(2+\theta)^2} \times \exp\left\{\frac{-(2+3\theta)T}{2}\right\}, \end{aligned}$$

$$\begin{aligned} p(D \mid \tau_2) &= \exp\left\{\frac{-\theta T}{2}\right\} \times \int_0^T \frac{\theta}{(1+\theta)} \cdot (1+\theta)\exp\left\{-(1+\theta)a\right\} \\ &\quad \times \exp\left\{-(1+\theta)(T-a)\right\}\,\mathrm{d}a \times \frac{\binom{1}{1}\binom{2}{2}}{\binom{3}{1,2}} \times \frac{2}{(2+\theta)} \cdot \frac{1}{2} \times \frac{\theta}{(1+\theta)} \cdot \frac{1}{(1+\theta)} \\ &= \exp\left\{\frac{-\theta T}{2}\right\} \times \frac{\theta^2}{3(1+\theta)^3(2+\theta)} \times \int_0^T (1+\theta)\exp\left\{-(1+\theta)T\right\}\,\mathrm{d}a \\ &= \frac{\theta^2 \cdot T}{3(1+\theta)^2(2+\theta)} \times \exp\left\{-\left(\frac{2+3\theta}{2}\right)T\right\}, \end{aligned}$$

$$\begin{aligned} p(D \mid \tau_3) &= \exp\left\{\frac{-\theta T}{2}\right\} \times \int_0^T \int_0^{T-a} \frac{\theta}{(1+\theta)}(1+\theta)\exp\left\{-(1+\theta)a\right\} \\ &\quad \times \frac{1}{(1+\theta)}(1+\theta)\exp\left\{-(1+\theta)b\right\} \times \exp\left\{-\frac{\theta}{2}(T-b)\right\}\,\mathrm{d}b\,\mathrm{d}a \\ &\quad \times \frac{\binom{1}{1}\binom{1}{1}}{\binom{2}{1,1}} \times \frac{\theta}{(1+\theta)} \times \frac{1}{(1+\theta)} \\ &= \left[\frac{2+\theta}{\theta(1+\theta)}\exp\left\{-\left(\frac{2+3\theta}{2}\right)\right\} - \frac{2}{\theta}\exp\left\{-(1+\theta)T\right\} + \frac{1}{(1+\theta)}\exp\left\{-\frac{\theta}{2}T\right\}\right] \\ &\quad \times \frac{\theta^2}{(1+\theta)^2(2+\theta)} \times \exp\left\{\frac{-\theta T}{2}\right\}, \end{aligned}$$

$$
\begin{aligned}
p(D \mid \tau_4) &= \exp\left\{\frac{-\theta T}{2}\right\} \times \int_0^T \int_0^{T-a} \int_0^{T-b} \frac{\theta}{(1+\theta)}(1+\theta)\exp\left\{-(1+\theta)a\right\} \\
&\quad \times \frac{1}{(1+\theta)}(1+\theta)\exp\left\{-(1+\theta)b\right\} \times \frac{\theta}{2}\exp\left\{-\frac{\theta}{2}c\right\} \times \exp\left\{-\frac{\theta}{2}(T-c)\right\} \mathrm{d}c\,\mathrm{d}b\,\mathrm{d}a \\
&\quad \times \frac{\binom{1}{1}\binom{1}{1}}{\binom{2}{2}} \\
&= \left[\frac{T}{\theta(1+\theta)} - \frac{1}{(1+\theta)^2} + \left(\frac{1}{(1+\theta)^2} - \frac{T^2}{2}\right)\exp\left\{-(1+\theta)T\right\}\right] \\
&\quad \times \frac{\theta^2}{2(1+\theta)^2} \times \exp\left\{\frac{-\theta T}{2}\right\}.
\end{aligned}
$$

We will not explain in detail all these conditional probabilities, but only the third one, the probability of the sequences conditional on τ_3, to give an idea of how the calculation is made. Looking at Figure 2.1(c), we can see that no event happened in population 1 and two events happened in population 2 before T. Since there is only one sequence in population 1, the total rate for the next event is simply the mutation rate $\theta/2$. For this portion of the genealogy, the likelihood is simply the probability that nothing happens during the time interval $[0, T]$, giving a $\exp\left\{\frac{-\theta T}{2}\right\}$ conditioned probability. In population 2, there are two events before T: first a mutation and then a coalescence event. We need to consider all the possibilities for the times of those events. For the mutation event, the time can be anywhere between 0 and T, and for the coalescence event, the time between the mutation and the coalescence events can be between 0 and T minus the time of the mutation event. Therefore, we have to do a double integral to cover all the times possible for those two events. In both cases, there are two lineages present before the events and consequently both times will follow an exponential distribution of rate $(1 + \theta)$. The probability of the mutation event is $\theta/(1 + \theta)$, and since the type of the resulting sequence is the same as the other one, the probability of observing this particular mutation event –knowing that the event is a mutation– will be one. A similar argument gives the probability of the coalescence event. Follows the term to account for unlabelled sequences. And finally after T, we have a mutation event followed by a coalescence event.

Figure 2.2(a) shows the analytic likelihood surface for this sample. The maximum is situated at $\theta = 0.45$ and $T = 2.56$. For the estimation of the likelihood function we have evaluated the likelihood at each point in a grid of 25 values of $\theta \in [0, 4]$, and 50 values of $T \in [0, 10]$. At each point, we have built a certain number of genealogies to obtain the estimation of the likelihood. We obtained the surface in Figure 2.2(b) using 1,500 trees. Figures 2.2(c) and 2.2(d) show how using fewer trees, specifically 100, can increase the variability in the estimates. Finally, 2.2(e) and 2.2(f) were estimated using 500 trees.

The maximum likelihood estimates (MLE) obtained (constrained by the grid) using the estimation of Figure 2.2(b) are $\theta = 0.48$ and $T = 2.4$. We need to take into consideration that our maximum will lie on a point of the grid; thus, if the real maximum is not on the grid, we will not find it. The likelihood surface is similar to the truth.

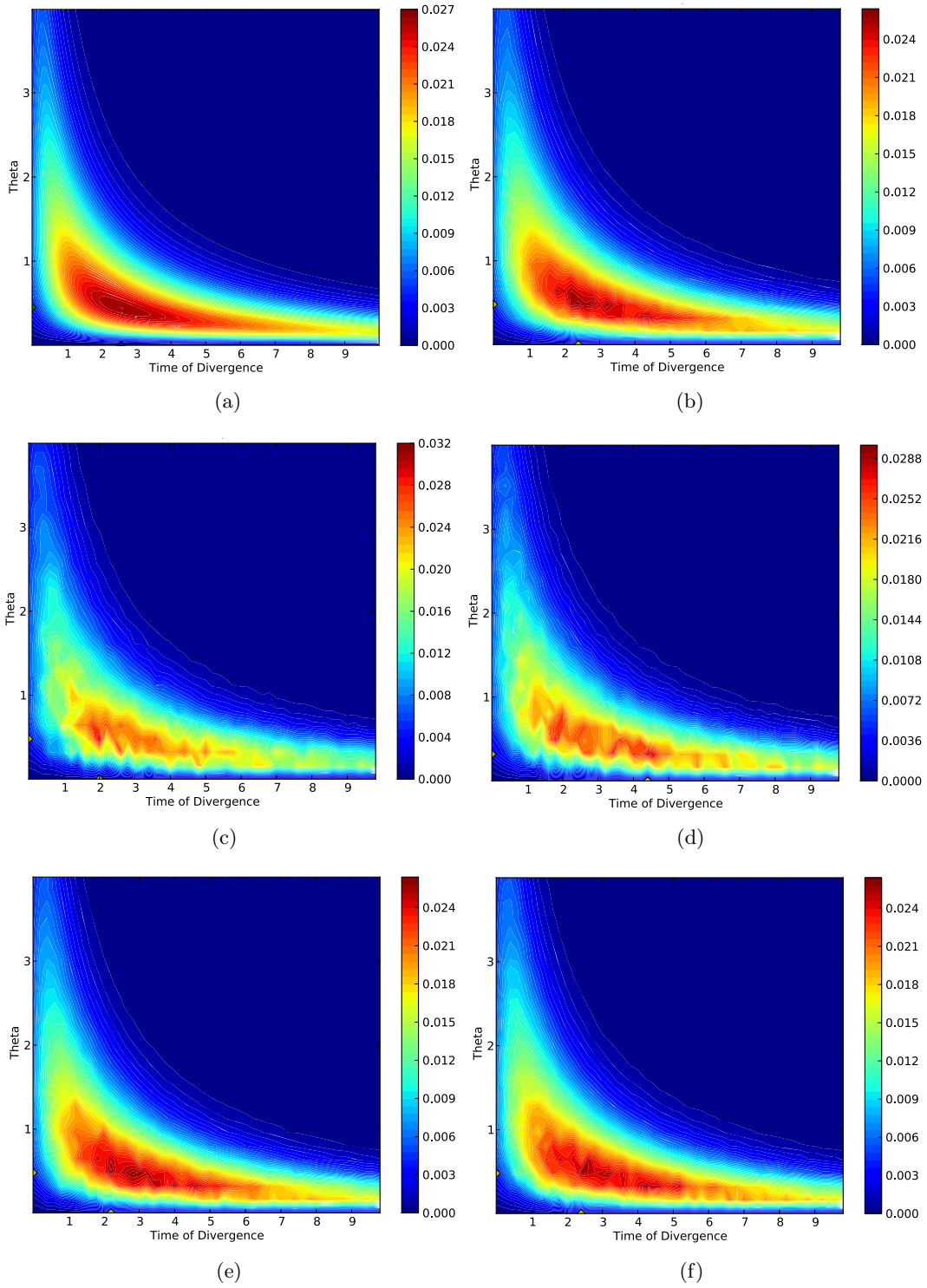


Figure 2.2: (a) the analytic likelihood function for the data, (b) an importance sample-based estimation of the likelihood function based on 1,500 trees; (c) and (d): as (b) but based on 100 trees, (e) and (f) as (c),(d) but based on 500 trees.

2.2.1.2 Comparison with *Genetree*

The first results of our validation test were encouraging. To examine a larger dataset, we used *Genetree*, a software that estimates the likelihood of θ (and others parameters depending of the model assumed) using the proposal distribution of Griffiths and Tavaré (19) in a model with no split. Using another proposal distribution should provide us with an independent estimation of the likelihood at the value $T = 0$. Stephens and Donnelly (64) showed that their proposal distribution improves upon the one used by *Genetree*.

In our sampler, if we fix T to a small value such that the probability that an event happens before T is nearly zero, then the estimate of the likelihood function of θ should converge to the *Genetree* estimates (after normalisation by a correction term corresponding to sampling two, rather than one, groups).

We simulated a dataset with the software *ms* developed by Hudson (34) which allows the simulation of datasets according to different models based on the coalescent process. The dataset is composed of twenty sequences of eight SNPs, with ten sequences per population. The scaled time of divergence was set to 0.1 and θ to 2. If all the sequences are remaining at time T (as $T \to 0$), the additional term, for unlabelled sequences, is:

$$\frac{\binom{10}{2,2,1,4,1}\binom{10}{3,1,3,1,2}}{\binom{20}{5,3,4,4,1,1,2}} \approx 0.000649505 \tag{2.4}$$

Genetree does not evaluate the likelihood for a list of values of θ, it uses a driving value to build the trees and then evaluates the likelihood for a range of values. The likelihood function is then smoother. Figure 2.3 shows the estimation of the likelihood function based on 100,000 trees using *Genetree* in blue, with the driving value of θ set to two.

We have evaluated the likelihood with our method for 50 values of $\theta \in [0, 6]$ using 100,000 trees. To be able to compare the two estimates of the likelihood function, we have divided our estimate by the additional term of Equation 2.4 and plotted both estimates on the same axis. The result is presented in Figure 2.3 by the red line.

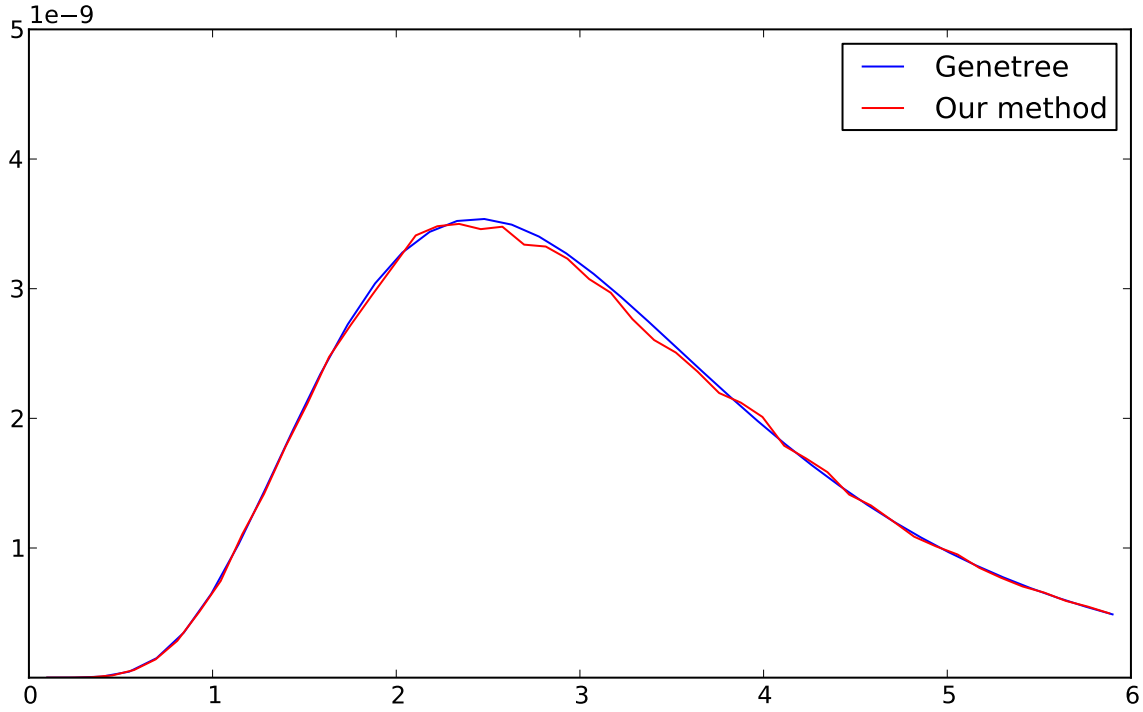


Figure 2.3: The figure shows the estimate of the likelihood of θ using *Genetree* (blue) and our importance sampler (red). Both estimates are based on 100,000 trees.

The results demonstrate the likely validity of our program and proposed distribution. Our estimate is not as smooth as *Genetree*, since we are evaluating the likelihood for a range of different values of θ. Therefore, it shows the variation we could have in our estimates. Repeated *Genetree* runs –not shown here– give a likelihood estimate with a slightly lower maximum.

2.2.2 A simulation study

To assess the strengths and weaknesses of our method, we decided to do a simulation study. First, let us mention that we wish to use our method on whole-genome data, but our method makes the assumption that no recombination occurred. The idea is to use multiple regions of the genome with a very low recombination rate, such that we can assume no recombination occurred. The regions are typically well separated; hence, we

can consider them independent. Therefore, the likelihood function is simply the product of the likelihood of every region.

Our first simulation study is far from realistic in size, but we wanted to try a large number of different parameter combinations; hence, to reduce the time of calculation we needed to use smaller sample size and sequence length. We first explain the design of the simulation study, then we present the results and finally we will present results from some extra simulations. This simulation study was performed, in part, using the resources of the Oxford Supercomputing Centre.

2.2.3 Design of the simulation study

All of the datasets were simulated using *ms* (34). The model used is a split model with no migration or recombination: an ancestral population of size N has diverged at time T into two descendant populations of equal size N. The time is rescaled such that one unit of time is equivalent to N generations. The unknown parameters are the scaled mutation rate θ and the time of divergence T.

We have first chosen eleven parameter combinations for which we varied three parameters: the mutation rate θ, the time of divergence T and the total sample size n (where $n/2$ is the number of sequences sampled in each subpopulation), where $\theta \in \{2, 3, 5\}$, $T \in \{0.25, 1, 1.5, 2.5\}$ and $n \in \{10, 20, 50, 100\}$. In Table 2.1 each line represents a combination of the parameters.

We have simulated fifty datasets for each combination of parameters. For every dataset we have evaluated the likelihood of θ and T on a 20×20 grid (twenty values of θ and twenty values of T). Then for each point we have built one hundred thousand trees to estimate the likelihood at this position. To assure ourselves that we have built enough trees, we have repeated the estimation of the likelihood function for all the datasets.

Theta	T	n
2	1.5	50
3	1.5	50
5	1.5	50
3	0.25	50
3	1.0	50
3	1.5	50
3	2.5	50
3	1.5	10
3	1.5	20
3	1.5	50
3	1.5	100

Table 2.1: The different parameter combinations used in the simulation study. The columns represent the mutation rate θ, the time of divergence T and the total sample size n (where $n/2$ is the number of sequences sampled in each subpopulation). We highlighted in yellow the parameter that is variable in the possible combinations.

2.2.3.1 Results : Using one dataset at a time

We first present the results we obtained without combining the different datasets. For each dataset we have plotted the log likelihood surface estimate and drawn the 95% confidence interval based on the likelihood ratio. Table 2.2 shows the percentage of coverage of those confidence intervals, *i.e.* the percentage of datasets for which the confidence interval contains the real values. We observe that the percentage is similar for the two repetitions. Only the datasets with $\theta = 5$ have a clearly lower than expected coverage. A possible explanation for this is that we did not build enough trees to obtain a good estimation of the likelihood surface.

We have plotted the maximum likelihood estimates of the first repetition for every combination of parameters to visualise their distributions. We point out that the maximum likelihood estimates (MLE) will be situated on points of the grid of estimated values. Hence, to establish that we see multiple estimates at the same position, we have used sunflower plots: multiple points at the same position are represented by petals. Figures 2.4, 2.5 and 2.6 represent the plots of the MLE for the combinations of parameters with θ, T and n varying, respectively. The real values of θ and T are represented by dotted lines.

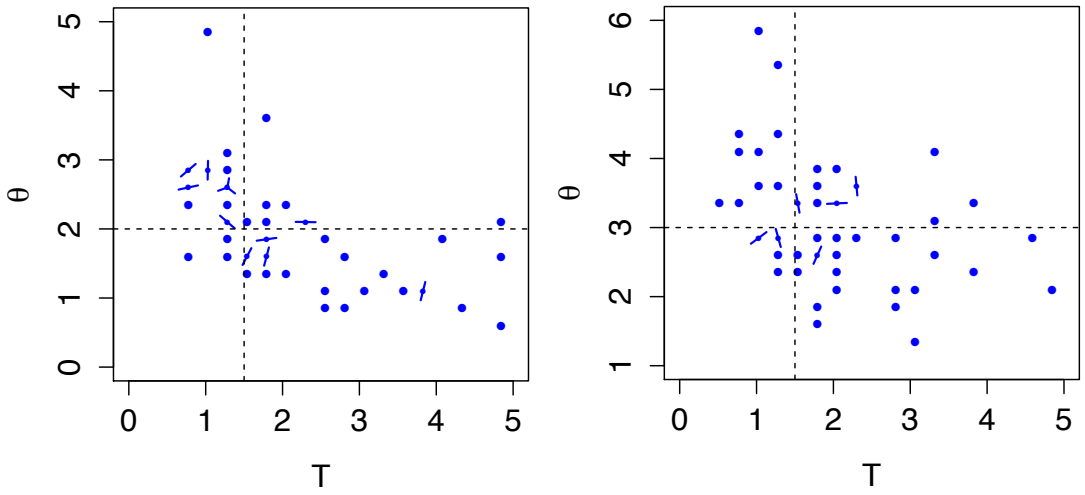
Theta	T	n	% of coverage (2.99 log 95% CI)	
			rep. 1	rep. 2
2	1.5	50	94%	94%
3	1.5	50	90%	90%
5	1.5	50	78%	78%
3	0.25	50	94%	92%
3	1	50	94%	92%
3	1.5	50	94%	92%
3	2.5	50	98%	98%
3	1.5	10	98%	98%
3	1.5	20	94%	94%
3	1.5	50	94%	94%
3	1.5	100	94%	94%

Table 2.2: Percentage of coverage of the 95% confidence intervals.

Looking more closely at Figure 2.4, we observe that the maximum likelihood estimates of T do not seem to be symmetrically distributed around the real value, with a longer tail for greater values of T. The MLE of θ seems, however, to be distributed symmetrically around their real values. This observation still holds when we contemplate Figures 2.5 and 2.6. We can also detect a negative correlation between the MLE of T and θ, *i.e.* a tendency to overestimate one parameter when we underestimate the other one.

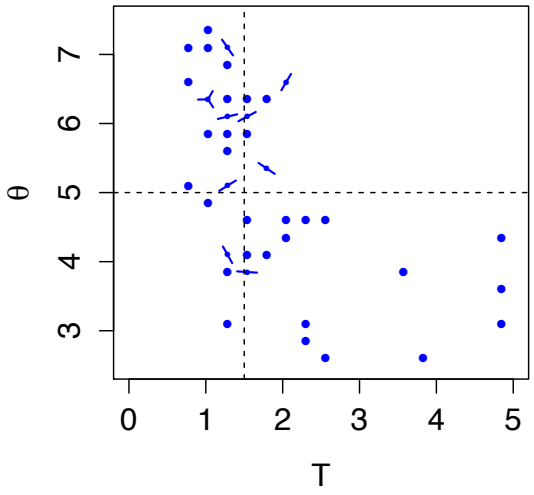
Since these two parameters are both scaled using the population size, we can expect a correlation between them, but there is more to it. When we underestimate (overestimate) the time of divergence T, we will usually overestimate (underestimate) θ and vice versa. Our model for mutation is the infinite-sites model, which means that each time a mutation occurs, it will occur at a new position in the sequences: there are no repeated mutations. Furthermore, θ is, by definition, the scaled mutation rate by sequence in a population, so all the mutation events in one population will occur independently of the mutation events in the other population (before T). Hence, the expected number of mutation events happening before T in both populations will be twice that expected in one population.

If we are evaluating the likelihood function for a value of T smaller than the real one, we are allowing fewer mutations to occur during the time when there were two populations.



(a) $\theta = 2$

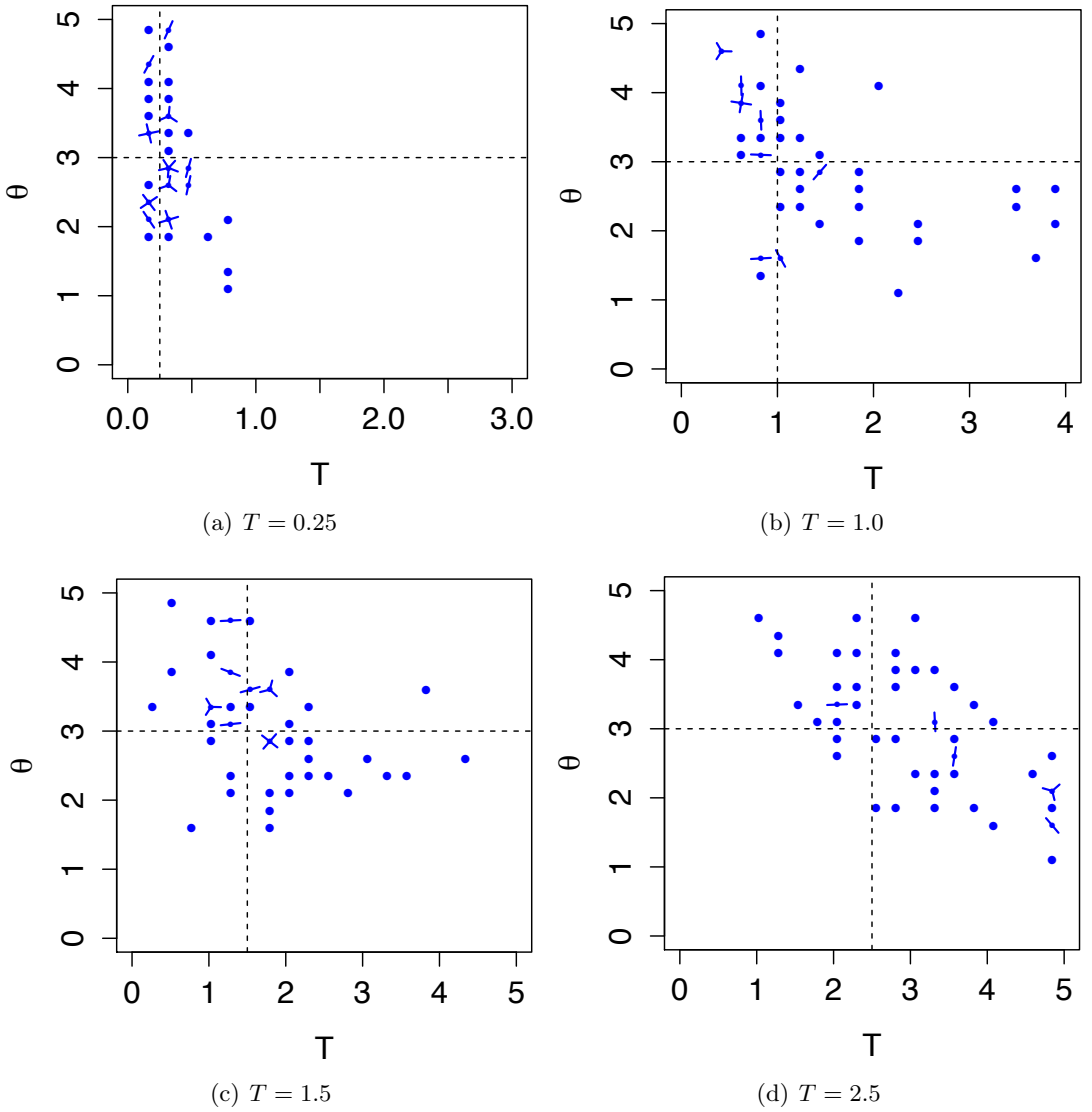
(b) $\theta = 3$



(c) $\theta = 5$

Figure 2.4: Sunflower plots of the maximum likelihood estimates for the combinations of parameters with θ varying. Multiple points at the same position are represented by the number of petals. The intersection of the dotted lines represents the real values of θ and T.

Hence, more mutation events will be required to occur in the ancestral population to account for the mutations in the sample and the likeliest value of θ will then be larger than its true value. This is why we think that T and θ are negatively correlated.

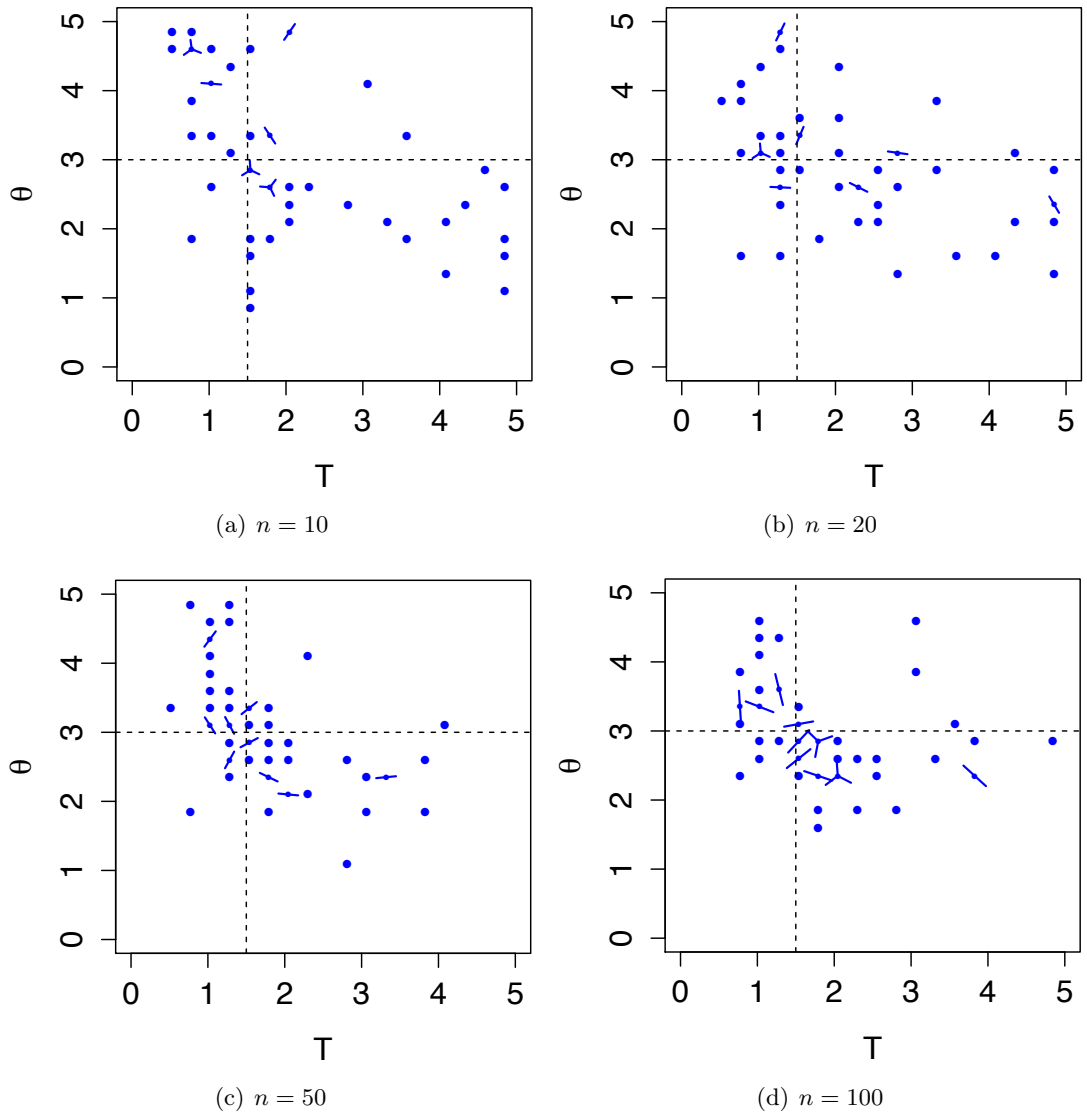


(a) $T = 0.25$ (b) $T = 1.0$

(c) $T = 1.5$ (d) $T = 2.5$

Figure 2.5: Sunflower plots of the maximum likelihood estimates for the combinations of parameters with T varying. Multiple points at the same position are represented by the number of petals. The intersection of the dotted lines represents the real values of θ and T.

By simply looking at those plots, it is difficult to see whether the mean of the MLE is near the real value, and to qualify its variance. In Table 2.3, we have listed the mean and variance of the MLE of θ. We remark that the means of the MLE of θ are usually close to the real values, except in the datasets with $\theta = 5$. The variances are similar, but larger for the datasets with $\theta = 5$ and the ones with a sample size of ten.



(a) $n = 10$ (b) $n = 20$

(c) $n = 50$ (d) $n = 100$

Figure 2.6: Sunflower plots of the maximum likelihood estimates for the combinations of parameters with n varying. Multiple points at the same position are represented by the number of petals. The intersection of the dotted lines represents the real values of θ and T.

Table 2.4 presents the mean and variance of the MLE of T, for which we observe a consistent upward bias. This is consistent with the asymmetric distribution of the MLE. The variance is smaller for the datasets with smaller sample size, and the ones with $\theta = 2$.

Theta	T	n	Mean theta MLE		Variance theta MLE	
			rep. 1	rep. 2	rep. 1	rep. 2
2	1.5	50	1.9800	1.9650	0.6307	0.5796
3	1.5	50	3.0700	2.9600	0.7822	0.6790
5	1.5	50	5.1750	5.1700	1.8425	1.4873
3	0.25	50	2.9800	2.9400	0.8424	0.7826
3	1	50	2.9950	2.9550	0.9518	0.8905
3	1.5	50	3.1100	3.1050	0.6504	0.6645
3	2.5	50	2.9500	2.9900	0.8214	0.7530
3	1.5	10	3.0100	3.0100	1.3463	1.3259
3	1.5	20	2.9150	2.8700	0.7750	0.6348
3	1.5	50	3.0000	3.0000	0.7219	0.6148
3	1.5	100	2.9400	2.9600	0.4994	0.4698

Table 2.3: Means and variances of the maximum likelihood estimates of θ.

Theta	T	n	Mean T MLE		Variance T MLE	
			rep. 1	rep. 2	rep. 1	rep. 2
2	1.5	50	2.0715	2.0562	1.3569	1.3587
3	1.5	50	2.0053	2.0613	0.9685	1.0295
5	1.5	50	1.7813	1.7457	0.9966	0.8372
3	0.25	50	0.3128	0.3221	0.0263	0.0307
3	1	50	1.3761	1.3802	0.8713	0.8812
3	1.5	50	1.7559	1.7406	0.6927	0.6583
3	2.5	50	3.1149	3.0182	1.1712	1.1624
3	1.5	10	2.0816	2.0664	1.7211	1.5911
3	1.5	20	2.2140	2.1783	1.6777	1.5498
3	1.5	50	1.7762	1.7202	0.7387	0.6848
3	1.5	100	1.8831	1.8322	0.9088	0.8496

Table 2.4: Means and variances of the maximum likelihood estimates of T.

We also remark that the results are analogous for the two repetitions. There is no large distinction between the means and variances of the MLEs. Comparing the likelihood surfaces and MLEs of the two repetitions for each dataset (results not shown), no large differences were observed, which suggests that we have built enough trees to obtain good estimates of the likelihood surfaces.

The time to build one 100,000 trees varies for the different combinations of parameters. As a reference point, the first repetition was done on a 2.8GHz Intel Core i7 processor and we have seen that an increase in the number of SNPs (equivalent to an increase in the value of θ from 2 to 5), increase the average time from 38 to 84.7 seconds. An increase in the sample size from 10 to 100, increases the time from 26 to 65.8 seconds. For this simulation study, we have built a total of 2.2×10^{10} trees.

2.2.3.2 Results : Using multiple sections of the genome

We have decided to combine estimation of the likelihood function of different datasets in the same combination of parameters to obtain a new, hopefully better, estimate. To be able to study how combining different sections of the genome changes our estimation without simulating more datasets and having to estimate their likelihoods, we have decided to resample the datasets we have already created. For each combination of the parameters, we have resampled fifty times three, five, and ten datasets and we have combined their likelihoods to get a new estimate. Tables 2.5, 2.6 and 2.7 show the percentages of coverage of the confidence intervals based on three, five, and ten datasets, respectively.

As expected, using more dataset improves the coverage. Again, the worst combination of parameters is the one with $\theta = 5$, perhaps resulting from a poor estimation of the likelihood caused by not using enough trees for the number of SNPs in the datasets. Using more datasets also reduces the sizes of the confidence intervals, as shown in Figure 2.7. Thus, each subfigure represents the fifty 95% confidence intervals obtained when using three (2.7(a)), five (2.7(b)) and ten (2.7(c)) datasets. The confidence intervals are represented by a light blue region and a region with darker blue means that this region is composed of many overlapping confidence intervals. An analogous pattern is observed for the others combinations of parameters.

We can also combine all 50 datasets for each combination of parameters. In doing this, we again reduce more the surface of the confidence interval and still obtain good estimates of the likelihood. Figure 2.7(d) represents the log likelihood estimate when the fifty datasets are used. Figure 2.8 represents the divergence of the maximum likelihood estimates from the real values. The blue asterisks represent the positions of the real values of the parameter and the tip of the arrows the position of the maximum likelihood estimates. The longer arrow, starting at $\theta = 3$ and $T = 1.5$, is in fact made of two superimposed arrows. One represents the datasets for which the sample size was only ten, and the other the combination where T was varied. We could expect to have worse estimates when we have

Theta	T	n	% of coverage (2.99 log 95% CI)	
			rep. 1	rep. 2
2	1.5	50	98%	100%
3	1.5	50	96%	94%
5	1.5	50	70%	68%
3	0.25	50	96%	94%
3	1	50	92%	98%
3	1.5	50	80%	96%
3	2.5	50	98%	100%
3	1.5	10	96%	94%
3	1.5	20	96%	92%
3	1.5	50	94%	96%
3	1.5	100	92%	96%

Table 2.5: Percentages of coverage of the 95% confidence intervals when using three random datasets to estimate the likelihood surface.

Theta	T	n	% of coverage (2.99 log 95% CI)	
			rep. 1	rep. 2
2	1.5	50	96%	96%
3	1.5	50	92%	92%
5	1.5	50	70%	52%
3	0.25	50	94%	92%
3	1	50	96%	96%
3	1.5	50	86%	94%
3	2.5	50	98%	94%
3	1.5	10	98%	96%
3	1.5	20	98%	96%
3	1.5	50	88%	100%
3	1.5	100	98%	100%

Table 2.6: Percentages of coverage of the 95% confidence intervals when using five random datasets to estimate the likelihood surface.

fewer sequences, but there is no straight-forward explanation for the other arrow. We can note that the second repetition gave us a closer maximum likelihood estimate; in fact, it is identical to the smaller arrow starting at the same asterisk. The other long arrow on the plot represents the datasets with $\theta = 5$, which we already know gave poor estimates. There is no apparent directional bias in our estimates, *i.e.* the arrows do not all point in the same direction.

Theta	T	n	% of coverage (2.99 log 95% CI)	
			rep. 1	rep. 2
2	1.5	50	98%	98%
3	1.5	50	94%	96%
5	1.5	50	62%	58%
3	0.25	50	88%	88%
3	1	50	96%	98%
3	1.5	50	90%	100%
3	2.5	50	98%	96%
3	1.5	10	96%	100%
3	1.5	20	98%	98%
3	1.5	50	100%	100%
3	1.5	100	100%	100%

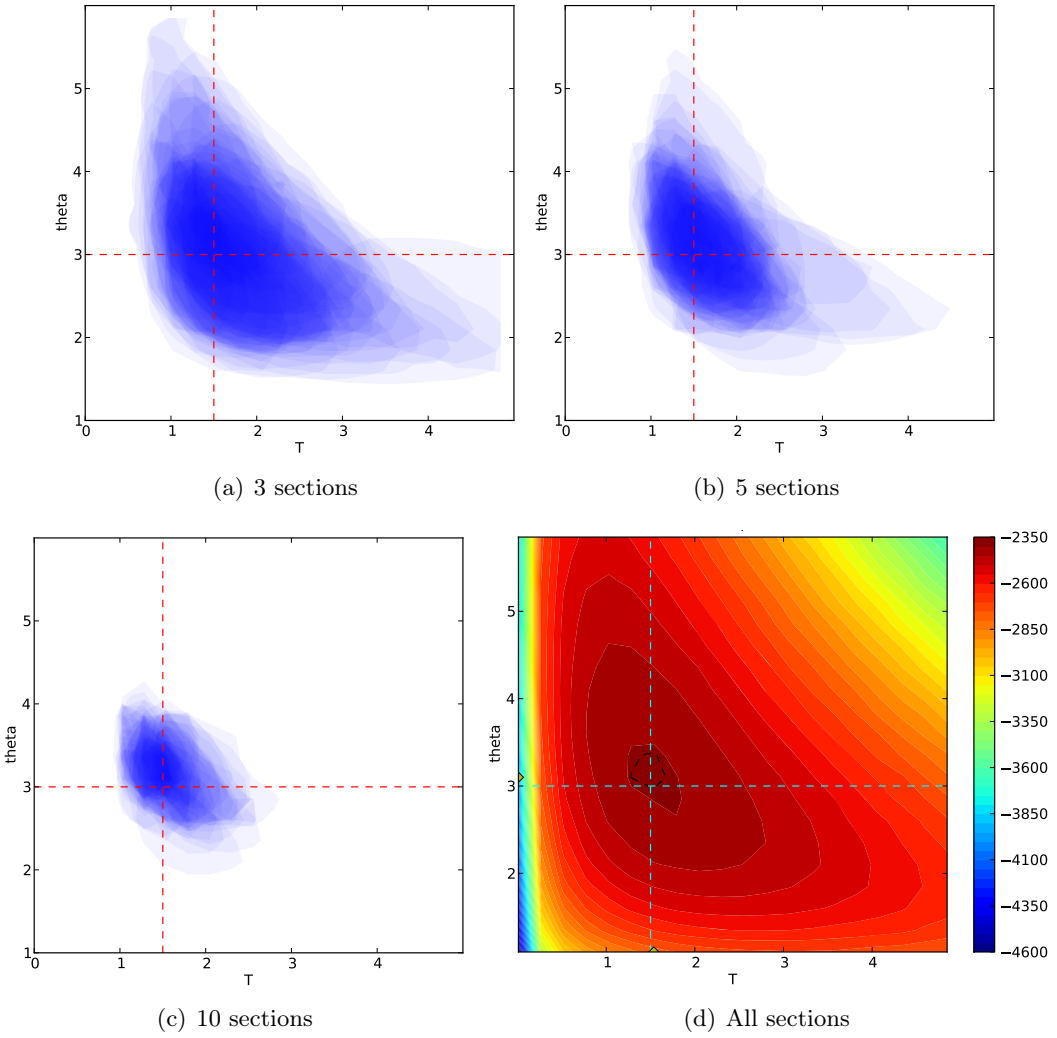
Table 2.7: Percentages of coverage of the 95% confidence intervals when using ten random datasets to estimate the likelihood surface.

2.2.3.3 Extra simulations

Following this simulation study, we thought about doing a few more simulations. First, we wanted to see how the likelihood surface would have looked if we had used more points on our grid. Then, we built more trees for the datasets with $\theta = 5$ to see whether we could get better results. And finally we thought that our range of values of T was missing values between 0.25 and 1.0; therefore, we simulated datasets for $T = 0.5$ and $T = 0.75$ using $\theta = 3$ and $n = 50$ again.

We have used a grid composed of 400 points, with 20 values of both θ and T. To see whether adding more points to the grid changes the likelihood surface, we have picked five datasets with $\theta = 3$ and reevaluated the likelihood surface again using 100,000 trees but on a grid of 2,500 points, using 50 values of both θ and T. Figure 2.9 shows the comparison for two datasets. Figures 2.9(a) and 2.9(c) are the likelihood surfaces when 400 points are used, and Figures 2.9(b) and 2.9(d) when 2,500 are used. We can see that the likelihood is not as smooth when we are using more points, but the shape of the surface is very similar.

We have also built more trees for the dataset with $\theta = 5$. When using 250,000 trees we obtain a percentage of coverage of 82% (compared to the 95% expected), and with



(a) 3 sections (b) 5 sections

(c) 10 sections (d) All sections

Figure 2.7: Representation of the fifty 95% confidence intervals when using three (a), five (b) and ten (c) datasets. Each confidence interval is represented by a light blue region, with darker blue regions being those in a larger number of confidence intervals. The real values are represented by the red dotted lines. Plot (d) represents the log of the likelihood when using all 50 datasets. The 95% confidence interval is represented by the black dotted line.

450,000 trees we get 84% (compared to the 95% expected). The MLEs of θ and T are the same even if we are using more trees. For the extra combinations of parameters we have done, the results are similar to what we observed in the simulation study. We obtain a 92% coverage for $T = 0.5$ and 90% coverage for $T = 0.75$. When using all the sections, we obtain a MLE of $\theta = 3.1$ and of $T = 0.4735$ for $T = 0.5$, and a MLE of $\theta = 3.1$ and of

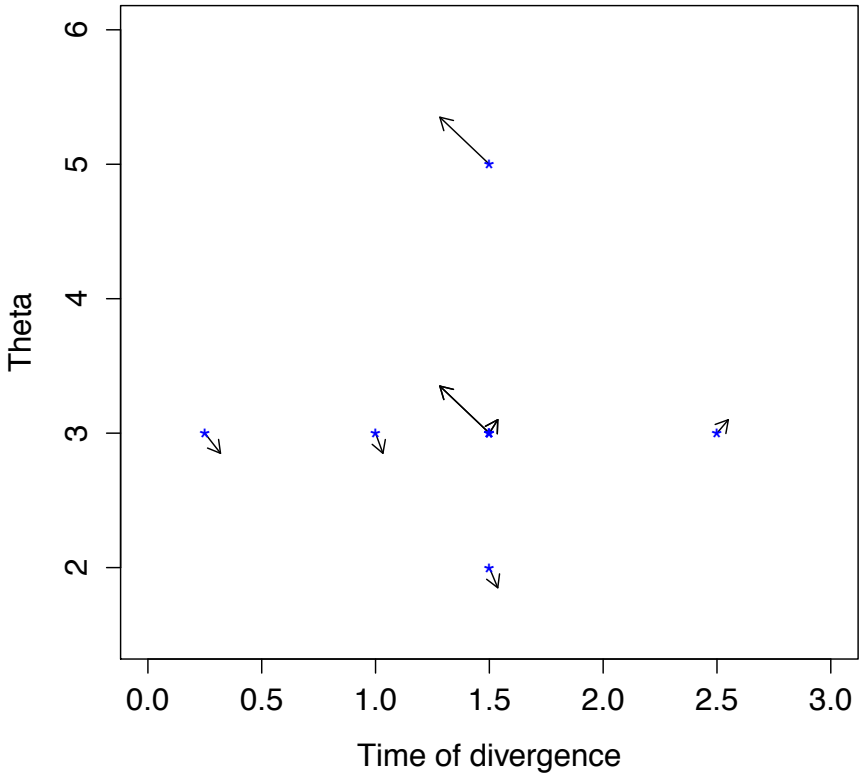


Figure 2.8: Representation of the divergences of the maximum likelihood estimates using all 50 sections for all combinations of parameters of the first repetition. The blue stars represent the real values, and the tip of the arrows the maximum likelihood estimates.

$T = 0.7825$ for $T = 0.75$.

Overall, those results are very encouraging: when using multiple sections we can reduce the confidence intervals a lot whilst still including the real values. Unfortunately, if we were analysing a larger dataset, this method is still very time-consuming. We are hoping to be able to still get good estimates when using fewer trees but more regions. The next section will present a simulation study of this type.

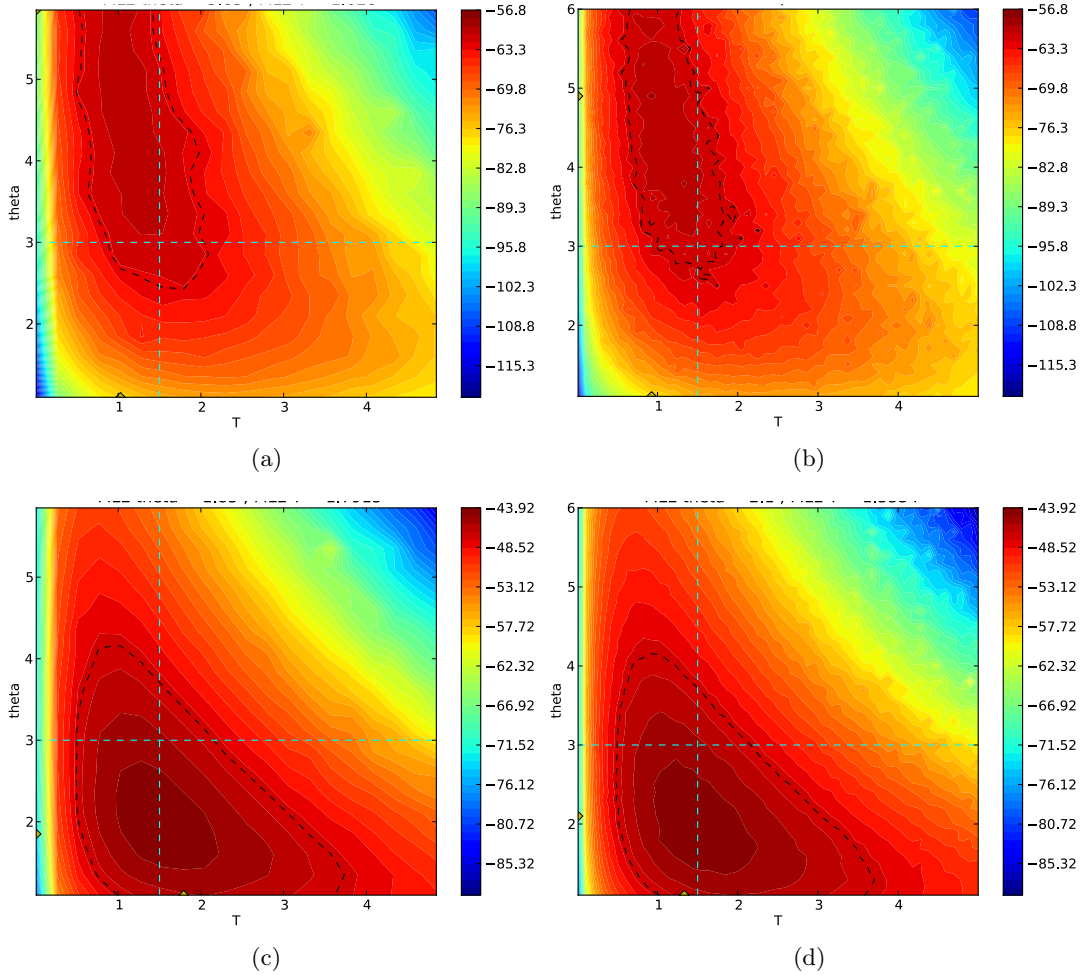


Figure 2.9: The effect of adding more points to the grid. Figures (a) and (c) represent the likelihood surfaces using a grid with 400 points and Figures (b) and (d) using a grid with 2,500 points.

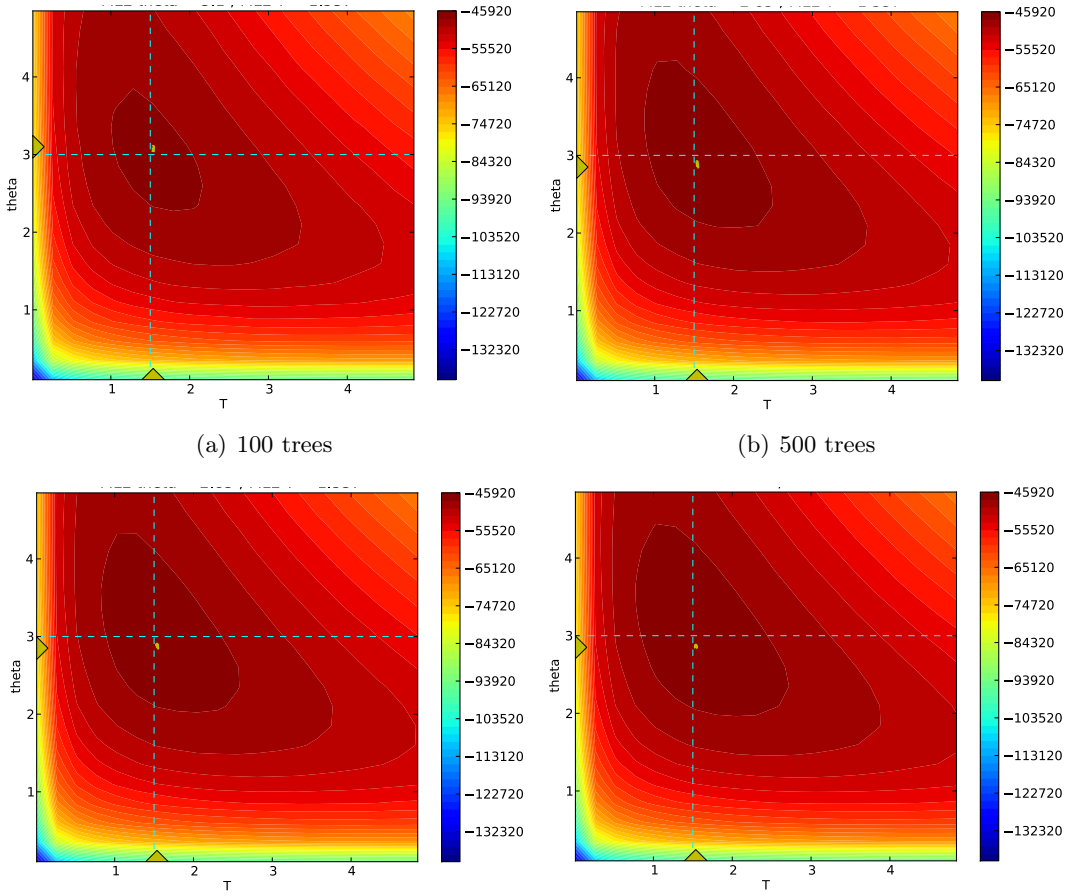
2.2.4 Closer to reality: a new simulation study

The main drawback of the full likelihood importance sampling method is the time required for the simulation of the trees. When using a large sample covering a small genome region, large numbers of trees are necessary to obtain good parameter estimates. We saw in the last section that we can obtain precise estimates when using multiple genome regions. The number of regions we used was small compared to reality. On the human genome, we would expect to be able to use thousands of regions. In this case, we could ask ourselves if we still need hundreds of thousands of trees. We have conceived a simulation study to try to answer this question. This simulations study was performed, in part, using the resources of the Oxford Supercomputing Centre.

We have simulated 1,000 datasets with *ms*, using the parameters $\theta = 3$ and $T = 1.5$, sampling 25 sequences in each sub-population. We are supposing, as before, that the population sizes are equal. Using a grid of 400 points, we have built 100,000 trees per points. The likelihood surface was then estimated using the first 100, 500, 1,000, 5,000, 10,000, 50,000, and finally all of the trees in the datasets. Figures 2.10 and 2.11 show the resulting likelihood estimates.

From these figures, we observe that the confidence intervals are very small (a black dot), and do not include the real values. The MLEs are indicated on their respective axes by a yellow triangle. The confidence intervals are situated at the intersection of the MLEs. They are all the same: $\theta = 2.85$ and $T = 1.537$ except for the estimates based on 100 trees for which the MLE of θ is equal to 3.1. Those two values of θ are the two grid points closest to the truth. The next closest grid point to $T = 1.5$ is 1.2825. This probably explains why the real values are not included in the confidence intervals. To be certain, we should have verified that the real values were in the grid points or used a more robust confidence interval approach.

We have also estimated the likelihood by gradually increasing the number of regions used. We observe that only 50 regions seem to be needed to obtain good estimates of



(a) 100 trees

(b) 500 trees

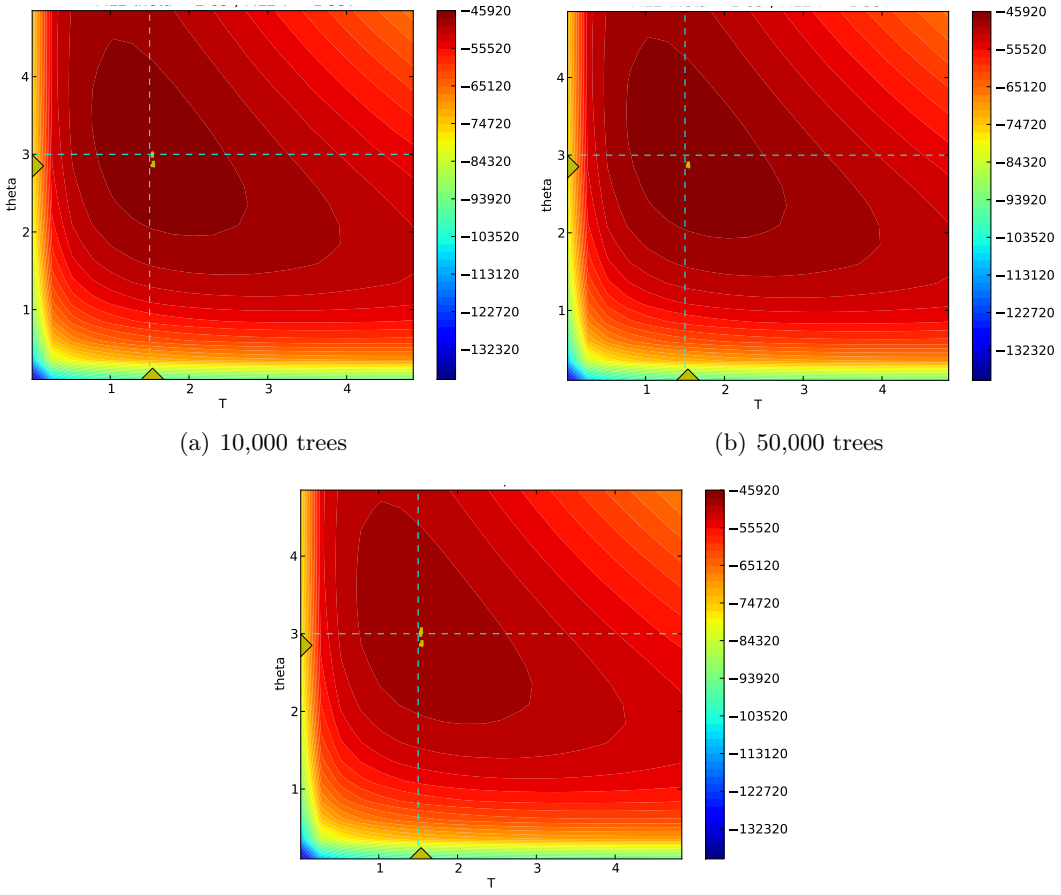
(c) 1,000 trees

(d) 5,000 trees

Figure 2.10: Estimates of the likelihood surface using 1,000 regions and the first: (a) 100 trees, (b) 500 trees, (c) 1,000 trees and (d) 5,000 trees. The real values are represented by the dotted lines, and the MLEs by the yellow triangles on the axes.

the parameters, even if we are using only 100 trees. But, when using fewer regions, we need more trees, around 10,000 for 25 regions, to obtain good estimates and confidences intervals that cover the real values.

These results shows that it is feasible to apply this method to real whole-genome data. In this section we have presented the results of two simulation studies; and established that our method gives promising results. The next section will present an optimisation algorithm to estimate more efficiently the likelihood surface.



(a) 10,000 trees

(b) 50,000 trees

(c) 100,000 trees

Figure 2.11: Estimates of the likelihood surface using 1,000 regions and the first: (a) 10,000 trees, (b) 50,000 trees, and (c) 100,000 trees. The real values are represented by the dotted lines, and the MLEs by the yellow triangles on the axes.

2.3 An optimisation algorithm for likelihood estimation

The use of a grid of points to evaluate the likelihood is inefficient, and even more so when using data larger than the one analysed previously. We need to reduce the number of points to evaluate, while continuing to find the maximum of the likelihood function. Thus, we have developed an optimisation algorithm based on hill climbing and the use of driving values that allows us to find a local optimum and greatly reduces the number of points of the grid to evaluate.

For most of our parameters we can use a driving value to build the trees and then we can estimate the likelihood for a range of surrounding possible values. This method give good estimates of the likelihood around the driving values. The only parameters for which we cannot use a driving value is T, the time of divergence because it could affect the history of the tree.

The algorithm is as follow:

1. The user needs to give a range of possible values for T and θ and, if desired, N_a/Ni (ex: $T : [0.0; 1.5]$ and $\theta : [15; 45]$) and a $jump$ is defined (ex: $jump = 0.25$).

2. A very fine grid of possible points is then determined using the ranges of possible values. (ex: all the points in the range for $T : [0.0; 1.5]$ with an increment of 0.0005 and all the points in the range for $\theta : [15; 45]$ with an increment of 0.02

3. The initial step of the algorithm consists of finding the three values of each parameter that are situated at the first, at the half and at the third quarter of the range of possible values (ex: 0.375, 0.75 and 1.125 for T and 22.5, 30 and 37.5 for θ).

4. Trees are then built for all of these points. The point (T_{max}, θ_{max}) that gives the maximum likelihood is kept in memory.

5. Then we repeat the following steps until the MLE is found:

   (a) We build trees at values $T_{max} \pm jump$ and θ_{max}

(b) The likelihood will be evaluated using those trees, for a range of values of θ surrounding θ_{max} (ex: 10 values below and above θ_{max} with a distance of 0.2 between each point). This is possible since θ does not appears directly in the probability of an event in the proposal distribution, but only in the probability of an event using the coalescence. Therefore, as we build the trees, using θ_{max} we can evaluate the likelihood (the importance weights) for different value of θ.

(c) The likelihood will be re-evaluated at T_{max} if the new maximum has a different value for θ to the previous maximum. This may occur because the evaluation of the likelihood is influenced by the value used to construct the trees. Among other things because θ is used to simulate the time of the events.

(d) If the maximum is situated at the same position for two consecutive steps, the size of $jump$ is decreased (ex: for T by 0.05 until $jump = 0.05$, then by 0.025 until $jump = 0.025$ and by 0.005 until $jump = 0.0025$).

6. The MLE is considered to be found when the maximum has remained the same for two consecutive steps and $jump$ is at its minimum (ex: $jump = 0.0025$).

Figure 2.12 presents the result obtained with the optimisation algorithm using the previously introduced dataset composed 1,000 regions. Multiple independent runs, not presented here, gave similar results. This optimisation algorithm allows us to obtain an good estimate of the likelihood surface for a finer grid of points near the local maximum.

In this chapter, we have presented a novel method for the analysis of a population split model, using an adaptation of the Stephens and Donnelly importance sampler. We have demonstrated its accuracy and its scalability to a larger dataset. The model used and assumptions are simple and might be unrealistic, but using a simpler model to start allows us to establish strong bases to our method to build on. In the next chapter, we present an extension to the method that allows the estimation of variable population sizes.

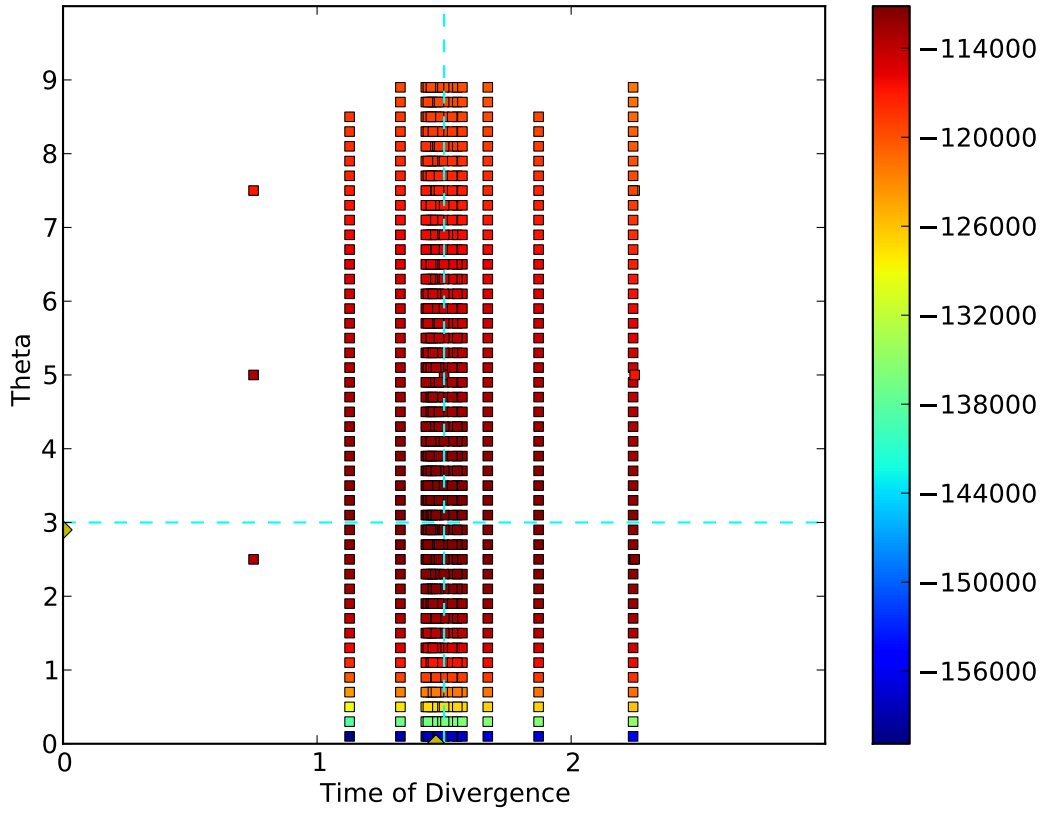


Figure 2.12: One result from the optimisation algorithm. The MLE of T is 1.47 and the MLE of θ is 2.9. The likelihood was evaluated assuming labelled sequences.

Chapter 3

Extension of the model to variable population sizes

One of the many interesting aspects of population history is the effective size of a population and how it changes over time. Estimating past changes in population size can give insight into how the population has been affected by major events, for example mass migrations, glacial periods and the advent of agriculture. It also provides valuable information for other studies that usually assume a constant population size.

We aim to extend our previous method to estimate the size of the populations in the model by allowing them to vary freely at fixed time intervals. Estimating the population sizes assuming they are constant within epochs will give a better picture of the changes compared to if we had modelled those changes, for example, as an exponential growth or decrease.

Our idea is to estimate the population sizes using trees built with our sampler. First, assuming that the time of divergence T and the scaled mutation rate θ have been correctly estimated, we explain how we can use built trees to obtain estimates of the population sizes. Then, we present simulation studies that demonstrate how the estimates are biased. We propose two different ways to correct for this bias and we present their performance.

Finally, we show how we can jointly estimate the time of divergence and the variable population sizes.

3.1 Estimation of the population ratio sizes

For the next two sections we will assume that the time of divergence T and the mutation rate μ are known. First, assuming that the trees for each region are known, we present the maximum likelihood estimates of the population sizes. Then, we explain the algorithm we use to obtain the population size estimates from trees built with our importance sampler. Here, time is rescaled according to $2N$ (we are using diploid individuals), a fixed arbitrary value that does not refer to the ancestral population size. In fact, for each epoch, in each population, we will estimate the population ratio size $q_{ij} = N/N_{ij}$, which is equivalent to the coalescence rate in epoch i of population j. From these estimates we are able to obtain estimates of the population sizes using N. By assuming μ is known and fixing N, the scaled mutation rate $\theta = 4N\mu$ is totally determined.

3.1.1 Maximum likelihood estimates

To simplify, we first suppose that the real tree of a region is known entirely (time of coalescence and time of mutation) and we present the maximum likelihood estimates of the population sizes. We first look at the likelihood function for an epoch in one of the descendant populations (population j).

We can estimate the population size for an epoch based only on the events occurring in that epoch, since all the events in the tree are independent. The population ratio size in epoch i is $q_{ij} = N/N_{ij}$, to simplify the notation we will drop the subscript ij of q. We suppose that the limits of the epoch, denoted by τ_1 and τ_2, are smaller than T. The likelihood is then the product of the probabilities of the events that occurred during the epoch. We denote by $\mathcal{E}$ the total number of events, by $\mathcal{C}$ the total number of coalescences

and by $\mathcal{M}$ the total number of mutations ($\mathcal{E} = \mathcal{C} + \mathcal{M}$). The likelihood for epoch i and population j is then:

$$\begin{aligned} L(q|\theta) = \prod_{\epsilon=1}^{\mathcal{E}} & \left[\left(\binom{n_\epsilon}{2} q + \frac{n_\epsilon \theta}{2} \right) \exp \left\{ - \left(\binom{n_\epsilon}{2} q + \frac{n_\epsilon \theta}{2} \right) t_\epsilon \right\} \right] \\ & \times \exp \left\{ - \left(\binom{n_{\mathcal{E}+1}}{2} q + \frac{n_{\mathcal{E}+1}\theta}{2} \right) \left(\tau_2 - \left(\tau_1 + \sum_{\epsilon=1}^{\mathcal{E}} t_\epsilon \right) \right) \right\} \\ & \times \prod_{c=1}^{\mathcal{C}} \frac{n_c(n_c-1)q}{n_c(n_c-1)q + n_c\theta} \times \prod_{m=1}^{\mathcal{M}} \frac{n_m \theta}{n_m(n_m-1)q + n_m\theta} \end{aligned} \tag{3.1}$$

where n_ϵ is the number of lineages remaining before the ϵ^{th} event and t_ϵ represents the time between events starting the time from τ_1. Therefore, $n_{\mathcal{E}+1}$ is the number of lineages remaining at the end of the epoch. The exponential on the second line represents the probability that no event occurs between the last event of the epoch and τ_2. The two last products are the probability that the event is either a coalescence or a mutation, and where n_c and n_m are the number of lineages remaining right before the c^{th} coalescence and the m^{th} mutation events.

The derivative of the log-likelihood function in regard to q is then:

$$\begin{aligned} \frac{\partial l}{\partial q} &= \sum_{\epsilon=1}^{\mathcal{E}} \left[\frac{n_\epsilon(n_\epsilon-1)}{n_\epsilon(n_\epsilon-1)q + n_\epsilon\theta} - \binom{n_\epsilon}{2} t_\epsilon \right] - \binom{n_{\mathcal{E}+1}}{2} \left(\tau_2 - \left(\tau_1 + \sum_{\epsilon=1}^{\mathcal{E}} t_\epsilon \right) \right) \\ &\quad + \sum_{c=1}^{\mathcal{C}} \left[\frac{1}{q} - \frac{n_c(n_c-1)}{n_c(n_c-1)q + n_c\theta} \right] - \sum_{m=1}^{\mathcal{M}} \left[\frac{n_m(n_m-1)}{n_m(n_m-1)q + n_m\theta} \right] \\ &= \sum_{\epsilon=1}^{\mathcal{E}} \left[-\binom{n_\epsilon}{2} t_\epsilon \right] - \binom{n_{\mathcal{E}+1}}{2} \left(\tau_2 - \left(\tau_1 + \sum_{\epsilon=1}^{\mathcal{E}} t_\epsilon \right) \right) + \sum_{c=1}^{\mathcal{C}} \frac{1}{q} \\ &= \sum_{c=1}^{\mathcal{C}} \left[\frac{1}{q} - \binom{n_c}{2} t_c \right] - \binom{n_{\mathcal{C}+1}}{2} \left(\tau_2 - \left(\tau_1 + \sum_{c=1}^{\mathcal{C}} t_c \right) \right) \end{aligned} \tag{3.2}$$

The second equality is obtained by simplifying the first term with the two last terms. We can then reorganise the summation in function of the coalescence events only, since the number of lineages remains the same after a mutation event. The maximum likelihood estimate for q is then simply:

$$\hat{q} = \frac{\mathcal{C}}{\sum_{c=1}^{\mathcal{C}} \binom{n_c}{2} t_c + \binom{n_{\mathcal{C}+1}}{2} \left(\tau_2 - \left(\tau_1 + \sum_{c=1}^{\mathcal{C}} t_c \right) \right)}.$$

Therefore, we only need the times between coalescence events (t_c) and the number of coalescences in an epoch ($\mathcal{C}$) to estimate the population ratio size (q). We use datasets composed of multiple independent regions of the genome, the likelihood is then simply the product of the likelihood of each region. And for each epoch and each population, we can also evaluate the likelihood independently.

However, when we built trees with our importance sampler, we only obtain a point estimate of the coalescence tree, but it is important to account for the variability in our estimates of the tree. To account for this, we could, for example, use multiple dependent trees per region. But there could be an issue with those estimates, since the trees are built assuming constant and equal population sizes. Therefore, the rate of coalescence events in the trees will be closer to the rate under the constant size population model rather than to the real rate.

There is still information in the coalescence events: for example, if, compared to N, the population size has increased drastically, q_j will be small and fewer coalescence events will have occurred in the true trees. This implies a larger number of mutation events on long lineages resulting in an excess of singletons. This makes it difficult to coalesce sequences in our sampler, perhaps resulting in fewer coalescence events than expected under a constant size model. We present in the next section a Monte Carlo EM algorithm to solve these issues and obtain good estimates of the population ratio sizes.

3.1.2 Implementing the extension: an MCEM algorithm

The EM algorithm is an iterative procedure for finding the maximum likelihood estimate of the parameters of interest when the underlying model depends on unobservable variables. It was presented in a paper by Dempster, Laird and Rubin in 1977 (11). The idea behind the algorithm was known but it is in this paper that the general algorithm was presented and where it got its name: EM for expectation–maximisation. This algorithm is particularly useful when the MLE is straightforward to compute in presence of the complete data. In our case, the unobservable variables are the underlying trees. We first describe the EM algorithm, then we use the trees built with our importance sampler to define a Monte Carlo EM algorithm to estimate the population ratio sizes. A review of the algorithm and its many extensions can be found in a book by McLachlan and Krishnan (46).

3.1.2.1 Expectation–Maximisation algorithm

First, we will define some notation : suppose that X represents the observed data, Z the missing data and θ the parameters of interest. The EM algorithm is composed of two steps. The first step, or the E step, is the establishment of the expectation, taken over Z conditioned on X, of the log likelihood of the complete data using the current guess at θ. This expectation is usually referred as the Q function, defined at the t^{th} iteration as:

$$Q(\theta, \theta^{(t-1)}) = \mathbb{E}_{Z|\theta^{(t-1)},X}\left[l(Z|X,\theta)\right]. \tag{3.3}$$

The second step, the M step, consists of finding the value of θ that maximises Equation 3.3. A new estimate of θ is proposed for each iteration of those two steps. The algorithm stops when the difference between successive estimates of θ, $\theta^{(i)}$ and $\theta^{(i+1)}$, is small enough, *i.e.* until convergence. Starting values for θ are needed and might influence the results if the likelihood surface is multimodal, and therefore it might be necessary to run the algorithm multiple times with different starting values.

3. EXTENSION OF THE MODEL TO VARIABLE POPULATION SIZES

Many different adaptations of the EM algorithm have been proposed over the years. In 1990, Wei and Tanner (73) published a Monte Carlo implementation of the EM algorithm. The idea simply consists of using a Monte Carlo method of integration to estimate the Q function (Equation 3.3) of the E step.

We have seen in the previous section that it is straightforward to obtain the MLE when we have the trees. Unfortunately, we do not observe those, but we can consider them as the missing data in the context of an EM algorithm. If we suppose that the time of divergence is known and θ is fixed (by fixing μ), the likelihood of a region D can be written as:

$$L(\lambda; D) = \int_H P(D, H|\lambda)\mathrm{d}H = \int_H P(D|H)P(H|\lambda)\mathrm{d}H, \tag{3.4}$$

where H represents a possible history (or tree) and λ the population sizes. The probability $P(D|H)$ will be equal to 1 if the tree H could have generated the data D or else it will equal to 0. And the probability $P(H|\lambda)$ is equivalent to the product of Equation 3.1 taken across all populations and epochs. Seeing the trees we build (H) as the missing data, we can write the Q function of an EM algorithm as:

$$Q(\lambda, \lambda^{(t-1)}) = \mathbb{E}_{H|\lambda^{(t-1)},D}\left[\log\{P(D|H)P(H|\lambda)\}\right]. \tag{3.5}$$

To evaluate this expectation we would need to look at all the possible trees that could have generated the observed data D. We have seen in the first chapter that even with a clever recursion the number of possible trees to evaluate gets rapidly too large. But we can approximate the likelihood function using importance sampling. Therefore, the expectation of Equation 3.5 can be approximated using any importance sampling method by:

$$\tilde{Q}(\lambda, \lambda^{(t-1)}) = \frac{1}{\sum_j w_j}\sum_{i=1}^{M} w_i \log\{P(D|H_i)P(H_i|\lambda)\}, \tag{3.6}$$

where the w_i are the importance weights of the M trees built using the parameters $\lambda^{(t-1)}$

and the proposal distribution. The importance weight of a tree is the ratio of the probability of this tree under the coalescent process over the probability to have built this tree with the proposal distribution. In Chapter 2, we have described our adaptation of the proposal distribution of Stephens and Donnelly to a split model where our importance weights were the product of the individual weights of all the events in a tree as described in Equation 2.3.

In this new context of using fixed epoch of time, the importance weight for an event in population p and in an epoch defined by the interval $[\tau_1, \tau_2]$ is:

$$\frac{p(\mathbf{D}_{j,t'}|\mathbf{D}_{i,t})}{q(\mathbf{D}_{j,t'}|\mathbf{D}_{i,t})} = \begin{cases} \frac{n_\alpha - 1}{n\theta} \cdot \frac{q_p}{(n-1)q_p+\theta} \cdot n^* & \text{for a coalescence of 2 sequences } \alpha \ (j = i+1) \\ \frac{\theta}{n((n-1)q_p+\theta)} \cdot n^* & \text{for a mutation resulting in a sequence } \alpha \ (j = i+1) \\ 1 & \text{if } t' \geq \tau_2 \ (j = i) \\ \exp\left\{ -\left(\tau_2 - \left(\tau_1 + \sum_{k=1}^{i} t_k \right) \right) \cdot \lambda_k \right\} & \text{if } n^* = 0 \ (j = i) \end{cases} \tag{3.7}$$

where $\mathbf{D}_{i,t}$ is the set of sequences present (in population p) after the i^{th} event of the epoch and at time t, $\lambda_k = \left(\binom{k}{2} q_p + k\theta/2 \right)$ and where t' is greater than t. And where n is the number of lineages in population p at that time, n^* the number of lineages that could be involve in the next event and n_α the number of lineages of type α. The weight of a tree w is then the product over all the events of the tree, across all epochs and population of their individual weights as defined by 3.7.

For multiple regions, the likelihood function is the product of the likelihood (Equation 3.4) of each region, since we considered them independent. Then $\tilde{Q}$ becomes:

$$\tilde{Q}(\lambda, \lambda^{(t-1)}) = \sum_{j=1}^{L} \frac{1}{\sum_k w_{jk}} \sum_{i=1}^{M} w_{ji} \log\{P(D_j|H_{ji})P(H_{ji}|\lambda)\}, \tag{3.8}$$

where D_j represents the data of the j^{th} region. With our proposal distribution we are building trees that always agree with the data, therefore $P(D_j|H_{ji})$ is always equal to one. The probability of a tree $P(H_{ji}|\lambda)$ is in fact equal to the likelihood of a tree as defined in the previous section by Equation 3.1. The estimate of the population ratio size of population p in epoch τ that maximise Equation 3.8 is equal to :

$$\hat{q}_{p\tau} = \frac{\sum_{j=1}^{L} \frac{1}{\sum_k w_{jk}} \sum_{i=1}^{M} w_{ji} \mathcal{C}_{ji}}{\sum_{j=1}^{L} \frac{1}{\sum_k w_{jk}} \sum_{i=1}^{M} w_{ji} \left[\sum_{c=1}^{\mathcal{C}_{ji}} \binom{n_c}{2} t_c + \binom{n_{\mathcal{C}_{ji}+1}}{2} \left(\tau_2 - \left(\tau_1 + \sum_{c=1}^{\mathcal{C}_{ji}} t_c \right) \right) \right]}. \tag{3.9}$$

Therefore, in practise to estimate the population sizes per epoch we use this MCEM algorithm that iterates between 1) building a number of trees per region with our importance sampler using the current estimates of the population sizes and 2) from those trees we obtain the new set of parameters λ^t using Equation 3.9. We iterate until the estimates of the population sizes per epoch of two consecutive iterations are close enough. As a starting value, we set all the parameters to 1, since they represent the population ratio sizes per epoch. This is equivalent to assuming that all the population sizes are equal to N. Any valid proposal distribution could in fact be used to build the trees, but in practise a good proposal distribution is useful since then fewer trees need to be built to obtain good estimates.

To obtain confidence intervals, we need to estimate the variance of our estimates of the population ratio sizes. We use the observed information matrix, which is an estimate of

the inverse of the variance. The information matrix is defined as:

$$I(\lambda, D) = -\frac{\partial^2 l(D;\lambda)}{\partial^2 \lambda}, \tag{3.10}$$

where l is the log of the likelihood. The estimate of the variance is the inverse of Equation 3.10 evaluated at $\hat{\lambda}_{MLE}$.

Different estimates of the observed information matrix have been proposed over the years. We will use the one proposed by Oakes in 1999 (54) (see also (35)), that uses the second derivative of the Q function. Oakes shows that the observed information matrix can be derived from:

$$-\frac{\partial^2 l(D;\lambda)}{\partial^2 \lambda} = -\frac{\partial^2 Q(\lambda, \lambda^{(t-1)})}{\partial^2 \lambda} + \frac{\partial^2 Q(\lambda, \lambda^{(t-1)})}{\partial \lambda^{(t-1)} \partial \lambda}. \tag{3.11}$$

An estimate of this function is obtained by evaluating it at $\lambda = \lambda^{(t)}$ the final estimate of the EM algorithm. Our estimate of the Fisher information matrix is obtained by replacing the Q function of Equation 3.11 by $\tilde{Q}$ (Equation 3.8). In our case, the last term of Equation 3.11 is equal to 0, since $\lambda^{(t-1)}$ does not appear in $\tilde{Q}$ and is only used when building the trees. Therefore, we can estimate the observed information matrix by differentiating twice with respect to λ our estimate of the $\tilde{Q}$ function.

Only the events happening in a particular epoch influence the estimation of the population size for this epoch. Nevertheless, the number of lineages remaining in an epoch depends on the events that happened in the previous epoch. Therefore, our confidence intervals and estimates of the population sizes are dependent.

3.1.2.2 Building trees using time intervals

This new extension has some implications in the way we build the trees. At any moment while building a tree, we need to make sure that the coalescence rate is accurate, *i.e.* that we are using the correct population size. We use a similar technique to the one used when

we get close to the time of divergence. We also need to explain how we find the limits of the time intervals used.

In our method, we use the population sizes on two different occasions: 1) when we simulate the time of the next event, and 2) when we evaluate the likelihood function. When building trees using estimates of the population sizes per epoch, we need to keep track of the epoch we are in to make sure we are simulating the time and estimating the likelihood using the correct population sizes. We also need to be careful when the time for the next event is outside the current epoch. Supposing we are in population i and epoch j when this happens, and denoting the ratio N/N_{ij} by q_{ij}, we then: 1) do not perform the event, 2) evaluate the probability that no event happens during the remainder of epoch j and add it to the likelihood, and 3) starting from the beginning of epoch $j+1$, we resample a time for the next event using $q_{i(j+1)}$. This is accounted for in the weights as seen in the third case of Equation 3.7.

When looking backwards in time to understand evolution it makes sense to use a logarithmic timeline. The further we go back in time, the fewer coalescence events that happen. Therefore, we want to have epochs that get larger as we go backwards in time. To define the epochs used during the estimation, we have adopted the method used by Li and Durbin (38). They used the formula:

$$t_i = 0.1 \exp\left\{\frac{i}{n}\log(1 + 10T_{max})\right\} - 0.1 \tag{3.12}$$

to define the limits of their intervals, where $i = 0, \ldots, n$ and T_{max} is the maximum value possible for the time of the MRCA in the rescaled time (where 1 unit is equivalent to $2N$ generations). On autosomal and simulated data Li and Durbin have used the value $n = 64$ and $T_{max} = 15$. They have regrouped some intervals following the pattern: $1 * 4 + 25 * 2 + 1 * 4 + 1 * 6$, which means that the first interval is formed of the four first intervals given by 3.12, then the following 25 intervals are formed each by regrouping two

intervals, *etc.* In doing this, they reduced the parameter space. Regrouping the first four intervals guaranteed enough coalescent events in this interval; because PSMC uses only one individual at the time, fewer coalescences will happen in recent time.

We have decided to use the same formula (Equation 3.12). However, our method will perform more coalescence events in the most recent past, since we are using multiple individuals. This might enable our method to obtain good estimates of the population sizes for recent years. Therefore, we allow more epochs close to the present, though we choose similar epochs to PSMC further back in time. To do this we have decided to use $n = 120$ and to not regroup time intervals at the beginning, and then we regroup the intervals following PSMC.

3.2 Simulations to assess performance

We present here the results of our simulation study to assess the performance of our method in recovering the changes in population sizes. The first result uncovers a problem with our importance sampler in the context of variable population sizes. We then explain the origin of this bias in our estimates and we propose two corrections that improve the results obtained with our importance sampler.

3.2.1 Design and results of the simulation study

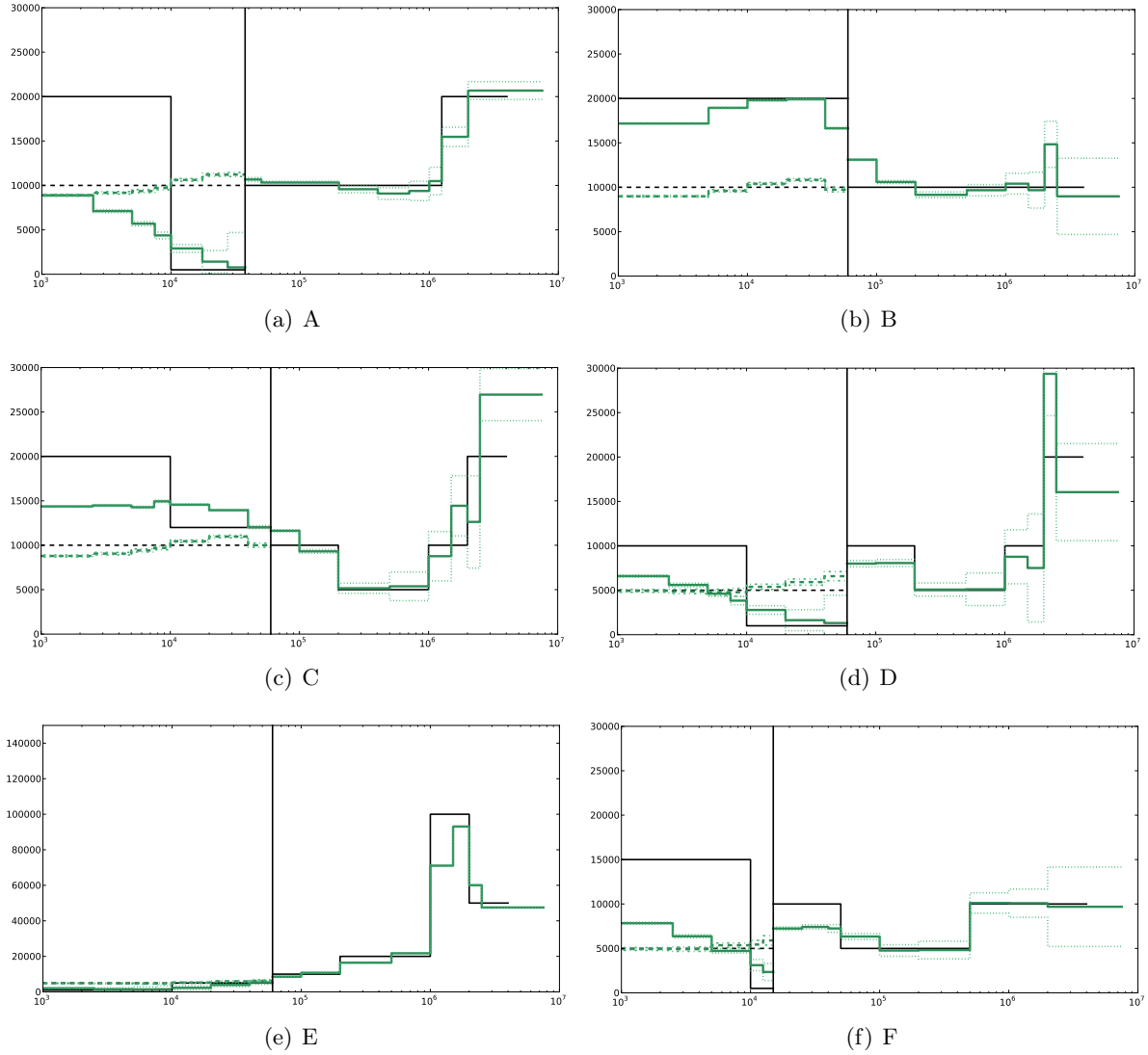
We have simulated six different datasets using different models for variable population sizes through time. Three of the datasets include a bottleneck after the split for one population and one has constant population sizes. In all of the datasets, only one of the descendant populations experiences population size changes; the other has a constant size. We used *ms* for the simulations, and each dataset consists of 300 haplotypes –150 per population– and 1,000 regions.

We used $\theta = 30$, where $\theta = 4N\mu^*$ and μ^* is the mutation rate per sequence. We used three values of T: 0.075, 0.12 and 0.03, where T is in unit of $2N$ generations. To rescale the parameters in years we need to define a value for N and to fix a number of years per generation. We decided to use the value $N = 10,000$ and a generation time of 25 years. This corresponds to times of divergence of respectively 37,500 years, 60,000 years and 15,000 years. A value of $\theta = 30$ is equivalent to a mutation rate per generation of 7.5×10^{-4} per sequence, and therefore equivalent to a mutation rate per site of $\mu = 2.5 \times 10^{-8}$ for sequences of length 30kb.

It is interesting to note that from the estimates of T and the population ratio size q_{ij}, we can rescale those estimates for any desired value of μ and years per generation without doing more computation. The idea consists of fixing the value of $\theta = 4N\mu^*$ in our importance sampler and then to deduce the value of N as a function of the desired μ. Once we know the values of N and the desired years per generation, we can rescale all our estimates of T and the populations sizes per epoch. This allows us to avoid the estimation of θ since the value chosen for N is arbitrary and any value of θ will lead to the same rescaled estimates of T and of the population sizes per epoch, as long as we are consistent.

Figure 3.1 shows the results when both T and θ are fixed to their true values. Twenty trees were built per region for each iteration of the MCEM algorithm presented in the previous section. The times of the events are situated on the x–axis on a logarithmic scale and the population sizes are on the y–axis. The vertical line represents the time of the split, the black lines the true population sizes and the green lines the estimates. The dotted lines represent the estimated size of the second descendant population, and the solid line the estimated sizes of the first descendant and ancestral populations.

We can deduce from these results that the method seems to smooth the estimates near drastic changes. We also have a tendency to slightly underestimate the population sizes, particularly when the population sizes are constant (Figure 3.1(b)). However, the most obvious concern we have with the results is the large underestimation of the population



(a) A (b) B

(c) C (d) D

(e) E (f) F

Figure 3.1: Estimates of the population sizes per epoch, using the real value of θ and T. The times of the events are situated on the x–axis on a logarithmic scale and the population sizes are on the y–axis. The vertical line represents the time of the split, the black lines the true population sizes and the green lines the estimates.

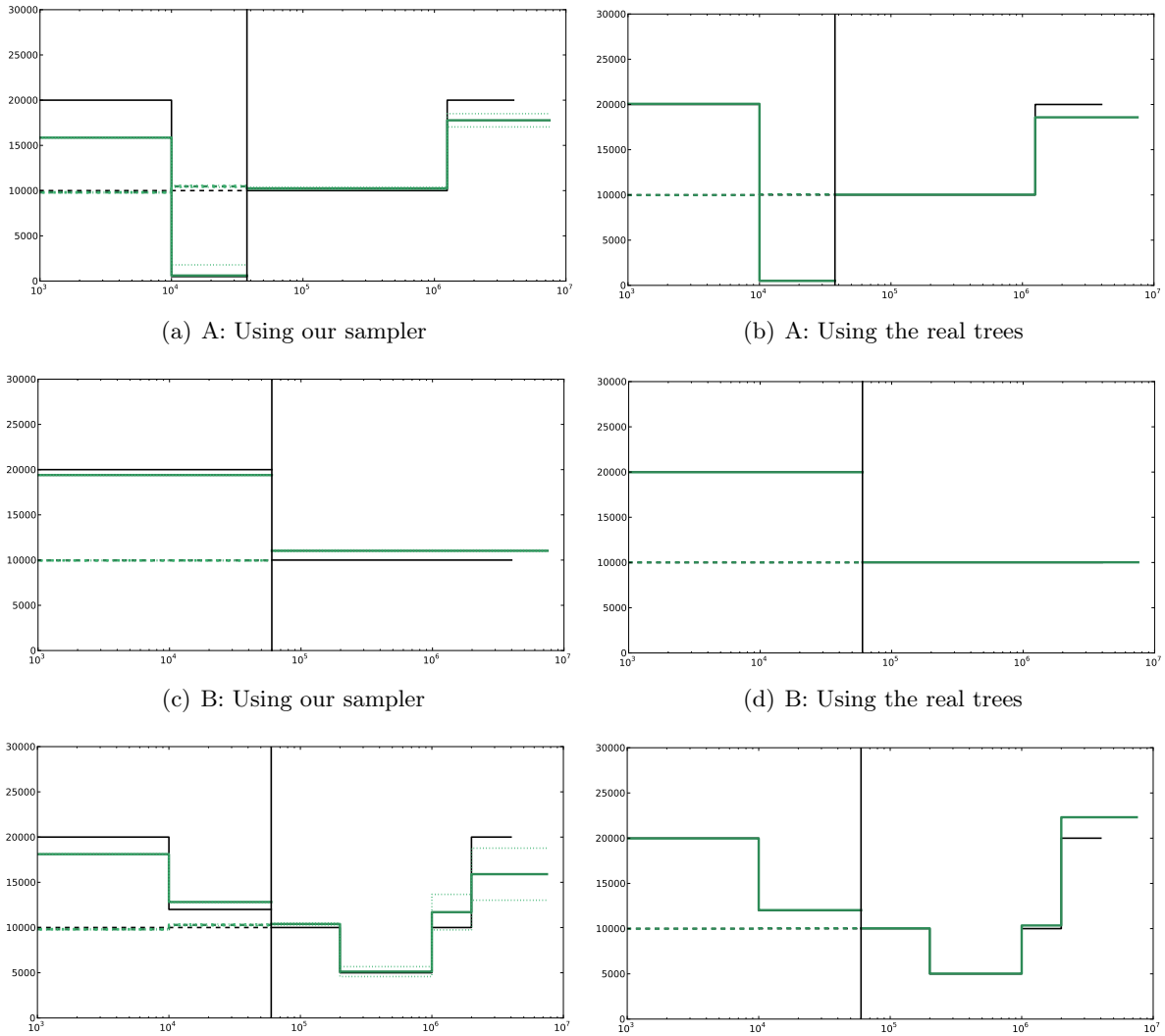
sizes after a bottleneck, when going forwards in time (Figures 3.1(a), 3.1(c) and 3.1(f)) we will later investigate this bias. In some cases, for example Figure 3.1(a), the estimate of the more recent population size is less than the half of the truth. Nevertheless, the method seems to capture some information about the changes in the ancestral population.

3.2.2 Investigation of the bias in presence of bottleneck

There are two main ways in which our method could have generated the observed bias: 1) the way we estimate the population size itself or, 2) the way our sampler works when building the trees. We can gain insight by building trees with our sampler using the true population sizes, T and θ, and then estimating the population sizes using these trees only for the time interval where there is changes. We did not used the MCEM algorithm here as we only built trees under the truth and from those we estimated the population sizes per epoch. For comparison, we also used the real trees from *ms* to obtain estimates of the population ratio sizes.

Figures 3.2 and 3.3 present the results. The first and second column of each figure use simulated and real trees, respectively. Fortunately, population sizes are estimated near perfectly when using the real trees, which suggests that the method has been correctly implemented. However, a bias is still apparent when the trees are simulated using the real population sizes (Figures 3.2(a), 3.2(e), 3.3(c) and 3.3(e)). Therefore, we can deduce that the importance sampler must be creating the bias.

To understand the bias we compared the number of coalescence events in the trees built with our sampler to the number of coalescence events in the real trees for each epoch. In general, in the earliest epochs the sampler performs more coalescent events than the real trees. In dataset B, for example, we are performing on average 144.66 coalescence events in the first epoch in the population with the largest size, compared to 123.35 for the real trees. The difference is much larger for datasets with a bottleneck, namely datasets A, D and F. In dataset A we are doing 101.72 coalescence events with our sampler and 64.05 with the real trees in the first epoch in the largest population. Performing too many coalescence events deflates estimates of the population ratio size. Therefore, our importance sampler is biased and the estimated probabilities of events are not what they should be. The Stephens and Donnelly importance sampler is correct when the population sizes are constant but appears to be biased otherwise.



(a) A: Using our sampler

(b) A: Using the real trees

(c) B: Using our sampler

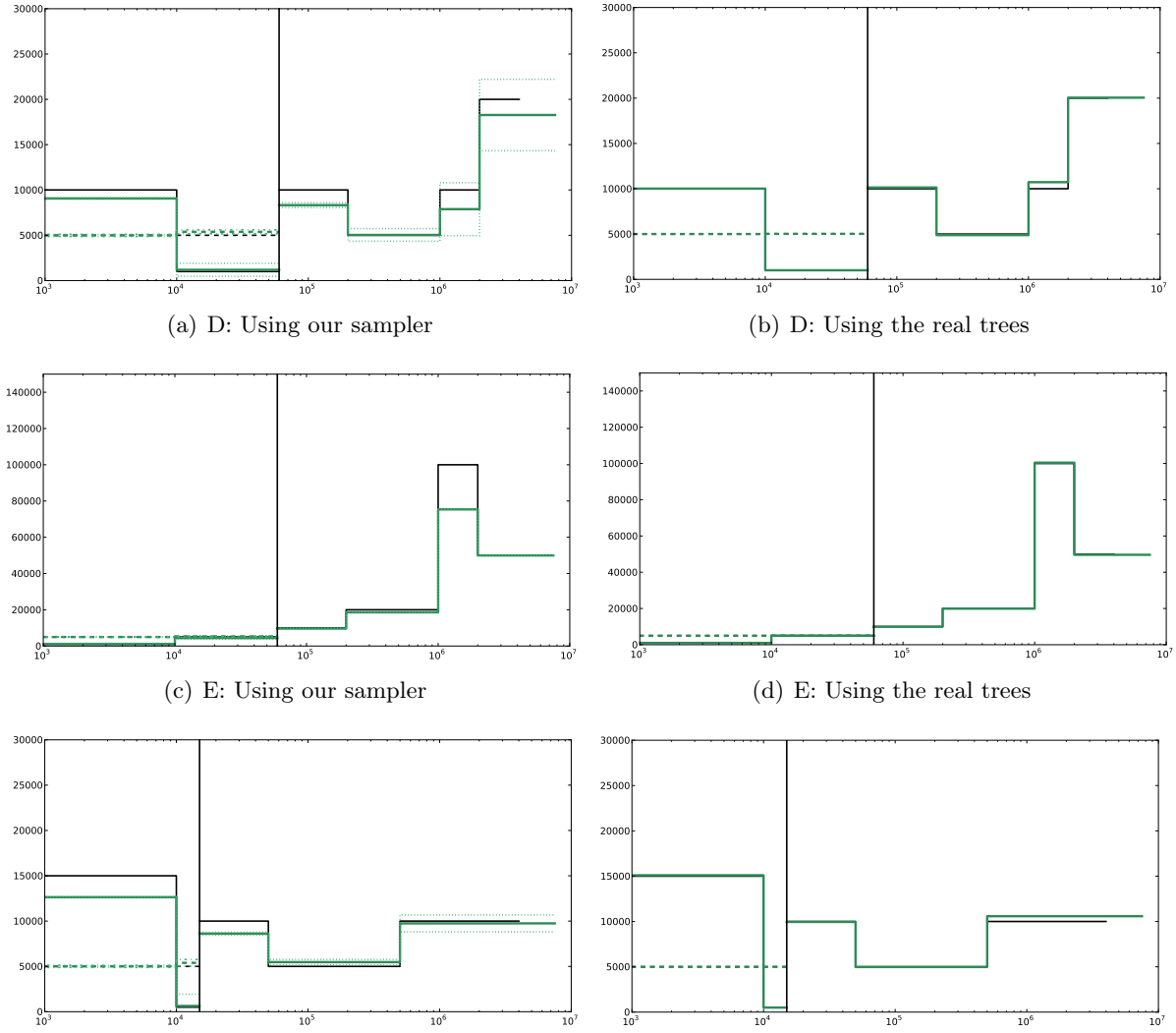
(d) B: Using the real trees

(e) C: Using our sampler

(f) C: Using the real trees

Figure 3.2: Estimates of the population sizes per epoch, using the real T, θ, and population sizes to build trees with our importance sampler. The MCEM algorithm was not used. On the left side are the estimates using our sampler and on the right side using the real trees. The times of the events are situated on the x–axis on a logarithmic scale and the population sizes are on the y–axis. The vertical line represents the time of the split, the black lines the true population sizes and the green lines the estimates.

In our importance sampler, we simply randomly pick the sequence to be involved in the next event from the ones that could have been involved. We do not use any information on the population size. This sampler has proved itself to be adequate when the coalescence



(a) D: Using our sampler

(b) D: Using the real trees

(c) E: Using our sampler

(d) E: Using the real trees

(e) F: Using our sampler

(f) F: Using the real trees

Figure 3.3: Estimates of the population sizes per epoch, using the real T, θ, and population sizes to build trees with our importance sampler. The MCEM algorithm was not used. On the left side are the estimates using our sampler and on the right side using the real trees. The times of the events are situated on the x–axis on a logarithmic scale and the population sizes are on the y–axis. The vertical line represents the time of the split, the black lines the true population sizes and the green lines the estimates.

rate is simply $\binom{k}{2}$ when there are k lineages remaining. But in our situation, the coalescence rate will rarely be equivalent to this, it can be smaller when the population size is larger than N, or larger when the population size is smaller than N. By not taking this into

account our proposal distribution will always built trees that have a coalescence rate closer to $\binom{k}{2}$. Even if we are correcting this in the importance weights via the probability of a tree (this is where we were using the correct coalescence rate in Figures 3.2 and 3.3), if the trees built are always wrong our estimates will always be biased.

3.2.3 Correcting for the bias

We now introduce two different strategies to correct the bias in our estimates. The idea behind both bias corrections is to establish at the beginning of an epoch the difference between the probabilities of an event in our sampler and the probabilities expected under the coalescent process. This bias correction is in fact a factor by which we adjust the probabilities of the events in our proposal distribution. All the events are still randomly selected when building the tree with the proposal distribution, but we modify the probabilities such that they are closer to the expected rate under the coalescent process.

To avoid over-fitting, we use simulated datasets to evaluate the bias correction so we are certain of the true parameters and of the event probabilities we should expect. Denote by b_j the bias correction applied to the rate of a mutation in our proposal distribution, and q_j the population ratio size for epoch j. Note that the bias correction could equally be defined as being applied to the rate of a coalescence. Also note that this correction will appear in our proposal distribution since they changes the probabilities of performing each events, and therefore our proposal distribution. Our implementation for both suggested corrections is as follows:

1. Simulate datasets with *ms* using the current values of T, θ and the population sizes.
2. Then for each epoch j:
   (a) Build trees with the sampler using the simulated data and all the b_k (that correct the mutation rate) such that $k < j$.
   (b) From those trees estimate b_j.

3. Build trees with the sampler using the real data and all the b_j to correct the mutation rate and estimate the population sizes from these trees.

In our method, the above steps are performed at the beginning of each iteration of the MCEM algorithm presented in the previous section.

Our first proposition is to find, for epoch i and population j, the correction b_{ij} such that the mutation rate of our sampler is equal to the mutation rate under the coalescent across all the L regions. Remember that under the coalescence process the mutation happens following an exponential distribution of rate $\frac{n_k\theta}{2}$. And under our sampler mutation happens following an exponential distribution of rate:

$$\frac{n_k((n_k-1)q_{ij}+\theta)}{2} \times \frac{m_k}{c_k+m_k}, \tag{3.13}$$

where n_k is the number of lineages remaining at the beginning of epoch i in population j, and m_k and c_k are, respectively, the number of sequences that can be involved in a mutation event and a coalescence event in our importance sampler. Therefore we want to find the bias correction b_{ij} that solve:

$$\sum_{k=1}^{L} \frac{n_k((n_k-1)q_{ij}+\theta)}{2} \times \frac{m_k b_{ij}}{c_k+m_k b_{ij}} = \sum_{k=1}^{L} \frac{n_k\theta}{2}. \tag{3.14}$$

The left side of Equation 3.14 represents the sum over all the regions of the rate of a mutation events at the beginning of epoch i in population j using our proposal distribution, while the right side is the equivalent rate under the coalescent. This equation can be simplified by:

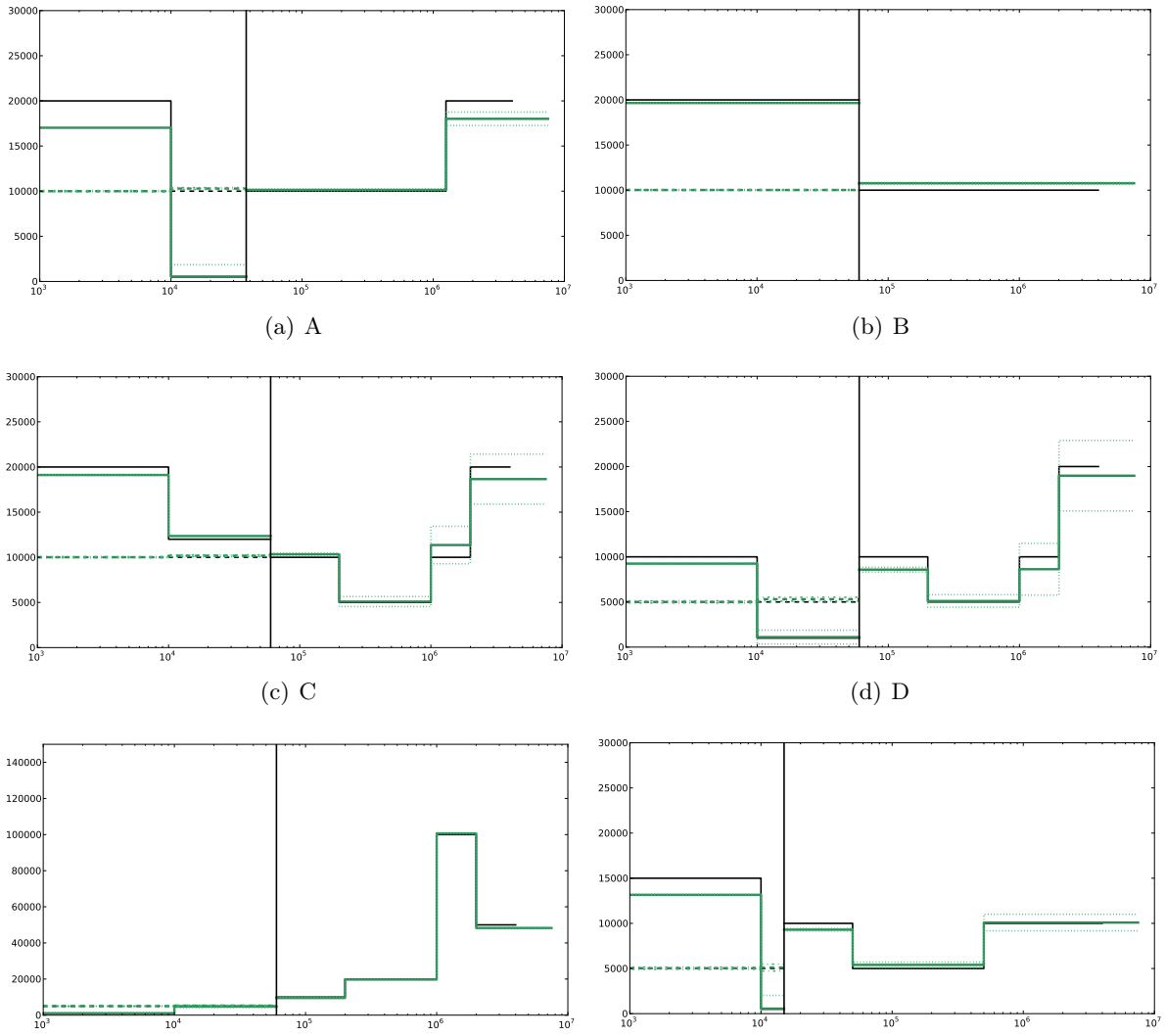
$$\sum_{k=1}^{L} n_k \left[\frac{((n_k-1)q_{ij}+\theta)}{\theta} \times \frac{m_k b_{ij}}{c_k+m_k b_{ij}} - 1 \right] = 0. \tag{3.15}$$

Equation 3.15 is then solved for b_{ij} using the Newton-Raphson method.

First, to understand if this correction helps us to get better estimates of the population ratio sizes, we have decided to use the true population sizes to simulate the data used to

estimate the bias. Therefore, after one iteration of the previous steps we should obtain nearly perfect estimates if the correction for the bias is adequate. Results are shown in Figure 3.4. The estimates we obtain are slightly closer to the truth in comparison to Figure 3.2 and Figure 3.3 (first column of both Figures), but they are not perfect. A possible explanation is that the correction is based on the probability of an event only at the beginning of an epoch. However, this probability changes during the epoch, and the correction might not be correct for the entire length of an epoch. To verify this possibility, we have estimated the population sizes using the correction and the truth but we have used smaller epochs. Results are shown in Figure 3.5. The estimates are then getting closer to the truth, with small smoothing around drastic changes. Finally, Figure 3.6 shows the results when constant and equal population sizes are used (instead of the truth) and using the MCEM algorithm. We can see that the correction improves the estimates of the populations sizes. There is still some smoothing, but the method is able to capture how the population sizes vary through time.

Correcting for the bias increases significantly the computation time, since datasets need to be simulated, and more trees need to be built to evaluate the bias correction. Moreover, the computation time will increase with the number of epochs, since the bias correction needs to be evaluated epoch by epoch.

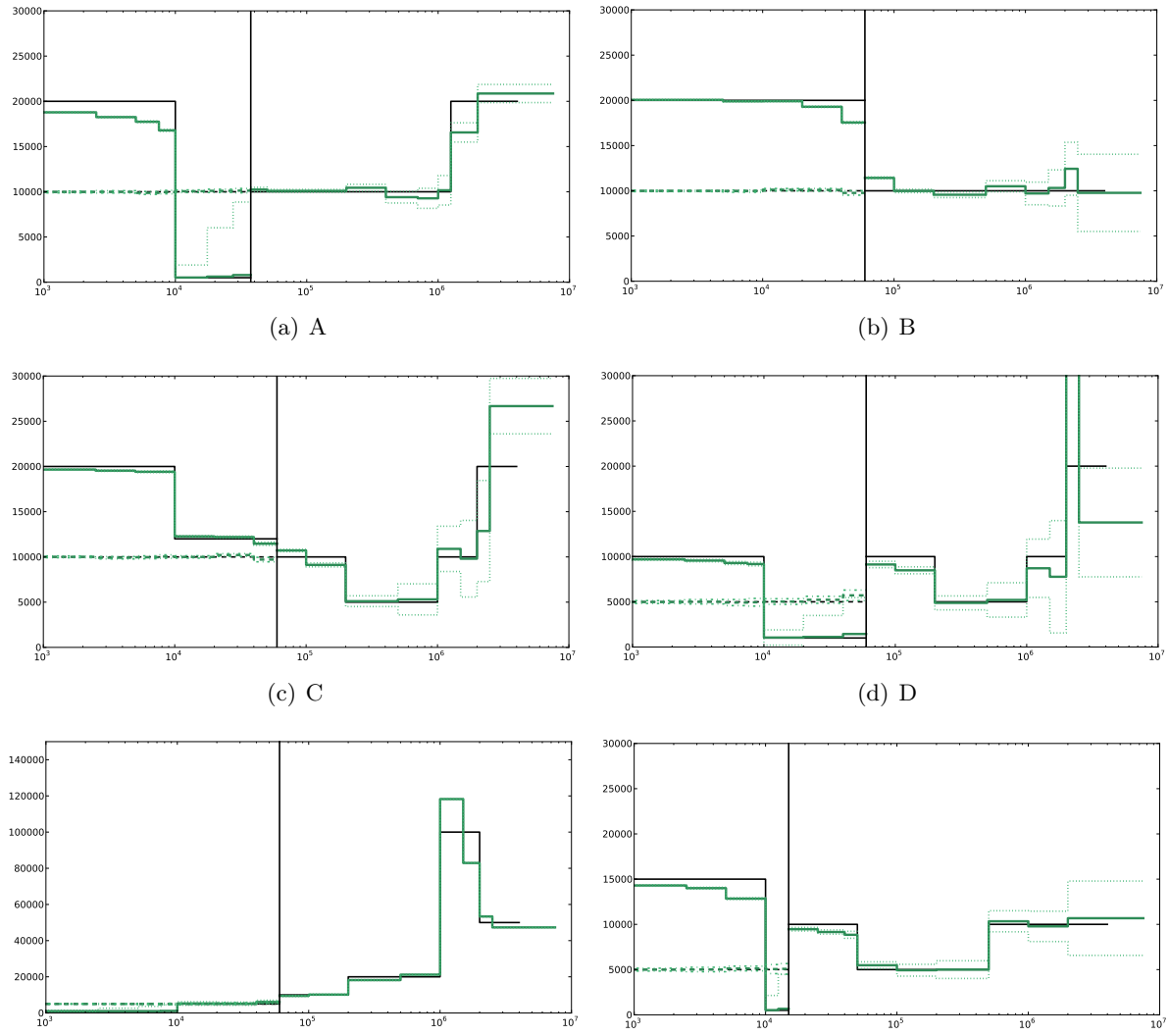


(a) A (b) B

(c) C (d) D

(e) E (f) F

Figure 3.4: Estimates of the population sizes per epoch using the first proposed bias correction. The trees are built with our importance sampler using the real T, θ, population sizes and epochs. The times of the events are situated on the x–axis on a logarithmic scale and the population sizes are on the y–axis. The vertical line represents the time of the split, the black lines the true population sizes and the green lines the estimates.

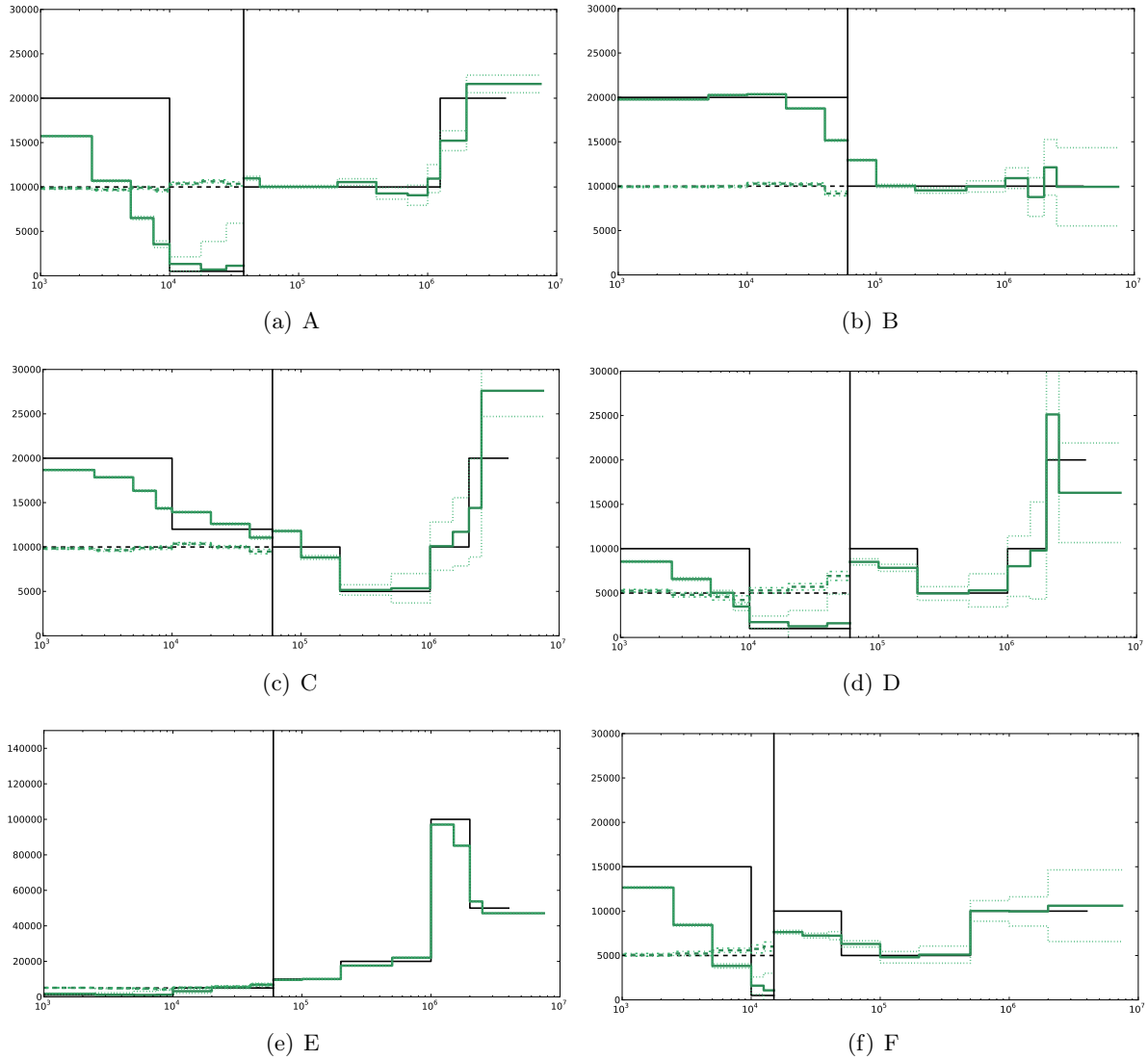


(a) A (b) B

(c) C (d) D

(e) E (f) F

Figure 3.5: Estimates of the population sizes per epoch using the first proposed bias correction. The trees are built with our importance sampler using the real T, θ and population sizes, using smaller epochs. The times of the events are situated on the x–axis on a logarithmic scale and the population sizes are on the y–axis. The vertical line represents the time of the split, the black lines the true population sizes and the green lines the estimates.



(a) A (b) B

(c) C (d) D

(e) E (f) F

Figure 3.6: Estimates of the population sizes per epoch using the first proposed bias correction. The trees are built with our importance sampler using the real T and θ. The population sizes are estimated using the MCEM algorithm with equal population sizes as initial values. The times of the events are situated on the x–axis on a logarithmic scale and the population sizes are on the y–axis. The vertical line represents the time of the split, the black lines the true population sizes and the green lines the estimates.

The second proposition is based on the observation that the true population size should make the expectation of the derivative of the log-likelihood function be zero. Therefore, a desirable property of our estimates of the population size based on trees built with our proposal distribution is that they should make the derivative of the log-likelihood function (Equation 3.2) vanish. The second proposition consists of finding the bias correction such that the derivative of the log-likelihood function is close enough to 0. If the trees built with our sampler respect the model under which we simulated the datasets, we would expect that the derivative of the log likelihood function, evaluated from our trees, be approximately 0. Unfortunately, we know that our sampler is biased when the population sizes are variable. This correction is not as straightforward to implement since the bias correction parameter does not appear in the likelihood function of the trees, therefore it cannot be found analytically. In fact, this correction adds another level of computation to the method and lengthens the execution time.

Here is an overview of the steps needed to find the bias correction. To find the bias correction we proposed to first simulate data using θ, T and the current estimate of q_{ij}. Then using those dataset, starting with the most recent epoch (i), evaluate the bias for this epoch by:

1. First, estimate the bias (b_{ij}) with the previous method (to obtain a decent starting point).

2. Then evaluate the derivative of the log likelihood function by building trees using the current bias correction. If it is not close enough to 0, apply a correction to b_{ij} (this correction is defined below).

3. Repeat the previous step until the derivative of the log likelihood is close enough to 0 (or until the maximum number of iterations allowed is reached).

4. Move to the next epoch, and repeat the above steps to find $b_{i+1,j}$ using all the b_{ij} for $j < i + 1$ when building the trees.

The correction applied to b_{ij} at step 2 is to multiply it by c when the derivative of the log-likelihood function is positive or else $1/c$, where c starts at 2 and is reduced by 0.05 each time the step 2 is repeated. We justify this correction by the fact that when we raise b_{ij} the derivative of the log-likelihood function should decrease. Hence, mutation events will have a higher probability of occurring, which increases the time between coalescence events which in turn will make the second term of Equation 3.2 bigger. Therefore, this will reduce the derivative of the log-likelihood function. Following simulation results, we have fixed the maximum number of iterations to 10.

Figure 3.7 presents the results when T, θ and the q_{ij} are fixed to their true values and using the second proposition for the bias correction. The results are now nearly unbiased and are comparable to the results obtained using the true trees. In Figure 3.8 we present the results when T, θ and the q_{ij} are fixed to their true values using smaller epochs while using the second proposition for the bias correction. The results are now on the truth for most epochs, but we can observe slight smoothing around drastic changes. In Figure 3.9, the population ratio sizes are estimated using again the real values of θ and T but starting with equal population ratio size (all $q_{ij} = 1$) and using the MCEM algorithm, where before each iteration of the MCEM a bias correction is estimated using our second proposition.

The second correction gives slightly better estimates of the population sizes than the first correction. Nevertheless, this correction is more difficult to apply, since we need to build trees at each iteration of the algorithm to find b_{ij}. Another difficulty is to define what is close enough to 0, since there is variability in the trees built with the same parameters and bias correction and therefore there is variability in the value of the log-likelihood function. It also significantly lengthens the computation time. Moreover, in most situations we found that our first proposed correction seems to bring the derivative of the log likelihood function close enough to 0 and give acceptable results. Therefore, we have decided to use the first proposition to correct the bias in our sampler.

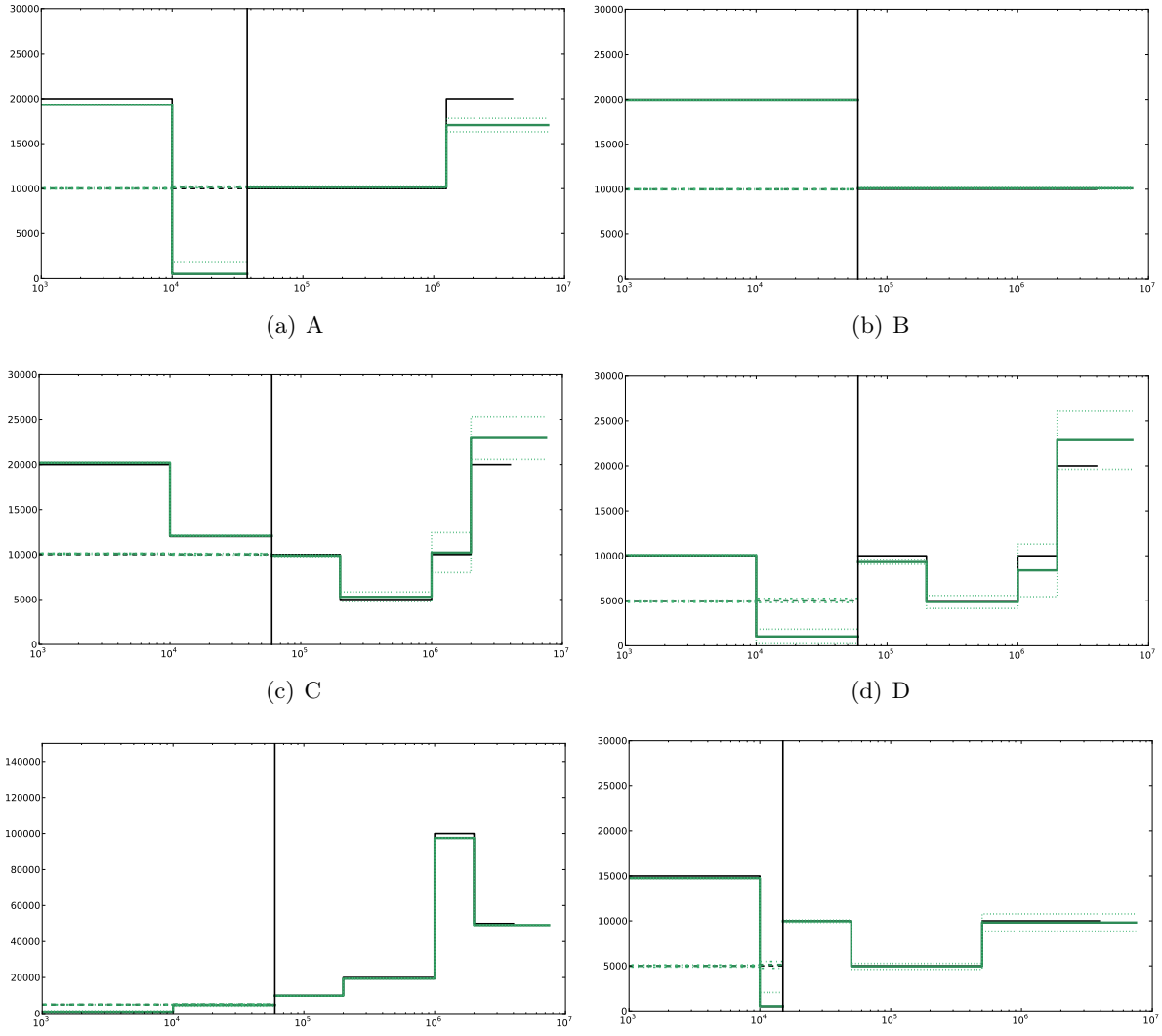
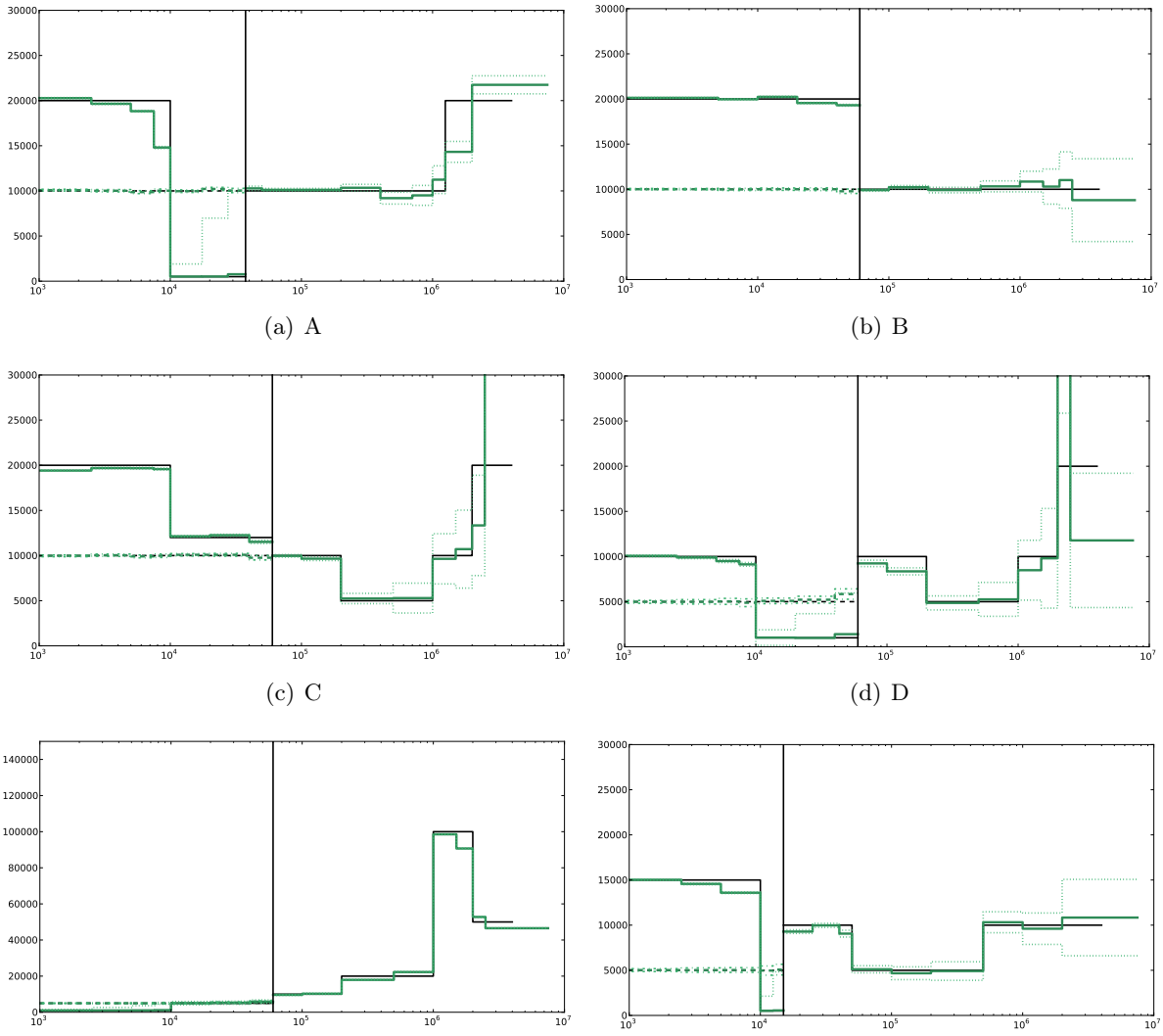


Figure 3.7: Estimates of the population sizes per epoch using the second proposed bias correction. The trees are built with our importance sampler using the real T, θ, population sizes and epochs. The times of the events are situated on the x–axis on a logarithmic scale and the population sizes are on the y–axis. The vertical line represents the time of the split, the black lines the true population sizes and the green lines the estimates.

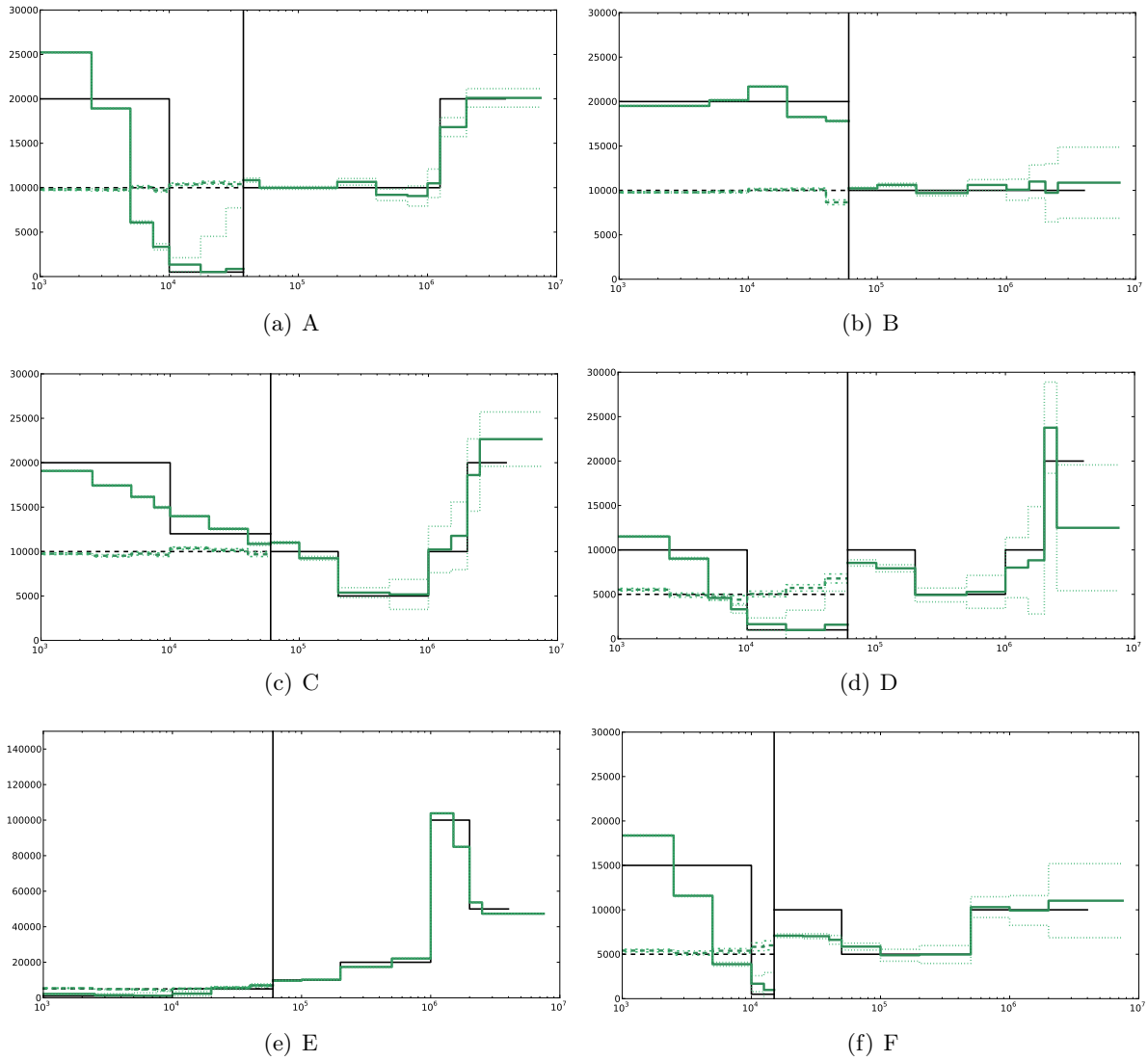


(a) A (b) B

(c) C (d) D

(e) E (f) F

Figure 3.8: Estimates of the population sizes per epoch using the second proposed bias correction. The trees are built with our importance sampler using the real T, θ and population sizes, using smaller epochs. The times of the events are situated on the x–axis on a logarithmic scale and the population sizes are on the y–axis. The vertical line represents the time of the split, the black lines the true population sizes and the green lines the estimates.



(a) A (b) B

(c) C (d) D

(e) E (f) F

Figure 3.9: Estimates of the population sizes per epoch using the second proposed bias correction. The trees are built with our importance sampler using the real T and θ. The population sizes are estimated using the MCEM algorithm with equal population sizes as initial values. The times of the events are situated on the x–axis on a logarithmic scale and the population sizes are on the y–axis. The vertical line represents the time of the split, the black lines the true population sizes and the green lines the estimates.

3.3 Joint estimation of the split time and the population sizes per epoch

We have presented how we can estimate the population sizes per epoch using trees built with our importance sampler. We now present how we combine this with our method to estimate the time of divergence. We also propose a way to obtain confidence intervals for the time of divergence using bootstrap resampling. Final results on the simulated datasets are then presented.

As mentioned in the last chapter, the trees need to be built conditionally on the time of divergence. For a fixed value of $T = T_0$, we have presented how we can obtain the MLE of the population sizes per epoch via an MCEM algorithm. Once the MLE of the population sizes are found, we can use them to build more trees and to obtain a point estimate of the likelihood function at T_0. Therefore, we can still use the optimisation algorithm to find the MLE of T presented earlier. Remember, also, that we don't need to estimate the parameter θ. The new algorithm is then:

1. A range of possible values for T is given (ex: $[0.0; 1.5]$) and a $jump$ is defined (ex: $jump = 0.25$).

2. A very fine grid of possible points is then determined using the range of possible T values (ex: all the points in the range $[0.0; 1.5]$ with an increment of 0.0005).

3. Find the three values T_1, T_2 and T_3 on that grid that are situated at the first, second and the third quarter of the range of possible values (ex: 0.375, 0.75 and 1.125).

4. Then repeat the following steps until the MLE of T is found:

   (a) The MCEM algorithm is run for the current T_i, until convergence of the estimates of the populations sizes per epoch (using the first bias correction proposed).

(b) Trees are then built for all of these points (T_i) using the population sizes and the bias correction estimated in (a) to obtain an estimate of the log-likelihood for all T_i (using our proposal distribution).

(c) The point T_{max} that gives the maximum likelihood is kept in memory.

(d) The new values of T_i to evaluate are $T_{max} \pm jump$.

(e) If T_{max} is the same as the previous one, the size of $jump$ is decreased (ex: by 0.05 until $jump = 0.05$, then by 0.025 until $jump = 0.025$ and by 0.005 until $jump = 0.0025$).

5. The MLE of T is considered to be found when the maximum has remained the same for two consecutive steps and $jump$ is at its minimum (ex: $jump = 0.0025$).

6. The MLE of the population sizes per epoch are the ones found using the MLE of T.

The estimate of the time of divergence is obtained using an estimated likelihood function. Therefore, there is some variability in the MLE of T due to the variability in our estimate of the likelihood. The confidence intervals presented in the last chapter assumed that we were using the real likelihood surface. Our proposal for confidence intervals is to estimate the standard deviation for each point estimate of the likelihood of T using a bootstrap method to resample the importance weights. For each value of T evaluated and each region, we resample with replacement M importance weights and evaluate the logarithm of the mean of the sample. By repeating a large number of times, we can evaluate the variance of the mean, *i.e.* our estimate of the log-likelihood at this value of T. Therefore, for each value of T where the likelihood is estimated we can draw confidence intervals around three standard deviations of the likelihood estimate. Our proposal for the confidence interval of the MLE of T is then the values of T for whose confidence interval overlaps the confidence interval of the MLE.

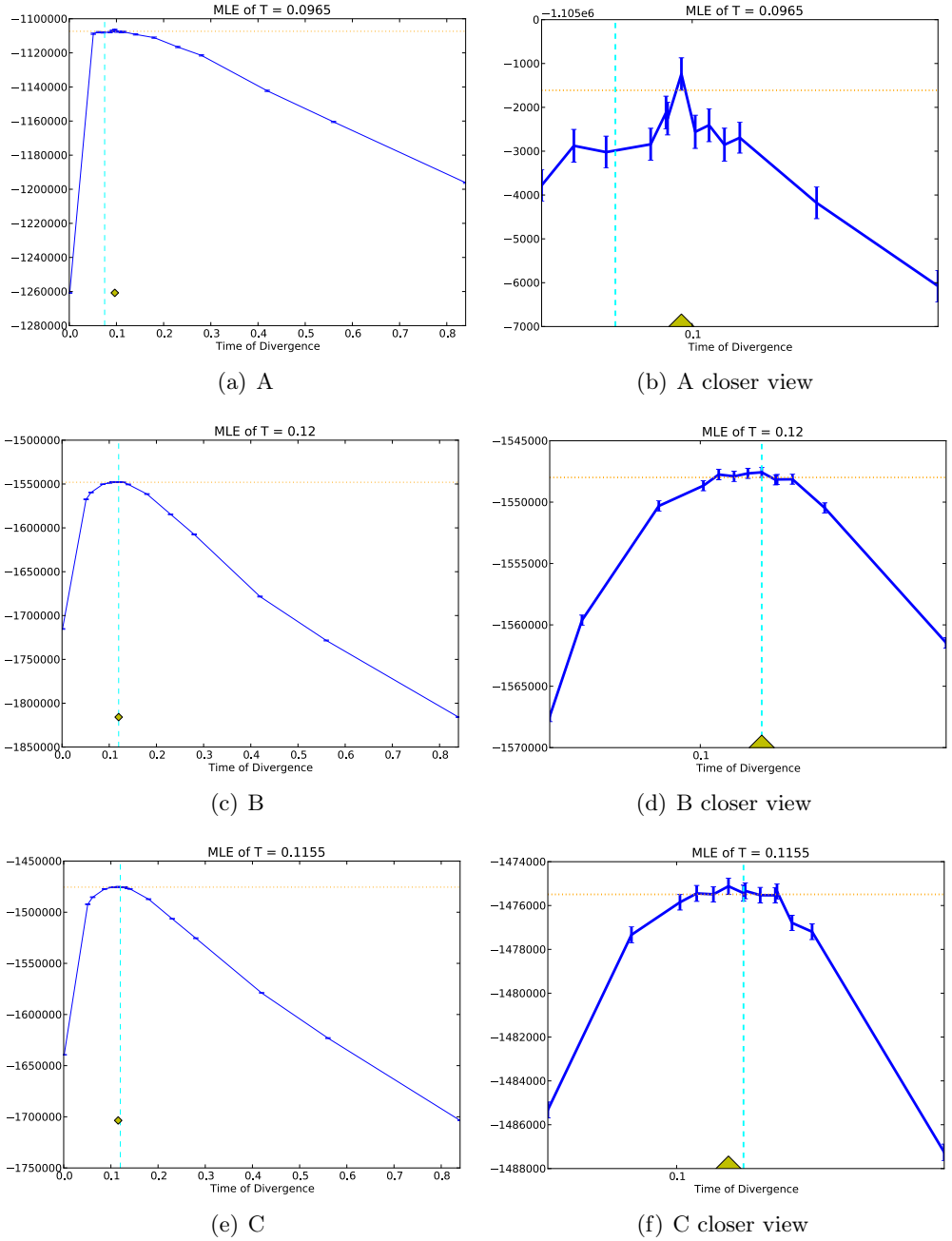
We have applied our method to jointly estimate the time of divergence and the variable population sizes of the six datasets presented previously. We built fifty trees per region

Data	T	$\hat{T}_{MLE}$	C. I.	
A	0.075	0.0965	(0.092)	(0.101)
B	0.12	0.12	0.106	0.13
C	0.12	0.1155	0.101	0.13
D	0.12	0.13	0.101	0.1345
E	0.12	0.096	(0.0915)	(0.1005)
F	0.03	0.041	0.0365	0.0455

Table 3.1: Results of the estimation of T for six datasets and their confidence intervals.Note that a confidence interval bound in parentheses means that this value is the closest one on that side of the MLE for which the likelihood was estimated, but this value is not included in the confidence interval. Therefore, the MLE is the limit of the confidence interval on this side.

in the MCEM algorithm and 200 trees when estimating the likelihood of T. Figures 3.10 and 3.11 present the estimates of the log-likelihood function of the time of divergence, where the second column is a closer view of the region near the mode of the log-likelihood function. The cyan vertical lines represent the position of the true value, and the yellow diamond the position of the MLE. The orange horizontal line represents log-likelihood values that are included in the MLE confidence interval. Therefore, if the confidence interval of the likelihood of a value of T includes the orange line, this value of T is included in the confidence interval of the MLE T. Table 3.1 summarises the results and gives the maximum likelihood estimates of T and the confidence intervals. Note that a confidence interval bound in parentheses means that this value is the closest one on that side of the MLE for which the likelihood was estimated, but this value is not included in the confidence interval. Therefore, the MLE is the limit of the confidence interval on this side.

From the results, we see that our estimates seem biased in the presence of a strong bottleneck as in datasets A and F (Figures 3.10(b) and 3.11(f)), where neither confidence interval includes the true T. Moreover, we seem to underestimate the variances since consecutive values of T sometimes have non-overlapping confidence intervals. This can cause strange bumps in the likelihood surface and, as a result, the confidence intervals might not include the true T, as in dataset E (Figure 3.11(d)). Another idea to estimate

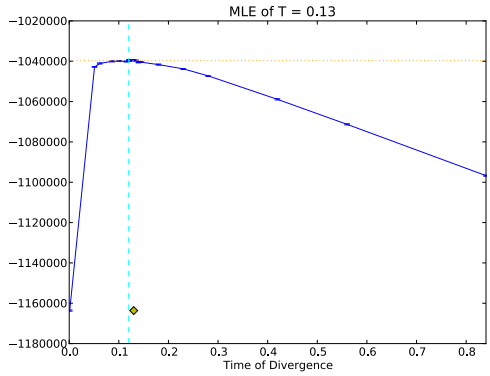


(a) A (b) A closer view

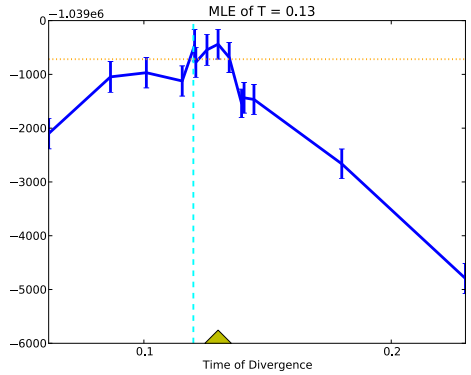
(c) B (d) B closer view

(e) C (f) C closer view

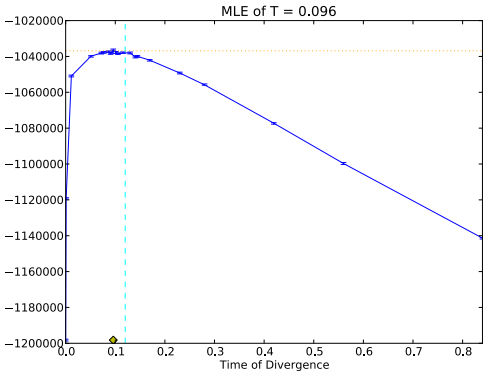
Figure 3.10: Log likelihood of T for three different scenarios of variable population sizes. The second column is a closer view of the region near the mode of the log-likelihood function of the figures in the first column.



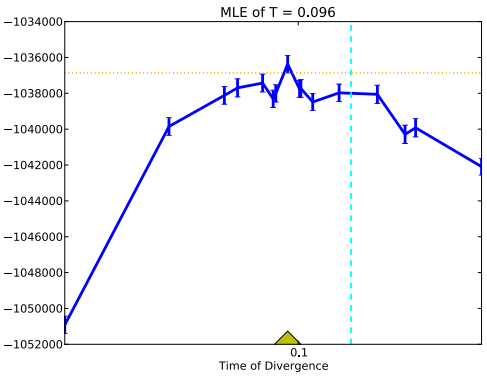
(a) D



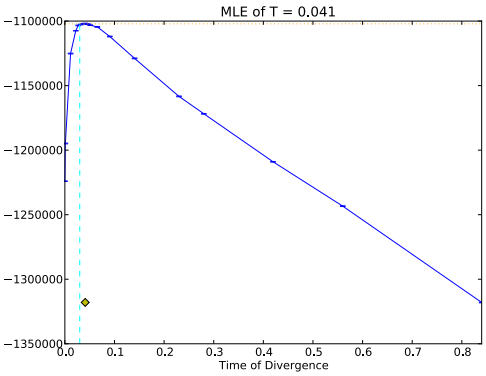
(b) D closer view



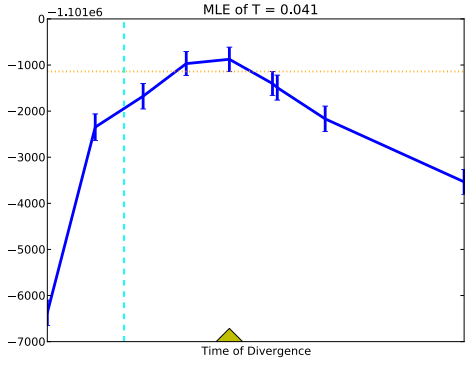
(c) E



(d) E closer view



(e) F



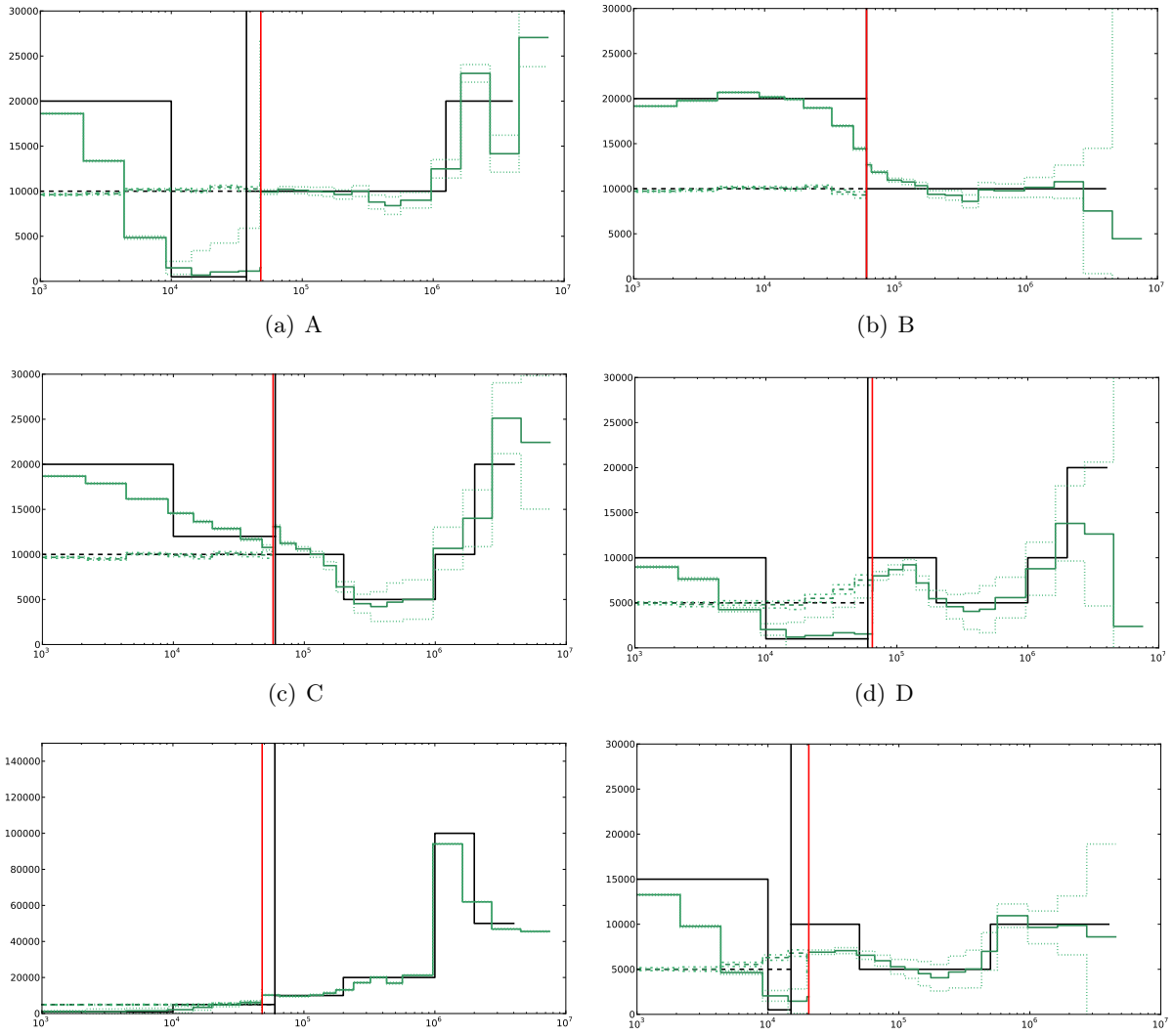
(f) F closer view

Figure 3.11: Log likelihood of T for three different scenarios of variable population sizes. The second column is a closer view of the region near the mode of the log-likelihood function of the figures in the first column.

the variance is to resample the region instead of the importance weights. This might better capture the variation in our estimates. However, more trees would need to be built for each new sample of regions, which would drastically increase the computation time.

Figure 3.12 presents the estimates of the variable population sizes and their confidence intervals. Overall, the method gives relatively accurate estimates that faithfully represent how the population sizes have varied through time. However, we still underestimate slightly the population size when the population size is larger than N before a bottleneck (going backward in time). Which have for effect a light overestimation of the population sizes during the bottlenecks. We see also that when a drastic change occurred during an epoch, the population size estimate for this epoch is an average of the population sizes before and after the change. The method seems also to smooth the population sizes near drastic changes in more recent times, and unsurprisingly the estimates become less accurate and more variable around two million years ago.

In this chapter, we have presented a novel method for the joint estimation of the time of divergence of two populations and their variable population sizes per epoch.



(a) A (b) B

(c) C (d) D

(e) E (f) F

Figure 3.12: Estimates of the population sizes per epoch in green and time of divergence in red. The times of the events are situated on the x–axis on a logarithmic scale and the population sizes are on the y–axis. The black lines represent the true population size and time of divergence.

Chapter 4

Robustness to model misspecification

The method presented in the last two chapters makes a variety of strong assumptions. The model assumes a clean split without any migrations between the descendant populations. The data is assumed to be free of recombination events and correctly phased with knowledge of the ancestral allele type. Of course, reality is more complex, and understanding the limits of our method is important. We have performed multiple simulation studies to see how the method behaves when different assumptions of the model are not met. The results are presented in this chapter. First, we will look at the impact of migration on the estimation of the time of divergence and on the population sizes per epoch. Then, we will see how the method is affected by an admixture event that occurs either with a population not sampled or with the other descendant population. Finally, we will see the impact of using data with a low recombination rate. We will also discuss briefly the possible impact of the sample size and genotyping and phasing errors.

4.1 Robustness in the presence of migrations

Our method does not allow migration events between descendant populations as some other methods do. This restriction can influence our parameter estimates since migration events

can introduce new alleles in a population and then increase the proportion of shared alleles between two descendant populations. As a result, our method might underestimate the time of divergence. Another realistic migration scenario is a smooth split where, following the split, the two descendant populations exchange migrants for a certain period of time and later diverge to total isolation.

We have designed two simulation studies to understand the effect of migration on our estimates. In the first, three parameters were allowed to vary: the time of the split T; the time at which the migration band begins; and the migration rate. But, in this scenario, the population sizes were kept constant and equal and we did not try to estimate them. For the second simulation study, we used a model with a migration band only, and we tried different scenarios of variable population sizes.

4.1.1 Effect of migration on the time of divergence estimates

The first simulation study was designed to see the effect of the migration rate on the estimated time of divergence. We used six migration rates M and we used five different values for the time of divergence T. Going backwards in time, three different scenarios were analysed: migrations starting at times 0, $T/3$ or $2T/3$. Note that migration starting at $T/3$ implies that there had been migration events in the interval $T/3$ to T. In total, we have simulated 90 datasets, one for each possible combination of M and T for all three scenarios. Each simulated dataset consists of 1,000 regions for 300 sequences (150 per population) using a scaled mutation rate $\theta = 30$.

The scaled migration rate is defined as $M = 4Nm$ (when time is rescaled using $2N$), where m represents the fraction of each population that is replaced by migrants from the other population at each generation. We used the following six scaled migration rates: $M =$ 0.01, 0.1, 1.0, 10, 100, and 1000. With $N = 10,000$, these scaled migration rates represent 0.0025, 0.025, 0.25, 2.5, 25, and 250 individuals migrating into the other population per generation. The time of divergence varied from 1,200 to 2,000 generations. We used the

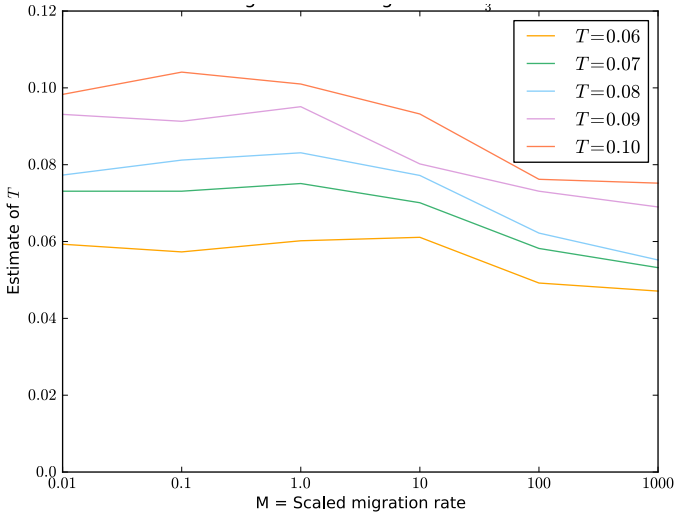


Figure 4.1: Estimates for the divergence times with migrations during time interval $[\frac{2T}{3}, T]$. The x axis represents the scaled migration rate and the y axis the estimates of T. Each line represents one of the true T.

version of the method presented in Chapter 2, where the population sizes are not estimated. We assumed knowledge of the mutation rate, and we built 100 trees per value of T.

The results from the estimation of the time of divergence are presented in Figures 4.1, 4.2 and 4.3. On each figure, the x axis represents the scaled migration rate and the y axis the estimates of T. Each line represents one of the true T values.

In Figure 4.1, migrations were simulated in the time interval $[\frac{2T}{3}, T]$. This can be viewed as a smooth split where the two descendant populations exchange individuals for a period of time and then are completely isolated. We see that the effect of migration events is subtle; our estimates seems to be biased only when the scaled migration rate is greater than or equal to ten (equivalent to 2.5 individuals per generation). When the migration rate is large, we underestimate the time of divergence, but our estimates are still larger than the $\frac{2T}{3}$ limit of the migration band.

When migrations are simulated in the time interval $[\frac{T}{3}, T]$ (see Figure 4.2), the effect of migration is greater: the two descendant populations exchange haplotypes for a longer time period. Similar to the above situation, we start to underestimate the population size

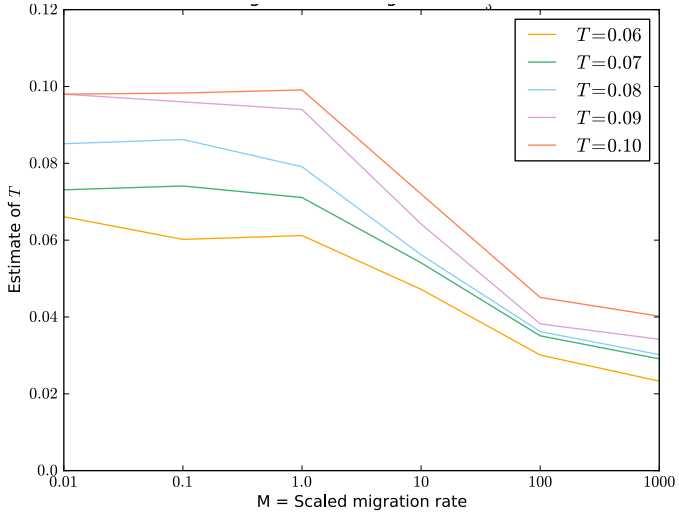


Figure 4.2: Estimates for the divergence times with migrations during time interval $[\frac{T}{3}, T]$. The x axis represents the scaled migration rate and the y axis the estimates of T. Each line represents one of the true T.

when M is greater than one. When the migration rate is at its largest ($M = 1,000$), our estimates are roughly a third of T; we estimate T to be the beginning of the migration band.

Finally, when the two descendant populations constantly exchange migrants (Figure 4.3), we rapidly see an effect on the estimated time of divergence. Our estimates are biased for migration rates greater than 0.1. In fact, our estimates converge toward 0 as the migration rate increases.

As expected, our estimates can be strongly biased in the presence of migration events. The longer the interval of time during which the populations exchange migrants and the higher the rate of migration, the stronger the bias. However, when we are in the presence of a migration band, instead of constant migrations, the migration rate needs to be high to affect our estimates. This is encouraging, since in reality an instant split might not be realistic for human populations. A more realistic scenario is that the descendant populations exchange migrants for a certain period of time followed by isolation. Our method has shown to be robust to this scenario as long as the migration rate is not too high.

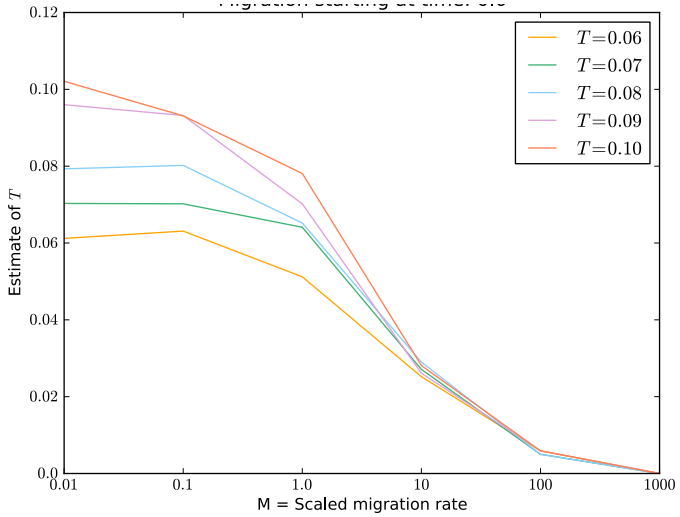


Figure 4.3: Estimates for the divergence times with migrations during time interval $[0, T]$. The x axis represents the scaled migration rate and the y axis the estimates of T. Each line represents one of the true T.

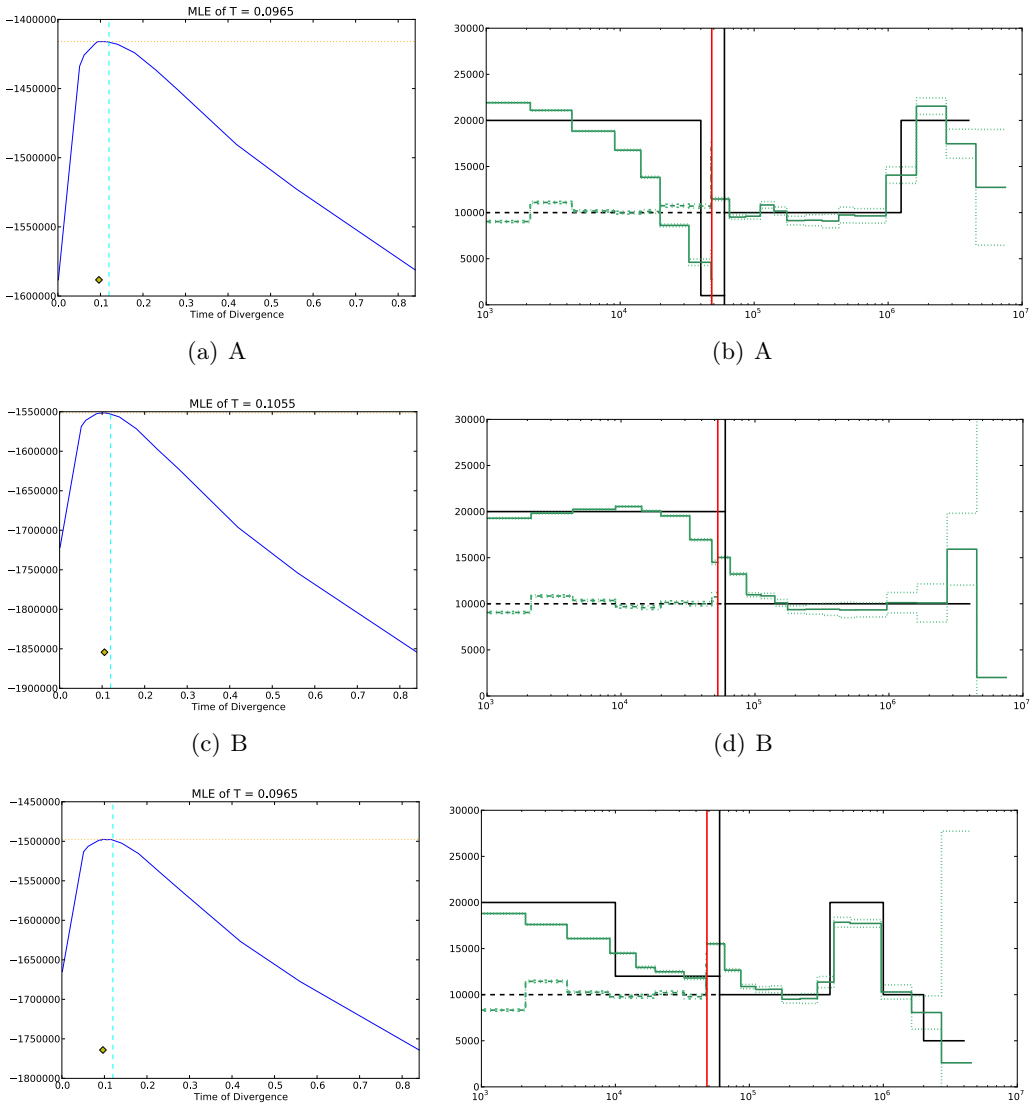
4.1.2 Effect of migration on the population sizes per epoch estimates

The second simulation study assesses the effect of migration events on our estimates of the population sizes per epoch. In particular, we want to see if our estimates are biased when there is a band of migrations of length $T/3$ after the split, since, as mentioned previously, this is a perhaps realistic scenario. We used one migration rate of 10 and two values for the time of divergence: 0.12 and 0.24 (equivalent to 2,400 and 4,800 generations (60,000 and 120,000 years ago for a generation time of 25 years)) and three different scenarios of variable population sizes.

The first scenario (A) consists of an ancestral population that decreases in size, then splits, and one of the descendant populations experienced a strong bottleneck of 20,000 years followed by a rapid expansion. In the second scenario (B) the population sizes are kept constant. And finally in the third scenario (C), an ancestral population experienced an expansion followed by a decrease and a split. Then, one of the descendant populations experienced an expansion. We have simulated a dataset for each combination of T and scenario. Each simulated dataset consists of 1,000 regions for 300 sequences (150 per

population) using a scaled mutation rate $\theta = 30$. We built 200 trees per value of T tried and 20 trees per region in the MCEM algorithm. Results are presented in Figures 4.4 and 4.5, where the likelihood of T and the estimates of the population sizes for one dataset are presented side to side.

Figure 4.4 presents the results for the three scenarios when the time of divergence is set to 0.12. As expected, the time of divergence is slightly underestimated; this can explain why the population is overestimated during the bottleneck in the first scenario. Our estimates of the population sizes do not appear to be biased and give a good sense of how the population sizes vary through time. Figure 4.5 presents the results when T is set to 0.24. The bias in the estimated time of divergence is greater since the duration of the band of migration is greater. There is no clear evidence of a bias in the estimates of the population sizes per epoch and the estimates still give a good idea of the changes in population sizes. Although in each case the population size estimates appear lightly inflated during the migration band (most obvious in Figure 4.5(f)). This is perhaps due to effective existence of ancient population structure (due to migration) during the migration band. This might be an interesting marker of gradual population separating. And the ancestral population sizes appear to be slightly underestimated.

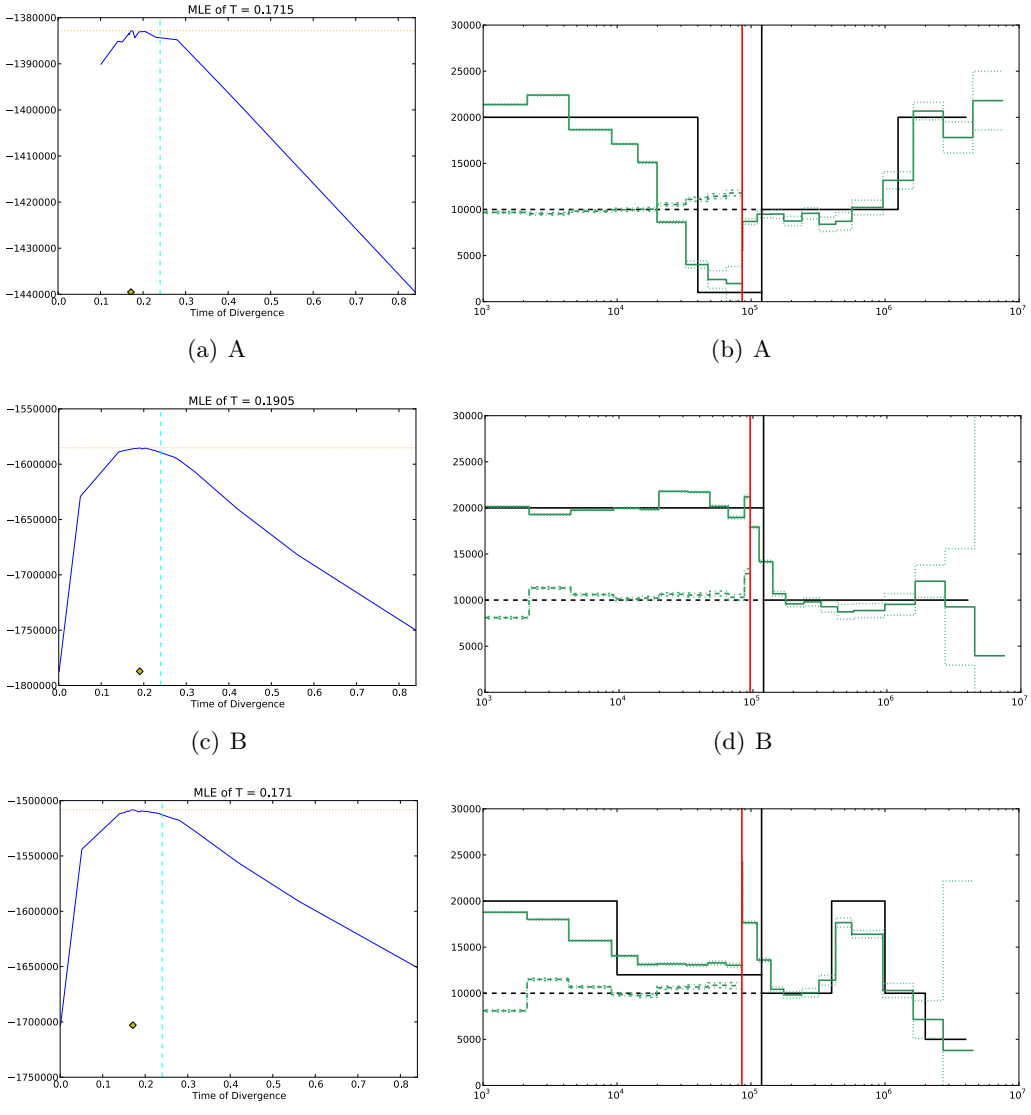


(a) A (b) A

(c) B (d) B

(e) C (f) C

Figure 4.4: Results in the presence of migration events for 20,000 years following the split. The estimate of the likelihood of T and the population sizes estimates. Each row represents the result for one of the datasets with $T = 0.12$. The dotted cyan lines in the figures on the first column represent the position of the true T. The population sizes estimates are rescaled using $N = 10,000$ and a generation time of 25 years.



(a) A (b) A

(c) B (d) B

(e) C (f) C

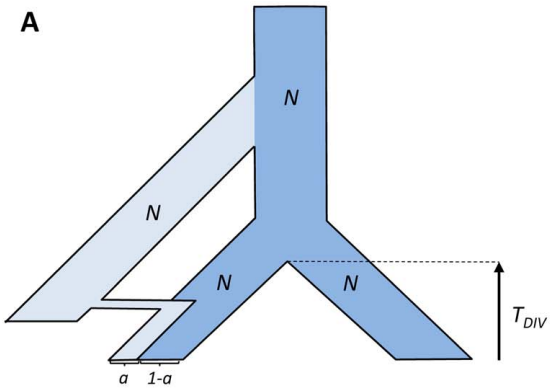
Figure 4.5: Results in the presence of migration events for 40,000 years following the split. The estimate of the likelihood of T and the population sizes estimates. Each row represents the result for one of the datasets with $T = 0.24$. The dotted cyan lines in the figures on the first column represent the position of the true T. The population sizes estimates are rescaled using $N = 10,000$ and a generation time of 25 years.

4.2 Robustness in the presence of admixture

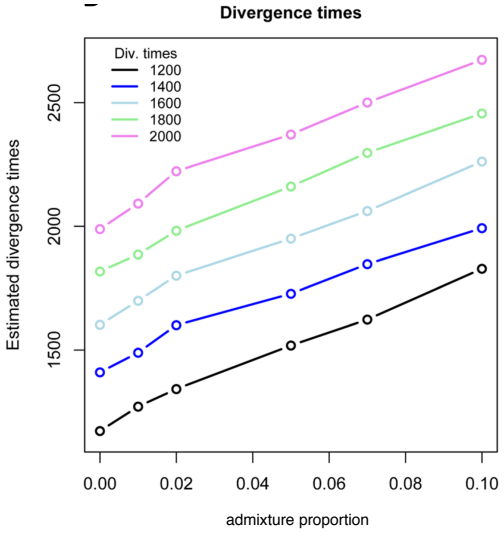
Genetic admixture is another demographic event that can introduce new genetic material to a population. It can be seen as one pulse of migration from an external population not sampled into the sampled population. At the time of the admixture event, a proportion of the sampled population is replace by lineages from a population not sampled. An example of an admixture event is presented in Figure 4.6(a), where one of the descendant populations has $a\%$ of its population replaced by a third population (not sampled) and represented in pale blue. We have done two different simulation studies to test the robustness of our method to admixture events. The first simulation study was inspired by a paper from Alves *et al.* (1) and is similar to the first test we have done with migration. The datasets were simulated assuming constant and equal population sizes and different proportions of admixture were used. The second simulation study looked at the effect of admixture on the estimates of the population sizes per epoch.

4.2.1 Effect of admixture on the time of divergence estimates

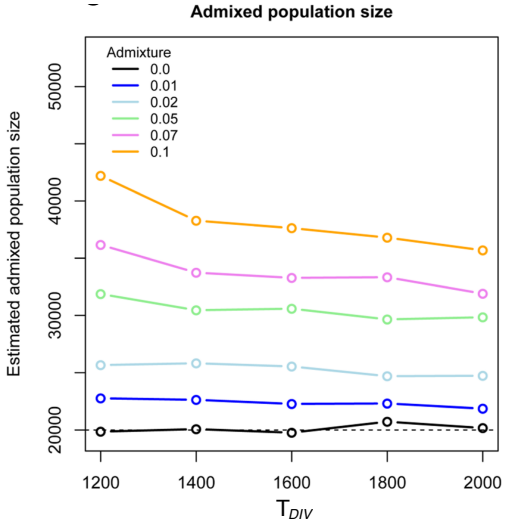
In an article published in 2012, Alves *et al.* (1) used a summary statistic method to estimate the time of divergence and the population sizes (assuming equality) of simulated data. They showed that in the presence of ignored admixture they obtained biased estimates of the divergence time and population sizes. The parameters were estimated by maximising the probability of the observed joint site frequency spectrum (SFS). The expected SFS is estimated by simulation following the approach of Nielsen *et al.* (51). We have used the same combination of parameters to simulate datasets with admixture. The model (Figure 4.6(a)) consists of two descendant populations that diverge at time T_{DIV}. One of the descendant populations is the result of an admixture event, at time T^*, with a population that diverged from the ancestral population 14,000 generations ago. At the time of the admixture, the descendant population received $a\%$ of its individuals from the non-sampled



(a) The model used



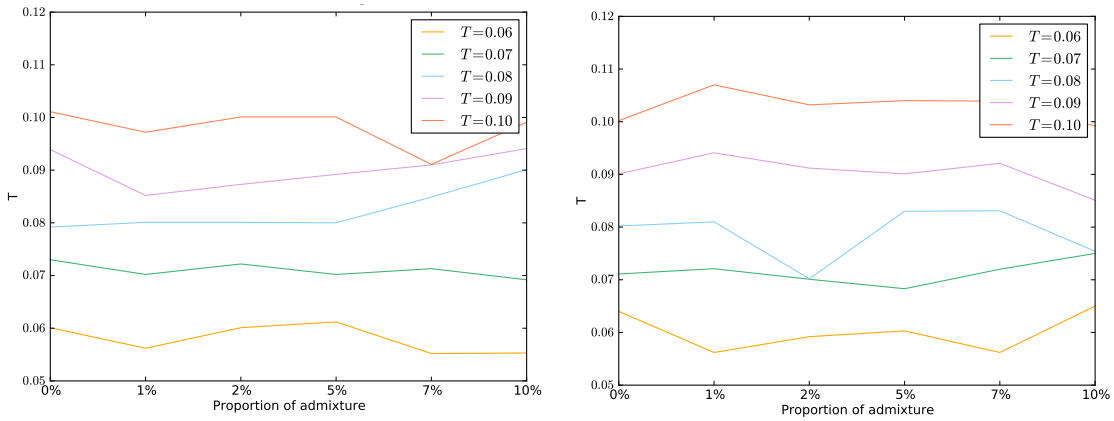
(b) Estimation of the divergence times



(c) Admixed population size

Figure 4.6: Figures from Alves *et al.*, (a) the model, (b) and (c) the results they obtained when estimating the divergence times and the size of the admixed population.

population. An example could be the estimation of the time of divergence between an European and an African population where it is believed that an admixture event occurred between the Neanderthal and the European populations, but 14,000 generations ago might be too recent in this case.

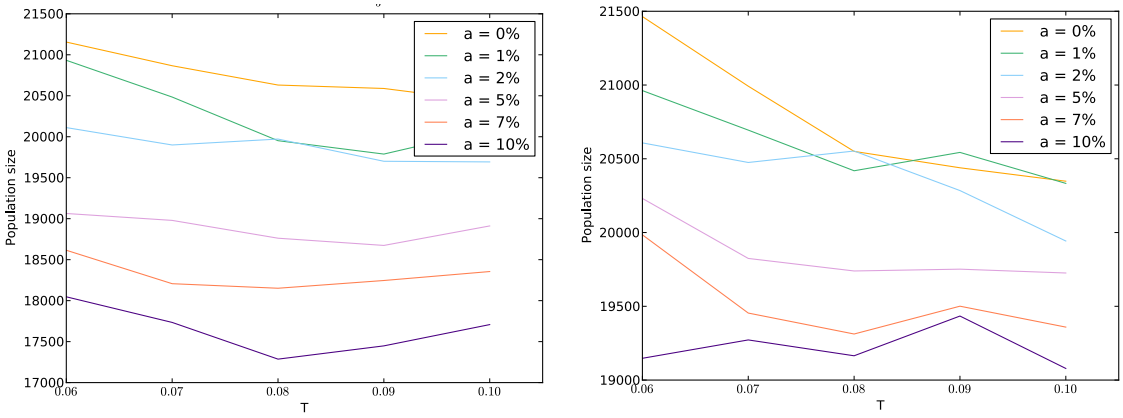


(a) (b)

Figure 4.7: Estimates for the divergence times with admixture event at times: (a) $T/3$ and (b) $2T/3$

Six different mixing proportions were used, ranging from 0% to 10%, as well as five different times of divergence, ranging from 1,200 to 2,000 generations. The only parameter not specified in the article is the time of the admixture event. The results they obtained are presented in Figures 4.6(b) and 4.6(c). Note that the population size was constant and equal to 20,000. Their simulated data consisted of 400,000 segments of 50bp, totalling a 20-Mb sequence. From their results, we see a clear bias when there is admixture: the time of divergence is overestimated, and the bias seems to grow linearly with the admixture rate. The population size is also clearly overestimated when there is admixture.

For each combination of the time of divergence and admixture rate we have simulated a dataset. Each simulated dataset consists of 1,000 regions for 300 sequences (150 per population) using a scaled mutation rate $\theta = 30$. This is equivalent to 1,000 regions of 30kb if $\mu = 2.5 \times 10^{-8}$, or 60kb if $\mu = 1.25 \times 10^{-8}$. For the admixture time T^*, we tried two different values: $T/3$ and $2T/3$. Figure 4.7 presents the results for the estimation of T when $T^* = \frac{T}{3}$ (4.7(a)) and $T^* = \frac{2T}{3}$ (4.7(b)). We do not observe any bias in our estimates of the time of divergence, but they do seem to be more variable than the ones obtained by Alves *et al.*.



(a) (b)

Figure 4.8: Estimates for the admixed population sizes with admixture event at times: (a) $\frac{T}{3}$ and (b) $\frac{2T}{3}$

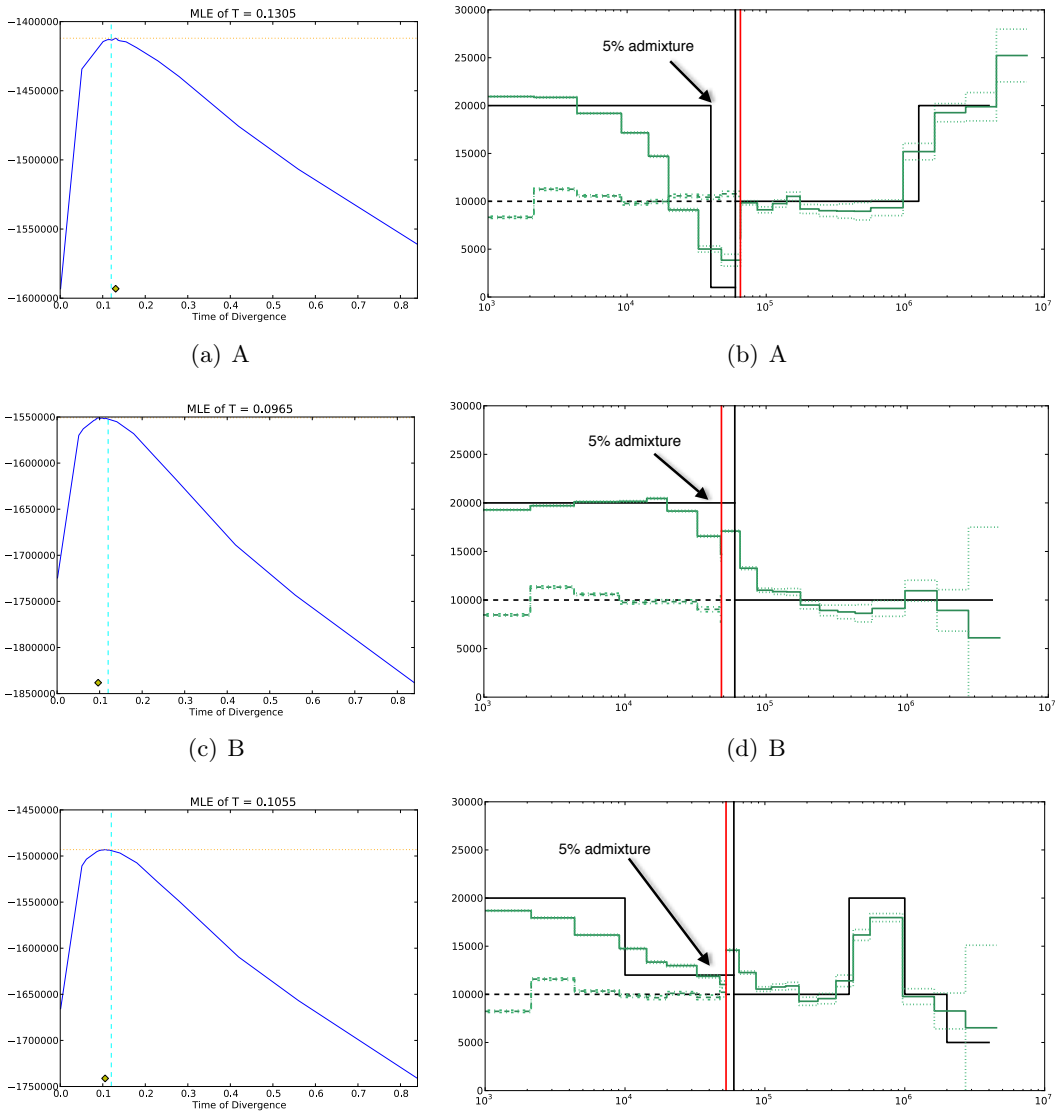
Figure 4.8 presents the results for the estimation of the population size when $T^* = \frac{T}{3}$ (4.8(a)) and $T^* = \frac{2T}{3}$ (4.8(b)). Those estimates were obtained, for one epoch, using the method presented in the last chapter using 100 trees to estimate the likelihood function and 20 for the MCEM algorithm. Note that the scale of the y-axis is not the same as the one in Alves *et al.*: it varies from 17,500 to 21,500 rather than 20,000 to 40,000. We can see that we are slightly underestimating the population size, rather than overestimating it. Admixture does not seem to bias our estimates of the time of divergence, but we see a small underestimation of the population sizes. This could be explain by the addition of new mutations in the admixed population as a results of the admixture event. Those mutations would decrease the coalescence rate and therefore reduce our estimates of the population size.

4.2.2 Effect of admixture on the population sizes per epoch estimates

We have also performed a simulation study to understand the effect of admixture on our estimates of the population sizes per epoch. This simulation is analogous to the second simulation study with migration events. We have used the same scenarios of variable

population sizes, the same two times of divergence ($T = 0.12$ and $T = 0.24$), and an admixture rate fixed at 5%. A dataset was simulated for each combination and consists of 1,000 regions for 300 sequences (150 per population) using a scaled mutation rate $\theta = 30$. However, for this simulation study, the admixture event is different and does not involve a third population. Instead, the admixture event is a pulse of migration from the second descendant population to the first population. The time of the admixture event was set to $\frac{2T}{3}$. Again an example could be the estimation of the time of divergence between Europeans and Neanderthals (but in this situation $\frac{2T}{3}$ might not be realistic).

Results are presented in Figures 4.9 and 4.10, where the likelihood of T and the estimates of the population sizes for the first population (the one that received migrants) are presented side to side. As expected from previous results, there is no clear bias in our estimates of the time of divergence for either $T = 0.12$ or $T = 0.24$. As for the population sizes per epoch, there is no drastic bias in our estimates. But, we can remark in Figure 4.10 a bump in the estimate of the size of the constant populations following the admixture event (forward in time). However, this bump is not visible in Figure 4.9, where the two descendant populations have been in isolation for a shorter period of time before the admixture event (forward in time). Because of the bottlenecks for scenario A and C it is more difficult to assess if a bump in the estimate of the population sizes is present. Also, there seems to be a slight underestimation of the ancestral population sizes, especially for datasets A and B, that is visible for both values of T.

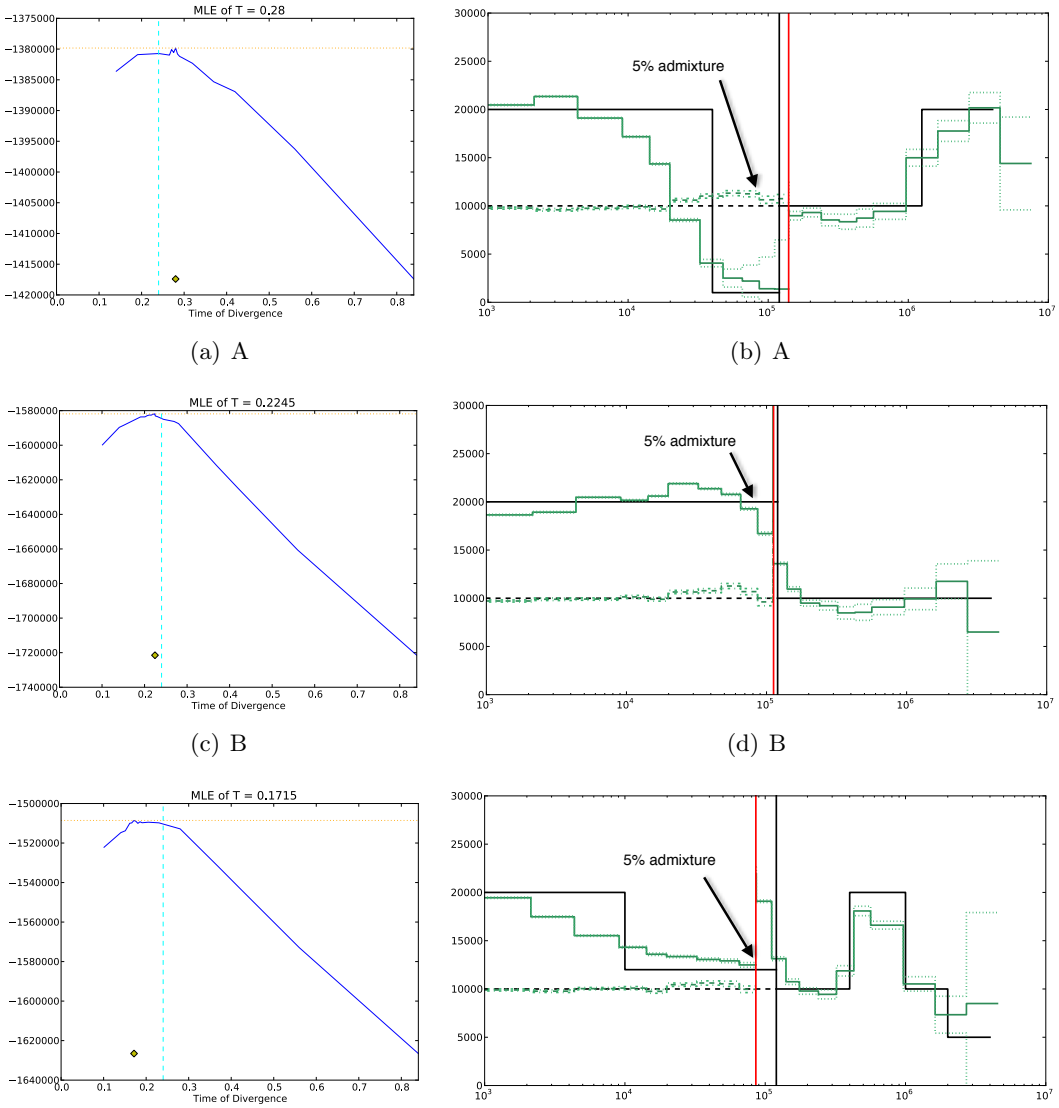


(a) A (b) A

(c) B (d) B

(e) C (f) C

Figure 4.9: Results in the presence of admixture; a pulse of migration 40,000 years ago, (represented by the tip of the arrow) from the population with constant size to the other population. The estimate of the likelihood of T and the population sizes estimates. Each row represents the result for one of the datasets with $T = 0.12$. The dotted cyan lines in the figures on the first column represent the position of the true T. The population sizes estimates are rescaled using $N = 10,000$ and a generation time of 25 years.



(a) A (b) A

(c) B (d) B

(e) C (f) C

Figure 4.10: Results in the presence of admixture; a pulse of migration 80,000 years ago (represented by the tip of the arrow) from the population with constant size to the other population. The estimate of the likelihood of T and the population sizes estimates. Each row represents the result for one of the datasets with $T = 0.24$. The dotted cyan lines in the figures on the first column represent the position of the true T. The population sizes estimates are rescaled using $N = 10,000$ and a generation time of 25 years.

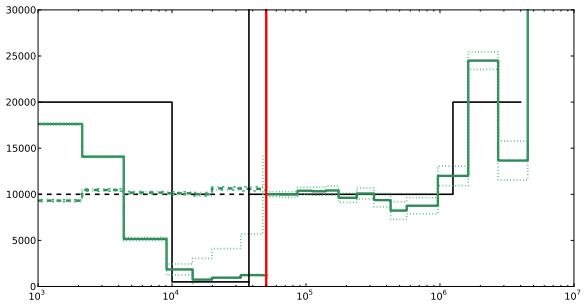
4.3 Robustness in the presence of recombination

Our last simulation study tests the robustness of our method to recombination events. The method assumed that no recombination occurred, and therefore the data necessarily had a tree-like genealogy. Moreover, the method assumes that the ancestral allele is known, and the infinite sites model of mutation. In simulation we can be certain that there is no recombination, but in reality, even if you use regions with low recombination rate, there is still a possibility that the data experienced recombination in the past. Therefore, when using real data, we need to ensure that the data agrees with a tree-like genealogy. To do so we can use the three gametes test.

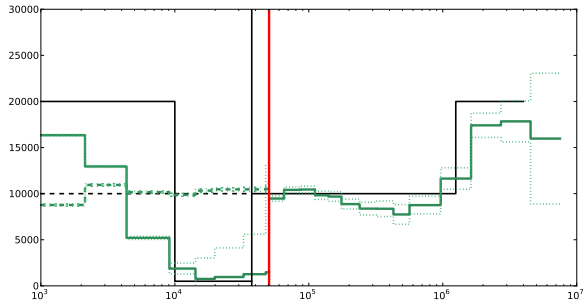
Recall the notation introduced in the first chapter, where all the sequences in a region are stored in a matrix $\mathbf{M}$. The idea behind the three gametes test is to look at all the possible pairs of columns and make sure that for any pair (j, k), we do not have all of $(0, 1)$, $(1, 0)$ and $(1, 1)$ when looking at all the (m_{ij}, m_{ik}). If this is the case, then no coalescence tree agrees with this region and a recombination event must have occurred. Our strategy to use the data is to remove columns in $\mathbf{M}$ (e.g. SNPs), until the remaining SNPs pass the three gametes test. We always first remove the SNP that has the most conflicts with others and repeatedly apply this rule.

In this simulation study, we have simulated datasets with a low recombination rate of 0.05cM/Mb to match plausible values of this rate for our real data after filtering based on recombination rate estimates (see chapter 5). We use the same six scenarios of variable population sizes as presented in the last chapter. Each simulated dataset consists of 1,000 regions for 300 sequences (150 per population) using a scaled mutation rate $\theta = 30$. For each region, we assessed whether it passed the three gametes test if not, we removed a SNP and repeated the test. On average we removed around seventeen percent of the SNPs per regions.

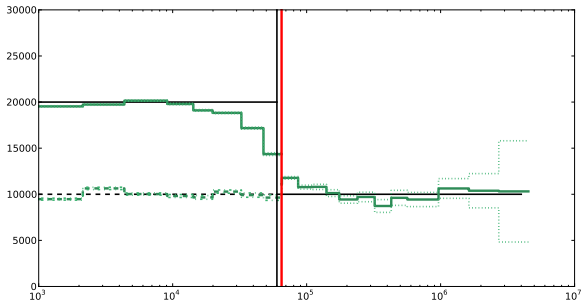
The results are presented in Figures 4.11, 4.12 and 4.13 where, for ease of comparison, the first column presents the results for datasets without recombination (the same datasets as those in the last chapter), and the second the results obtained for the datasets with recombination (but cleaned to be able to build trees). From the results, there is no evidence of bias in our estimates of the time of divergence. As for the estimates of the population sizes per epoch, we see for the more ancient population sizes, near a million years ago, that we seem to underestimate the population sizes and most strikingly our confidence intervals are larger. This might be due to removal of ancient SNPs which are more likely to have been affected by the modest amount of recombination we simulated here.



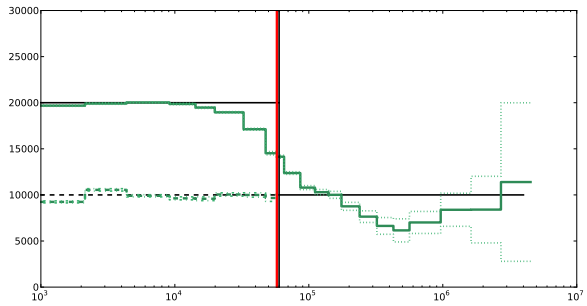
(a) A: Without recombination



(b) A: With recombination

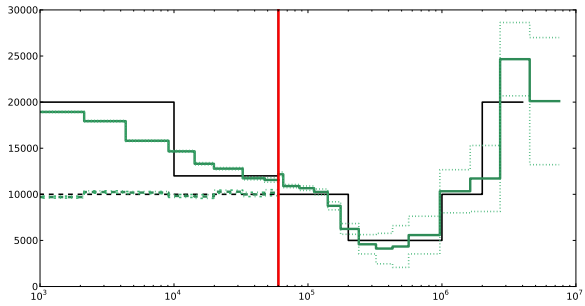


(c) B: Without recombination

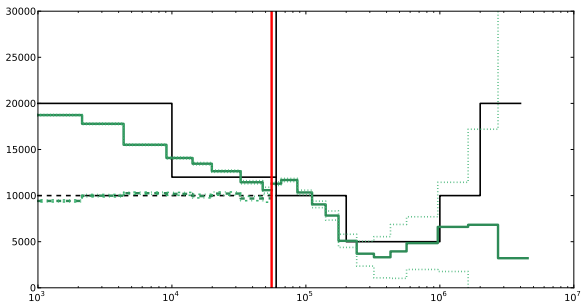


(d) B: With recombination

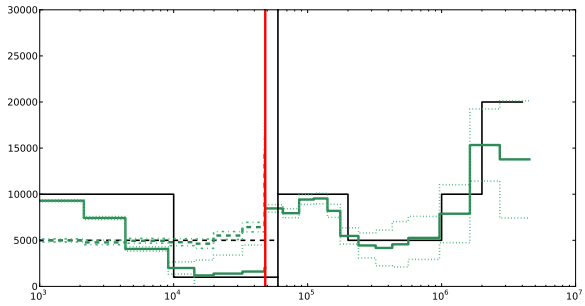
Figure 4.11: Joint estimation of the time of divergence and population sizes. The coloured lines represent the estimates and the black lines the truth. The first column is the results with datasets simulated without recombination. The second column is the results obtained with datasets simulated with a low recombination rate, and with SNP removed so that the remaining pass the three gametes test.



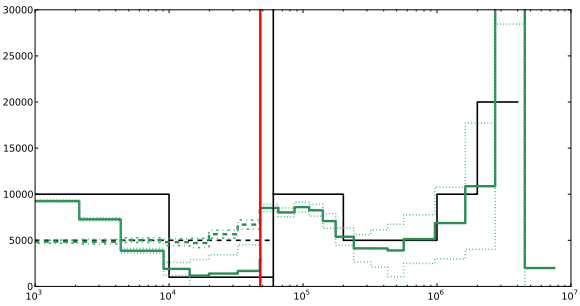
(a) C: Without recombination



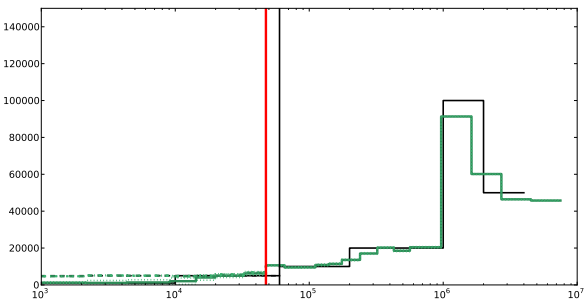
(b) C: With recombination



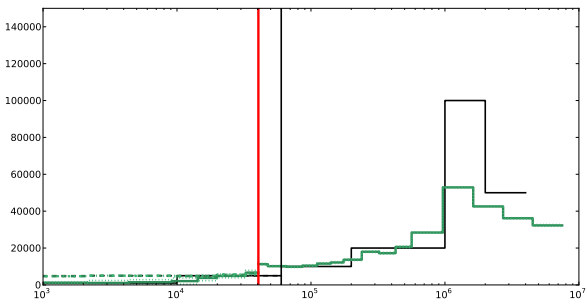
(c) D: Without recombination



(d) D: With recombination

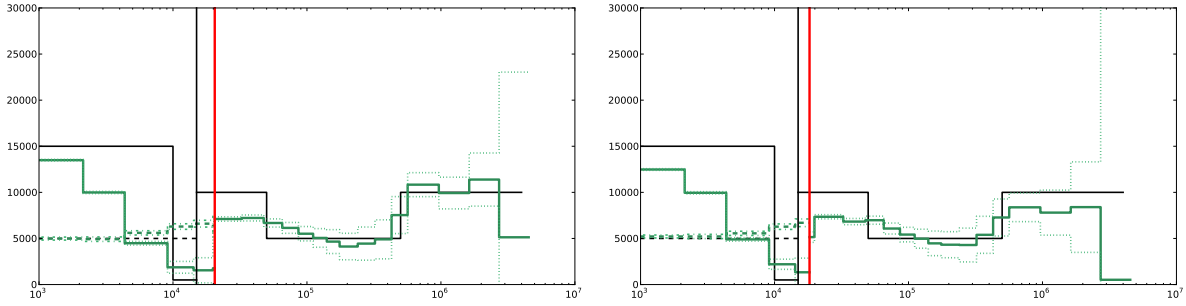


(e) E: Without recombination



(f) E: With recombination

Figure 4.12: Joint estimation of the time of divergence and population sizes. The coloured lines represent the estimates and the black lines the truth. The first column is the results with datasets simulated without recombination. The second column is the results obtained with datasets simulated with a low recombination rate, and with SNP removed so that the remaining pass the three gametes test.



(a) F: Without recombination

(b) F: With recombination

Figure 4.13: Joint estimation of the time of divergence and population sizes. The coloured lines represent the estimates and the black lines the truth. The first column is the results with datasets simulated without recombination. The second column is the results obtained with datasets simulated with a low recombination rate, and with SNPs removed so that the remaining pass the three gametes test.

4.4 Discussion

The method presented in this thesis is designed to estimate the time of divergence of two populations believed to be in isolation. The regions used need to come from parts of the genome with low recombination rates. We have performed several simulation studies to understand the limits and robustness of our method. From those, we can conclude that the method is robust to model misspecification, as long as the transgressions are not too extreme. The method is robust to some migration following the split, but not to constant migrations at a high level. We have seen that in some situations, in the presence of admixture or recombination, the method can underestimate slightly the ancestral population sizes for more ancient times.

It would have been interesting to do more simulation studies. We could in future use simulation to look at the effect of the sample size on the estimates of the time of divergence and population sizes per epoch. The aim could have been to set some boundaries on what to expect for a given sample size. We expect intuitively to have less power to estimate the recent population sizes when the sample size is smaller, since fewer coalescent events

will have occurred recently. We might ask also how the method performs using only two individuals from population trios, since trios allow better phasing of the genotypes. Another question is the effect of genotyping or phasing errors on our estimates. We believe that the current genotyping and phasing methods have relatively good accuracy and in the future these types of errors may not be an issue, as haplotypic phase becomes directly identified.

Another simulation study we would like to undertake is a comparison of methods. Unfortunately, this is not an easy task, since all the methods presented in the introduction chapter use different models and different types of data. For a fair comparison, we would need to simulate datasets on a whole genome scale for multiple individuals and then clean these datasets in a different way for each method. For our method, this would mean extracting all the regions with low recombination rate; PSMC uses the whole genome of only one individual; and for diCoal we would need to identify one large region. Direct comparison of the estimates can be difficult since, to our knowledge, our method is the only one that estimates both the time of divergence and the variable population sizes. Moreover, some methods use phased data and others genotype data. Therefore, we did not have the time to perform a fair comparison of methods.

Chapter 5

Analysis of real samples

In the last chapters, we have demonstrated that the method presented in this thesis performs well, even when the model assumptions are not precisely met. We saw that the presence of recombination only slightly affects the estimation of the ancient ancestral population size; that a split followed by a migration band will bias our estimates only when the migration band is quite long or when the migration rate is large; and that admixture does not have a great impact on our estimates. These deviations were important to test since it can be hard to assess if real data meet all of the model assumptions. In this chapter, we present the results on real samples from the 1000 Genomes Project (1KGP)(44). We are interested in how modern humans have colonised the world and how the size of these past populations have changed through time.

We have an overview of the peopling of the world by modern humans from fossil records. It is now an accepted fact that modern humans evolved in Africa around 200 thousand years ago (45) and that they were present in Europe and Australia around 40 thousand years ago (30)(55). There are still questions about the routes taken and the specific dates of the arrival of modern humans in India and the Middle East, thought to be roughly 70 thousand (59) and 100 thousand years ago (48), respectively. One possible model for the out of Africa migration is that a small population moved to the south Middle East and then

split and spread towards Asia and Europe (3). Using our method we seek to estimate the date of the out of Africa migration by comparing various European and Asian populations with an African population and to date the European-Asian split.

The 1KGP is a large collaborative effort with the goal of finding most of the genetic variants with a minor allele frequency $\geq 1\%$. For Phase 1, 1,092 individuals were sampled from 14 populations around the globe (East Asia, Africa, Europe and Americas) and their whole genomes were sequenced. The data were quickly made freely available, and soon phased haplotypes was also being released. These data are an incredible reference for researchers interested in human genetics and population histories. The final phase of the project will include the sequences of more than 1,000 additional individuals from populations in South Asia and Africa and the phased haplotypes should be available in mid 2014.

In this chapter, we first present the filtering procedure used on data to approximately meet the model assumptions of no recombination and known ancestral allele type. Then, we discuss the uncertainty in rescaling the parameter estimates and the impact of our filtering procedure on the mutation rate estimate. Finally, we present our results and compare them to previous estimates obtained from existing methods.

5.1 Data filtering

Following the model assumptions, we need to : 1) have phased haplotypes, 2) know the type of the ancestral allele and 3) identify regions of the genome with a low recombination rate. Phased haplotypes from the 1000 Genomes Phase 1 are available on the IMPUTE2 website (31) – they were phased using SHAPEIT2 (10). The 1,000 Genomes Consortium has also annotated most of their variant sites with ancestral alleles. This was done by comparing human sequences with sequences from two different types of primates. We have obtained a map of regions with low recombination rate in the human genome from Anna

Frangou (personal communication, December 2011). A region was defined as cold if the recombination rate was lower than or equal to 0.05cM/Mb across four recombination maps: YRI, CEU, deCODE, and AA. We identify 2,425 regions across the genome; they have a average length of 35kb (median 30kb).

In our regions, we kept only the SNPs for which the type of the ancestral allele was identified. Then, we needed to remove SNPs with a fixed allele, since they do not provide any information when building the trees. Finally, we remove either SNPs or haplotypes until those remaining passed the three gametes test. Therefore, for each pair of populations, we:

1. Removed all the SNPs with a fixed allele.
2. Then for each region:
    - We computed the number of SNPs to remove to pass the 3 gametes test.
    - We computed the number of haplotypes that needed to be removed to pass the 3 gametes test.
3. We decided which regions to keep by:
    - Keeping only the regions with more than 30 SNPs.
    - If more than a third of all SNPs and more than 10 haplotypes were removed from a region, then it was discarded.
    - Then for a region we choose between removing SNPs or haplotypes by:
        - Removing haplotypes only if the number of haplotypes to remove was lower than 10 and lower or equal to the number of SNPs to remove.
        - Otherwise we removed SNPs.

We chose to discard regions where we needed to remove more than a third of the SNPs to avoid regions with strong recombination signal. Overall, for our 10 pairs of populations,

we removed between 11% to 13% of the SNPs on average. We chose to remove haplotypes in a region, instead of SNPs, between 2% to 5% of the time, depending on the pair of populations used. Finally, we discarded around 45 regions for each pair analysed.

5.2 Scaling the estimates

With the use of real data comes the difficulty of rescaling the estimates into years. We have previously explained how this scaling is in part related to the mutation rate μ and also the length of a generation in years. There is no real consensus on the values to use: estimates for μ vary by a factor of two and possible generation times range from 20 years to 30 years. In this section, we discuss these issues as well as the effect of our filtering.

Scally and Durbin presented in 2012 (61) a nice review of the uncertainty in the mutation rate and generation time and its implication for the parameter estimates of population history models. There are two different ways to estimate the mutation rate: one uses fossil estimates of speciation and genetic diversity between species to deduce the mutation rate; the second uses deeply sequenced trios to directly count the number of de novo mutations.

Using fossils dates and genes from human, chimpanzee, gorilla, Old World monkeys and New World monkeys, Takahata and Satta obtained an estimate of the mutation rate of 1.0×10^{-9} per year (67). The estimates obtained from de novo mutations uses recent sequencing technologies and are based on the whole genome. The estimates obtained by those methods are per generations, since they are based on new mutations observed between the parents and infant. The 1KGP Consortium estimated a mutation rate of 1.2×10^{-8} per generation in an European population and of 1.0×10^{-8} per generation in the Yoruba population(44). Recent analysis of 131 trios of the 1KGP sequences in high depth has given a genome average of 1.35×10^{-8} per generation (Adam Auton, personal communication, 9 October 2013).

As for the estimate of the generation time, it is believed that it is closer to 30 years for recent human populations, and that it was closer to 20 years for primates(43). It was suggested to use a generation time between 28 to 30 years for autosomal DNA, and between 25 to 28 years for mitochondrial DNA (16). We have chosen to use a generation time of 28 years.

The data we used come from regions of low recombination and do not cover the whole genome, moreover we needed to remove between 11% to 13% of the SNPs with our filtering procedure. Therefore, we need to verify how these two restrictions affect the mutation rate. To understand the reduction in mutation rates at different steps of our filtering procedure, we looked at the decrease in diversity using a pairwise measure. Starting from the whole genome, we average the diversity of each individual (count the number of heterozygous SNPs per individual), removing regions of the genome larger then 2,000bp with no SNPs. When evaluating the same quantity using our cold regions, we saw a reduction of 32% of the diversity across all the pairs of populations. A de novo mutation rate estimate in these regions is 1.28×10^{-8} per generation with 95% confidence interval $[1.05 \times 10^{-8}; 1.52 \times 10^{-8}]$(Adam Auton, personal communication, 9 October 2013). Therefore, there is no evidence from these data that regions with low recombination have a reduced mutation rate. After filtering, using only the SNPs for which the type of the ancestral allele is known, we have reduced the diversity (compared to whole-genome estimates) by 0.49% to 0.54% for different pairs of populations. For the mutation rate we will show the results using different estimates, but we always apply a correction to μ to account for the loss in diversity due to our filtering.

5.3 Results

We have analysed five populations from the 1KGP, one from Africa (YRI: Yoruba in Ibadan, Nigera), two from Europe (TSI: Tuscany and GRB: British in England and Scotland) and

two from Asia (CHB: Han Chinese in Bejing and JPT: Japanese in Tokyo). We analysed all ten possible pairs. This gave us four different estimates of the out of Africa migration time and four different estimates of the European and Asian divergence time. We also obtained two estimates related to more recent divergence, one for the Chinese and Japanese divergence time, and another one for the Great-Britain and Tuscany divergence time. For each pair, we have run the method using 200 trees per value of T, 20 trees in the MCEM procedure to evaluate the population sizes and we used 1,000 bootstrap resamplings to estimate the confidence intervals for T. We used all the available individuals in each population.

The raw estimates of T, in units of $2N$ generations, and their confidence intervals are presented in Table 5.1. Note that a confidence interval bound in parentheses means that this value is the closest one on that side of the MLE for which the likelihood was estimated, but this value is not included in the confidence interval. Therefore, the MLE is the limit of the confidence interval on this side. The estimates of the time of divergence for specific events are consistent across the populations used. Our estimates for the out of Africa migration range from 0.0655 to 0.0665. For the estimates of the European and Asian divergence we have more variability: they range from 0.0365 to 0.046. Unfortunately, our confidence intervals are probably too narrow again, since most of the time they included only the MLE. Figures 5.1, 5.2, 5.3 and 5.4 present the log-likelihood of T and a closer view around the MLE for all pairs of populations.

In Table 5.2 we have rescaled those estimates using a generation time of 28 years and a mutation rate of $\mu = 1.25 \times 10^{-8}$, which uses our full correction for the lost in diversity in our data. Our estimate for the out of Africa migration is around 80 thousand years ago, and our estimates of the European and Asian divergence range from around 41.5 to 52 thousand years ago. Our estimates of the Japanese/Chinese divergence is around 6.8 thousand years ago, and around 1.1 thousand years ago for the Tuscany/Great Britain divergence. When comparing these results, we see that they make a lot of sense: for

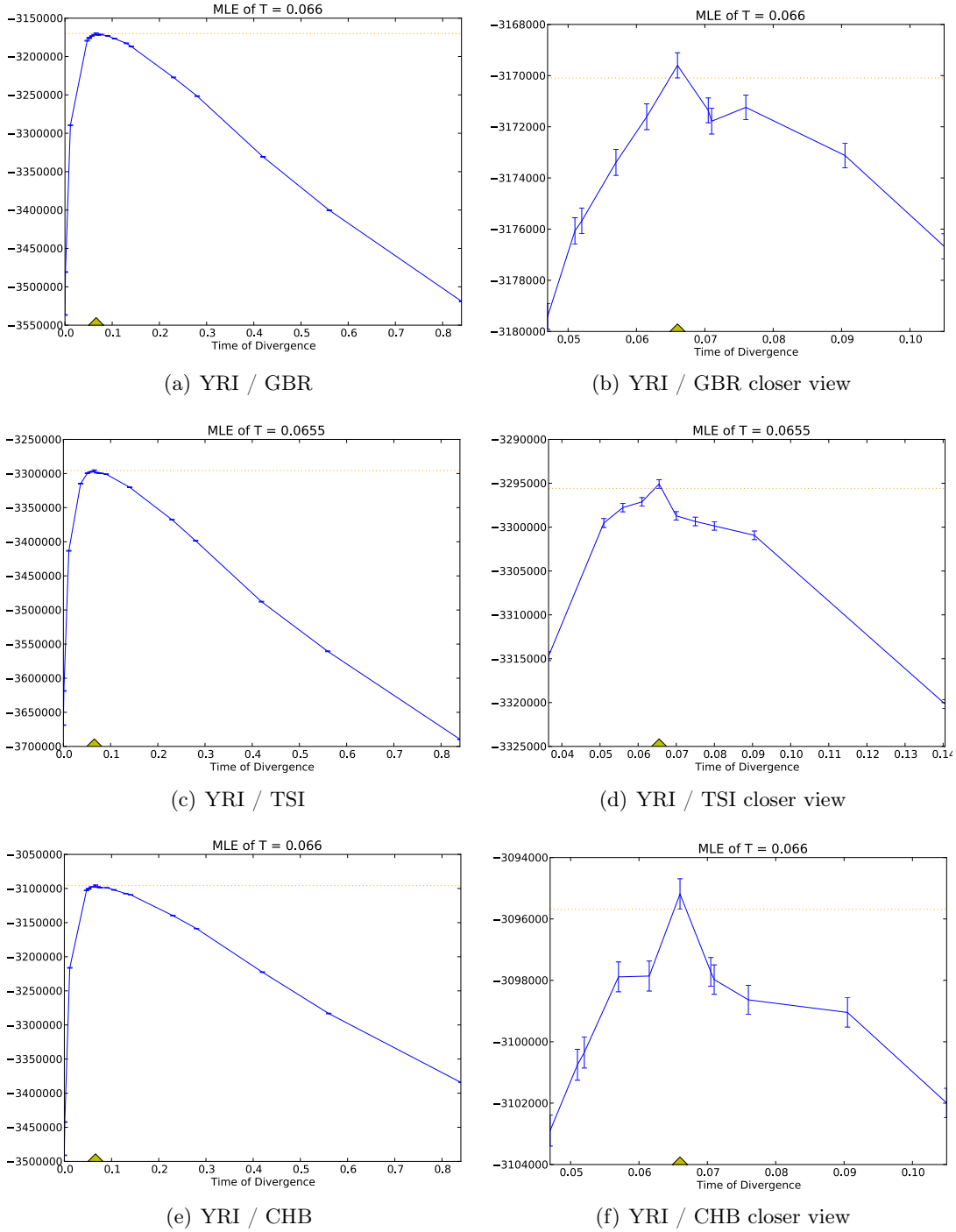
example, we can easily imagine the Tuscany and Great Britain population being much less diverse than Japanese and Chinese populations since the Japanese have been isolated for a longer period from the continent than Great Britain. Furthermore, it makes sense that the oldest divergence is that between the African population and the non-African populations. We will discuss the estimated dates in more detail in the next section.

Populations	$\hat{T}$	C. I.	
YRI / GBR	0.066	(0.0615)	(0.0705)
YRI / TSI	0.0655	(0.061)	(0.07)
YRI / CHB	0.066	(0.0615)	(0.0705)
YRI / JPT	0.0665	0.062	(0.071)
GBR / JPT	0.046	0.0415	(0.0505)
GBR / CHB	0.0365	0.032	0.041
TSI / JPT	0.041	(0.0365)	(0.0455)
TSI / CHB	0.0365	(0.032)	(0.041)
GBR / TSI	0.001	(0.0)	0.0055
CHB / JPT	0.006	(0.0015)	(0.0105)

Table 5.1: Results of the estimation of T, in units of $2N$ generations, for 10 pairs of populations and their confidence intervals. Note that a confidence interval bound in parentheses means that this value is the closest one on that side of the MLE for which the likelihood was estimated, but this value is not included in the confidence interval. Therefore, the MLE is the limit of the confidence interval on this side.

Populations	$\hat{T}$	C. I.	
YRI / GBR	79,741.10	(74,304.21)	(85,177.99)
YRI / TSI	79,825.90	(74,341.68)	(85,310.12)
YRI / CHB	81,017.10	(75,493.20)	(86,540.99)
YRI / JPT	81,666.67	76,140.35	(87,192.98)
GBR / JPT	52,379.02	47,254.98	(57,503.05)
GBR / CHB	41,629.33	36,496.95	46,761.71
TSI / JPT	47,242.80	(42,057.61)	(52,427.98)
TSI / CHB	41,988.50	(36,811.83)	(47,165.16)
GBR / TSI	1,110.89	(0.00)	6,109.90
CHB / JPT	6,804.37	(1,701.09)	(11,907.65)

Table 5.2: Rescaled results of the estimation of T for 10 pairs of populations and their confidence intervals, using $\mu = 1.25 \times 10^{-8}$ and a generation time of 28 years. Note that a confidence interval bound in parentheses means that this value is the closest one on that side of the MLE for which the likelihood was estimated, but this value is not included in the confidence interval. Therefore, the MLE is the limit of the confidence interval on this side.



(a) YRI / GBR

(b) YRI / GBR closer view

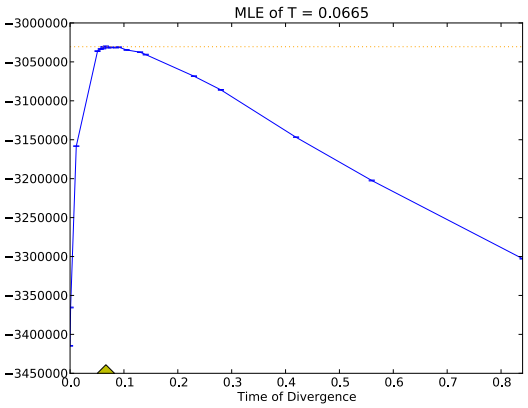
(c) YRI / TSI

(d) YRI / TSI closer view

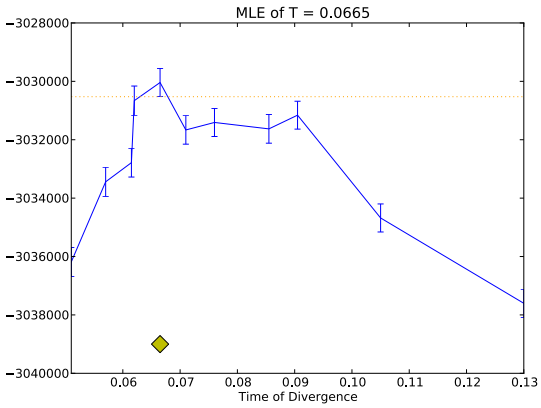
(e) YRI / CHB

(f) YRI / CHB closer view

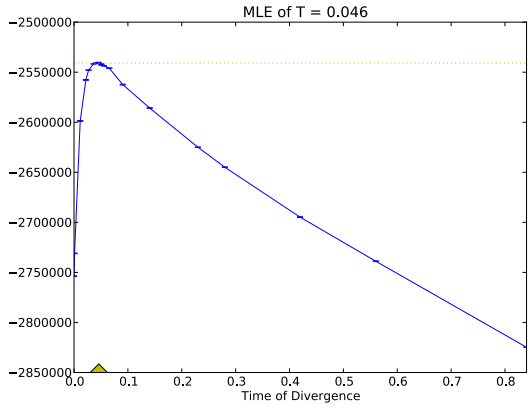
Figure 5.1: Estimates of the log likelihood of T for different pairs of populations. The second column is a closer view of the region near the mode of the log-likelihood.



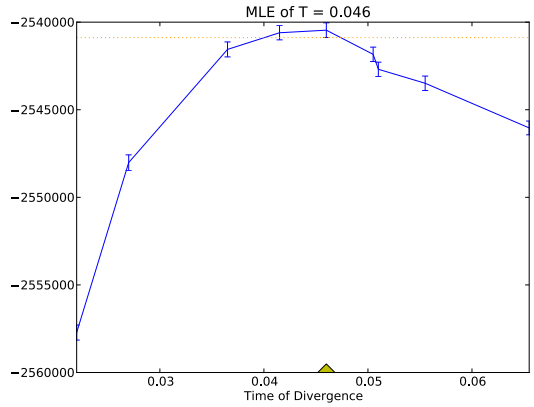
(a) YRI / JPT



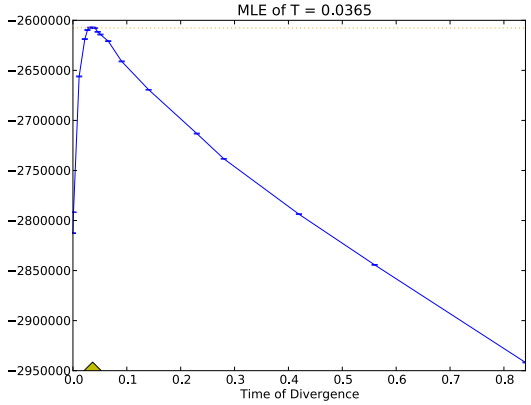
(b) YRI / JPT closer view



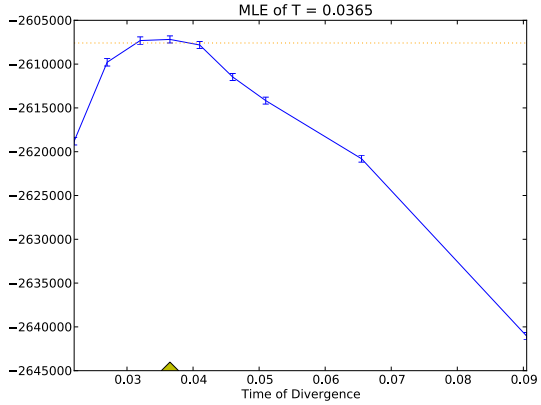
(c) GBR / JPT



(d) GBR / JPT closer view

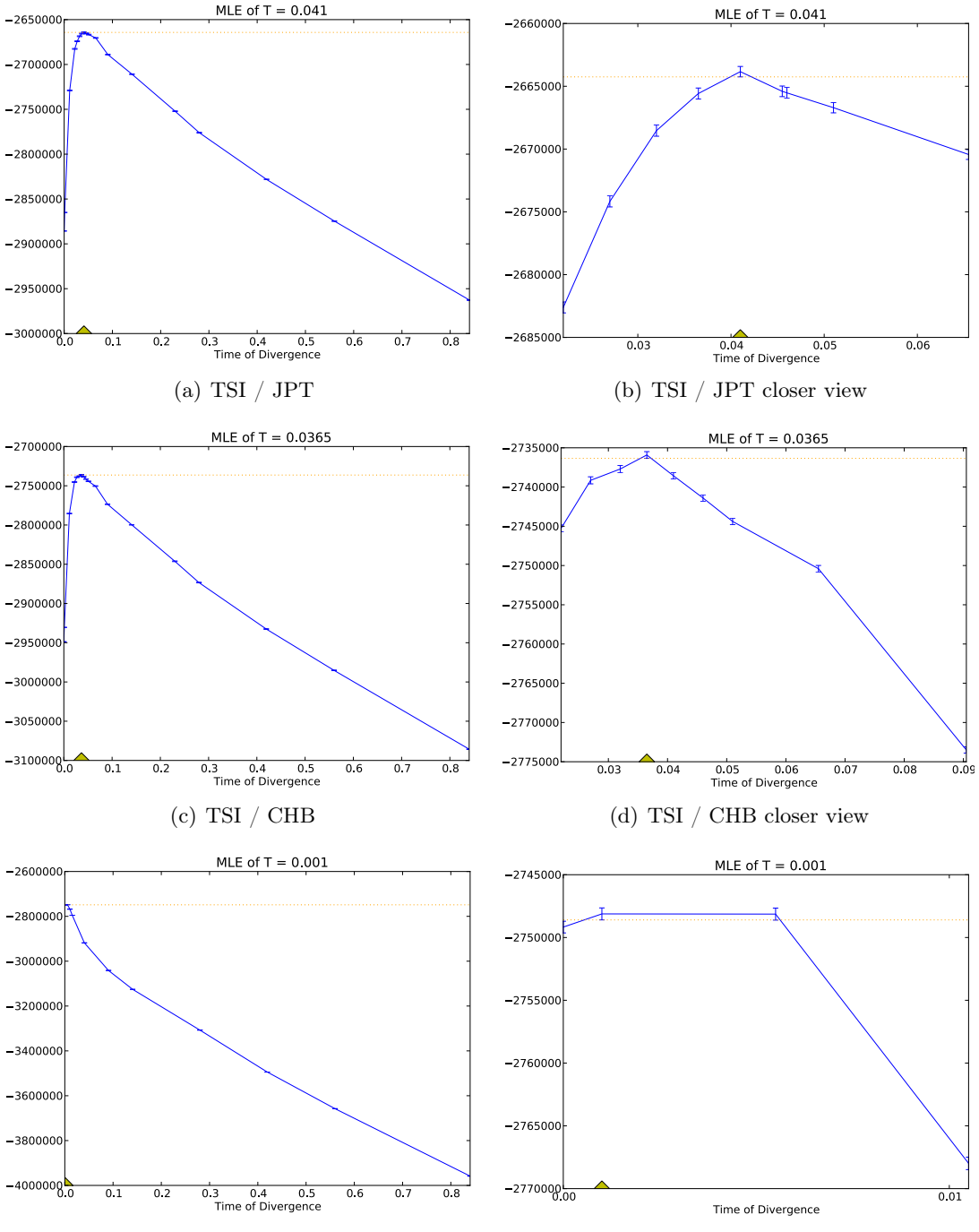


(e) GBR / CHB



(f) GBR / CHB closer view

Figure 5.2: Estimates of the log likelihood of T for different pairs of populations. The second column is a closer view of the region near the mode of the log-likelihood.



(a) TSI / JPT

(b) TSI / JPT closer view

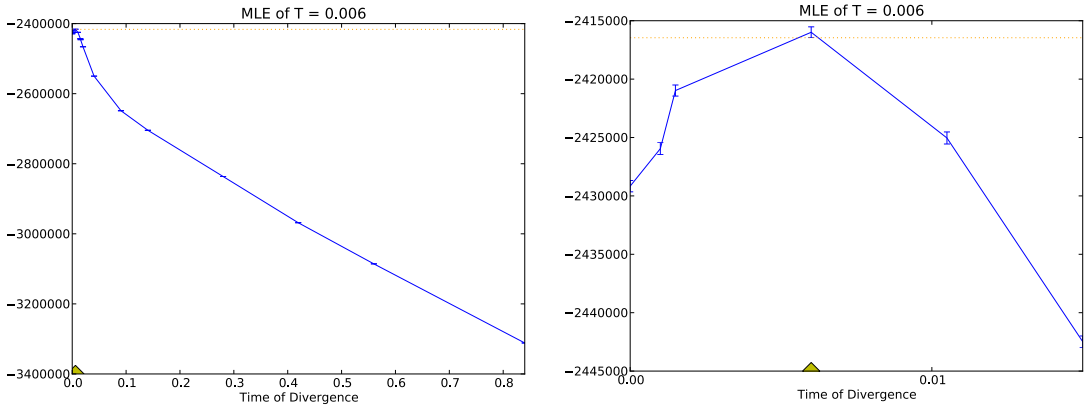
(c) TSI / CHB

(d) TSI / CHB closer view

(e) GBR / TSI

(f) GBR / TSI closer view

Figure 5.3: Estimates of the log likelihood of T for different pairs of populations. The second column is a closer view of the region near the mode of the log-likelihood.



(a) CHB / JPT

(b) CHB / JPT closer view

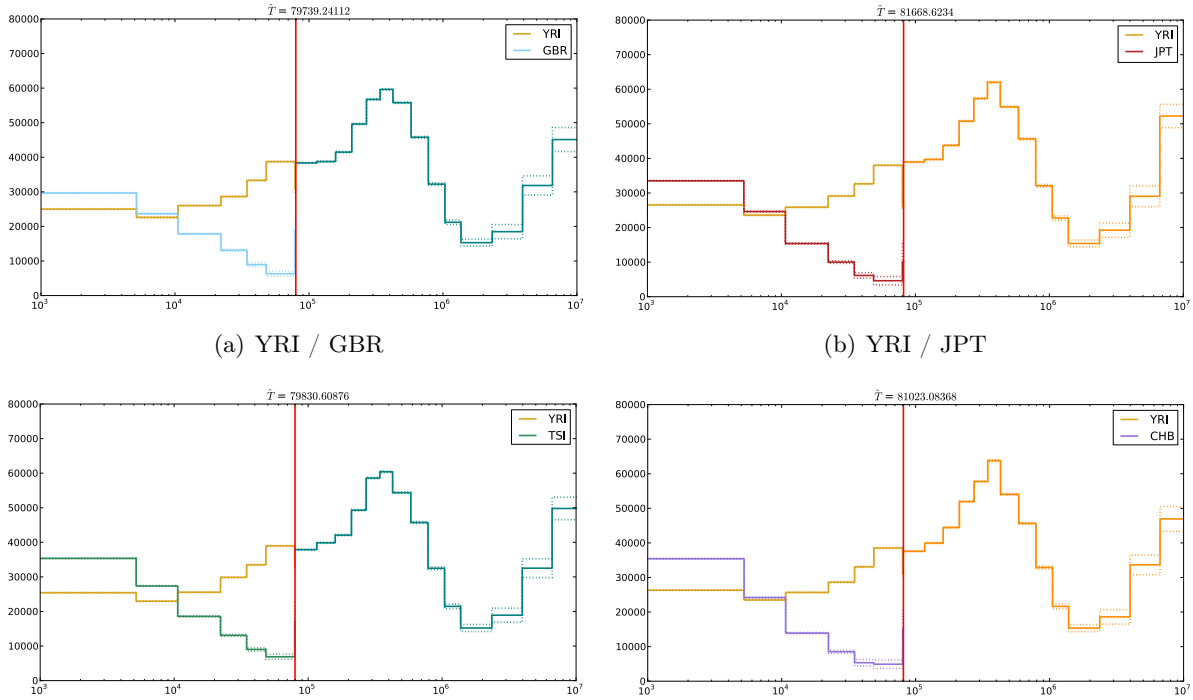
Figure 5.4: Estimates of the log likelihood of T for different pairs of populations. The second column is a closer view of the region near the mode of the log-likelihood.

Table 5.3 presents the estimates of T rescaled using the estimate of μ and the confidence intervals obtained by Adam Auton based on de novo mutations in our regions. We again use a generation time of 28 years, though this time we use our estimates of the loss of diversity related only to the filtering procedure rather than those based on the whole genome.

Populations	$\mu = 1.05 \times 10^{-8}$	$\mu = 1.28 \times 10^{-8}$	$\mu = 1.52 \times 10^{-8}$
YRI / GBR	64,950.46	53,279.67	44,867.09
YRI / TSI	64,927.93	53,261.20	44,851.53
YRI / CHB	65,837.13	54,007.03	45,479.60
YRI / JPT	66,099.84	54,222.53	45,661.08
GBR / JPT	42,220.19	34,633.75	29,165.27
GBR / CHB	33,719.46	27,660.49	23,293.05
TSI / JPT	38,033.16	31,199.07	26,272.90
TSI / CHB	33,960.34	27,858.09	23,459.45
GBR / TSI	900.86	738.99	622.31
CHB / JPT	5,469.23	4,486.48	3,778.09

Table 5.3: Results of the estimation of T for 10 pairs of populations and their confidence intervals, rescaled using three different value of μ as estimated by Adam Auton and using a generation time of 28 years.

Figures 5.5 and 5.6 present our population size estimates using a generation time of 28 years, a mutation rate of $\mu = 1.25 \times 10^{-8}$, and the full correction for the loss in diversity in our data. We again see a good internal consistency of our estimates of the population sizes. After the out of Africa migration, going forward in time, we have that the African population size starts around the size of the ancestral population just before the split, slowly decreases and then stays constant. For all the non-African populations, we see an initially small population sizes that quickly expand to be larger than the African population size in recent times.



(a) YRI / GBR

(b) YRI / JPT

(c) YRI / TSI

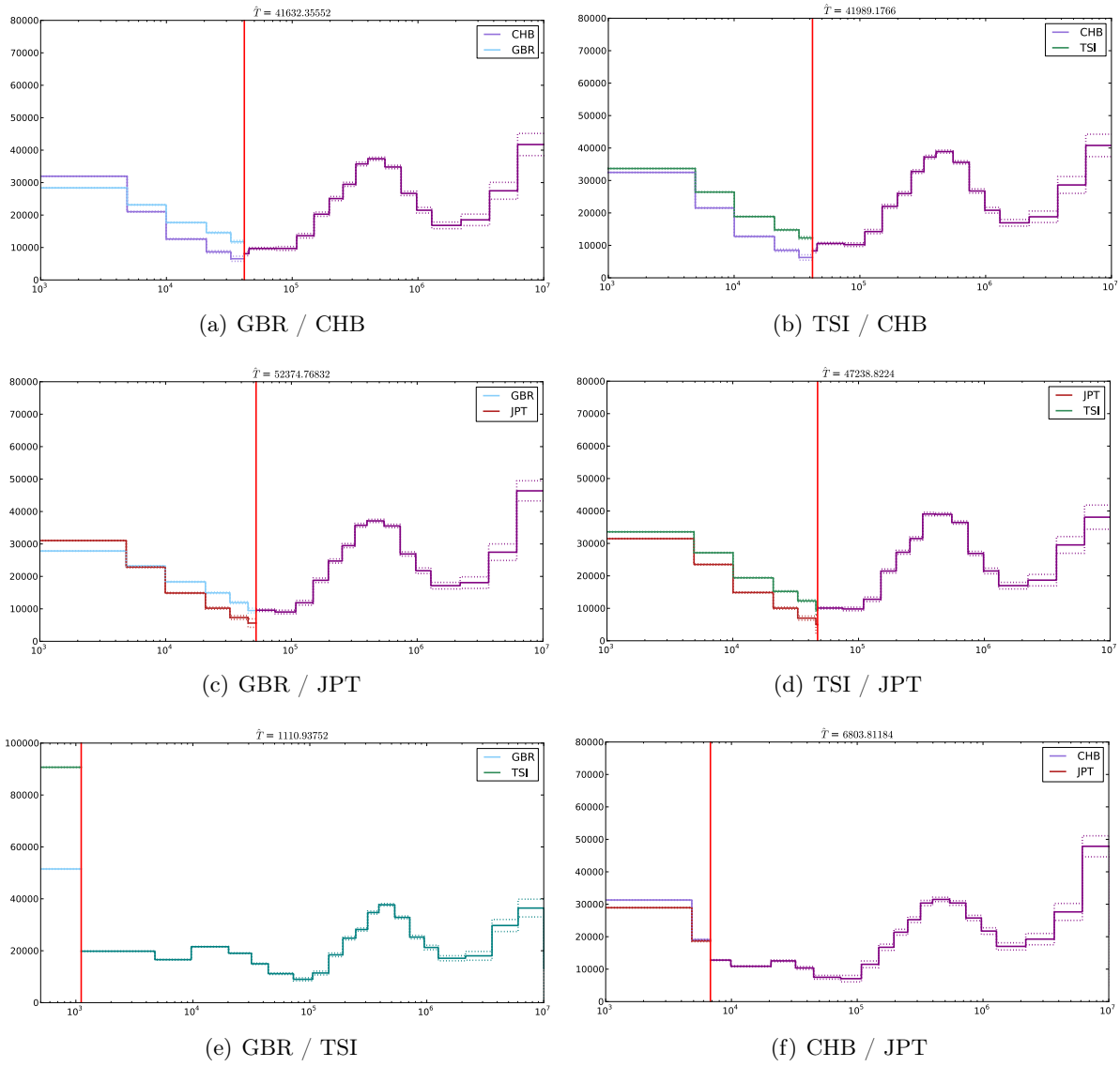
(d) YRI / CHB

Figure 5.5: Estimates of the variable population sizes, the red vertical line represents the position of the estimated time of divergence. The x-axis is the time on the log scale, going back. The y-axis is the estimates the the population sizes.

If we look at the population size estimates for the European/Asian split (Figures 5.6(a), 5.6(b), 5.6(c) and 5.6(d)), we see that the bottleneck clearly occurs before the split (going forwards in time) and that the expansion experienced by all populations starts right after the split. In all the Figures, we see that the ancestral population sizes experienced an expansion, followed by a retraction (going forward in time), with a peak around 400 thousands years ago. The population sizes are largest when the African population is included in the data. This can be explained by a loss of diversity due to our filtering procedure, since we had to remove more SNPs, proportionally, when the Yoruba population was included.

When looking at the two comparisons of closely related population (Figures 5.6(e) and 5.6(f)), it is interesting to see that the expansion following the bottleneck is far more drastic. In fact, in both cases, the population sizes remained mostly constant after the bottleneck and then jump after the split. The population sizes get drastically larger when comparing the two European populations, which could be due to an underestimation of the time of the split. Note that in this Figure 5.6(e) the x-axis covers a large time interval to be able to clearly see the population sizes after the split, since the time of the split is close to the lower limit of the x-axis usually used (1,000 years ago). We can also remark that the conditional confidence intervals for the population sizes are really narrow, even for the ancient ancestral population.

To show the internal consistency of our estimates of the population sizes, we have created two plots (Figure 5.7) where we have superimposed all the population sizes estimates for one population. In Figure 5.7(a), the four pair estimates with the Yoruba population are presented. We see that the four different estimates of the YRI sizes are nearly on top of each other. The estimates of the ancestral population sizes are also really similar across the four different runs. In Figure 5.7(b), we have superimposed the four pairwise estimates with Great Britain. Except for the estimates using the pair GBR/YRI, the different estimates of the ancestral population sizes lined up nicely. The difference for GBR/YRI can be explained by the higher lost in diversity due to the filtering procedure which compen-



(a) GBR / CHB (b) TSI / CHB

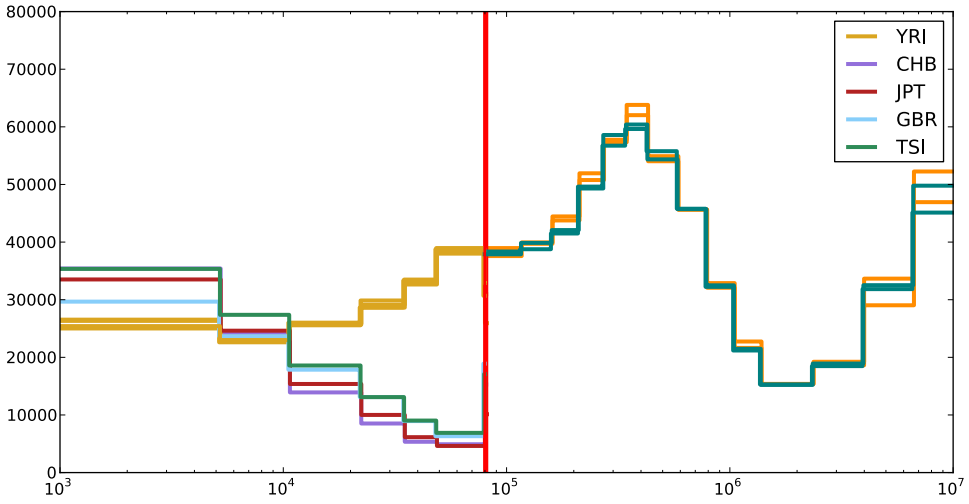
(c) GBR / JPT (d) TSI / JPT

(e) GBR / TSI (f) CHB / JPT

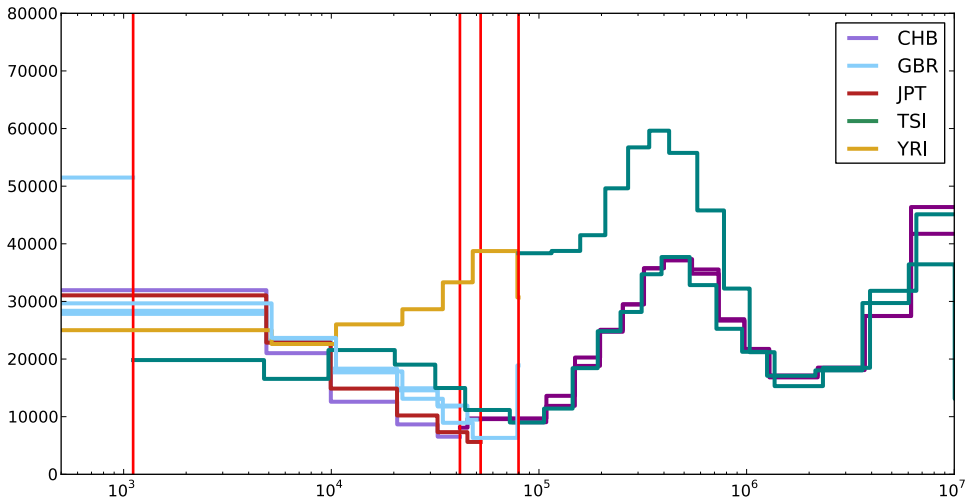
Figure 5.6: Estimates of the variable population sizes, the red vertical line represents the position of the estimated time of divergence. The x-axis is the time on the log scale, going back in time. The y-axis is the estimates the the population sizes.

sate with a larger correction. The estimates of the GBR sizes after the splits are close to each other, between 28,000 and 30,000 for the most recent epoch. Except for the estimates obtained while using TSI for which the population size is higher after the split (around 51,000) and is smaller before the split (around 20,000). A possible explanation, as

previously mentioned, could be an underestimation of the time of divergence. One other possible explanation could be a smoothing effect for the others estimates (others pairs).



(a) YRI *vs.* all others four populations.



(b) GBR *vs.* all others four populations.

Figure 5.7: Estimates of the variable population sizes, where the results of multiples runs are superimposed on each other. The red vertical lines represent the position of the estimated time of divergence. The x-axis is the time on the log scale, going back in time. The y-axis is the estimates the the population sizes.

5.4 Comparison with previous results

In this section, we compare our estimates of the times of divergence of the African/non-African populations and European/Asian population to previous estimates recently obtained with methods presented in the first chapter. These results, as presented in their respective papers, are summarised in Table 5.4. We choose to not rescale these results as we scaled ours. In this table, you have first the main author and reference, then the estimates of the divergence of an African population with a non-African population, followed by the estimates of the time of divergence of a European and an Asian population, when available. The next columns indicate if the model used included an expansion for the descendant population sizes and if it included a bottleneck in the European population. Finally, the last two columns indicate the type of data used and the method.

Most of the other results are based on summaries statistics methods. Those methods compared different models, including migrations, exponential expansions or a bottleneck event. They then simulated datasets under these models and compared the summary statics of the simulated data to the summary statistics of the real data. When the population sizes were not assumed to be constant, all methods used an exponential function to model a variable population size, except PSMC, which models the population sizes as piecewise constant.

In their paper (38), Li and Durbin used their estimates of the variable population sizes to infer the time at which different populations diverged. Arguing that when the sizes of two populations are no longer equal (or close enough), they have started to diverge. Combining two haplotypes from two different populations, they also argue that when the population sizes of this fake population goes to infinity, the two populations are in fact in isolation. Therefore, the authors infer that there must have been migrations between the African and non-African populations, since their population sizes start to differ earlier in the past than the time at which the fake population goes to infinity. PSMC does not

directly estimate the time of divergence, but we will use the range of possible times of divergence mentioned in the paper.

Nearly all methods used a generation time of 25 years, except the analysis of Garrigan (17) which used a generation time of 20 years with data from mitochondrial DNA and the Y and X chromosomes. Cox (8) used the IM method, as did Garrigan, but used a generation time of 28 years with genomic DNA data. All the methods used either a mutation rate of 2.5×10^{-8} per generation or assumed a chimp/human divergence time of 6 millions years ago to estimate the mutation rate, except Gronau (21), who assumed a chimp/human divergence time of 6.5 million years ago.

Overall, the estimates of the time of divergence of an African population and a non-African population range from around 40 thousand years ago to 140 thousand years ago, but are mostly around 50 to 60 thousand years ago. The estimate of 140 thousand years ago obtained by Gutenkunst (2009)(22) with ∂a∂i can be explained in part by the strong migration band inferred in their model. On the other end of the spectrum, the 40 thousand years ago estimate was obtained assuming a generation time of 20 years, which might not be sufficiently large. The estimates for the European/Asian divergence are more similar and around 23 thousands years ago, ranging from 22.5 to 30 thousand years ago, if we exclude the range 26 to 47 thousand years ago estimated with PSMC.

To compare our results, we can roughly double the estimates in Table 5.4 to rescale to the mutation rate we used. We can see that our estimates agree with the lower bound of the range of estimates of the other studies. In comparison, our estimates of the time of the out of Africa migration are a bit lower than the average, and around the average for the European/Asian divergence time. A possible explanation is that, unlike most other methods, we do not include migration events between the populations in our model.

Reference	T^a	T^b	Exp.	Bottl.	Data	Method
Fagundes (2007)(14)	51		✓	✓	50 loci of 50bp and 20 ind.	summ. stat. ABC
Garrigan (2007)(17)	$39.5^{\dagger}$	$25^{\dagger}$	✓		mDNA, X and Y 16.2kb 400 ind.	IM
Cox (2008)(8)	58^{*}	30^{*}	✓		98kb (20 loci), 90 ind.	IM
Gutenkunst (2009)(22)	140	23	✓	✓	5Mb, 68 ind.	∂a∂i
Wall (2009)(71)	120 / 80		✓	✓	5.3Mb, 58 ind.	summ. stat.
Laval (2010)(37)	60	22.5	✓		20 loci 6Mb, 213 ind.	summ. stat. ABC
Gravel (2011)(18)	51	23	✓	✓	1KGP Pilot	∂a∂i
Li and Durbin (2011)(38)	(60-80)		✓	✓	whole genome, 7 ind.	PSMC
Gronau (2011)(21)	$47(38-64)^{**}$	$(26-47)^{**}$			37,574 loci of 1kb, 6 ind.	G-Phocs

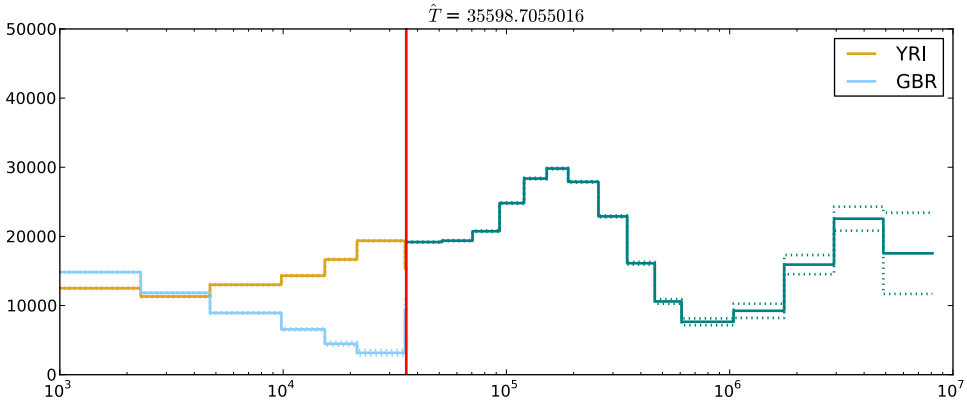
Table 5.4: A comparison of different estimates of the time of the out of Africa migration (T^a in thousands of years ago) and the time of divergence of European and Asian populations (T^b in thousands of years ago). All the time estimates are scaled using a generation time of 25 years (20 years for $\dagger$, 28 years for $*$), and assumed a mutation rate around 2.5×10^{-8} or a chimp/human divergence time of 6 millions years ago (6.5 for $**$). All studies, except Gronau, considered an exponential expansion of the sizes of at least one of the population sampled (or a freely variable population size for Li and Durbin). Some used a model with a bottleneck. Note that all studies modelled migration events or inferred migrations from their results (Li and Durbin).

Our method is the only one, other than PSMC, that estimates variable population sizes without making assumptions (other than piecewise constant). To compare our results with PSMC, we have rescaled our estimate of the population sizes for populations YRI and GBR using a generation time of 25 years and a mutation rate of 2.5×10^{-8} (using our correction for the loss of diversity). Figure 5.8 presents both our estimates (Figure 5.8(a)) and the results with PSMC (Figure 5.8(b) from the Li and Durbin paper (38)). We see that the estimates beyond 100 thousand years ago are similar in shape, peaking at around 150 thousand years ago. However, our estimates of the population sizes are roughly twice as large as Li and Durbin's estimates.

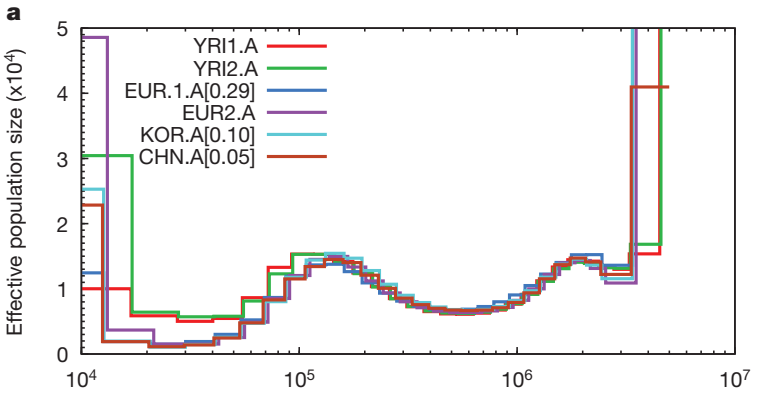
The main difference between the estimates of the methods is that our method does not infer a smooth decrease in population sizes around 100 thousand years ago. We see the beginning of this increase but then, going forwards in time, our estimate of the population size is more constant. After the split the the size of GBR is quite small (the bottleneck) while the YRI size decreases close to 10,000 and then go up again to 12,000. This is similar, but more spread in time, to the estimates of the population size of YRI from Li and Durbin, except that our population sizes are twice as big. Another remark is that our method gives more accurate estimates for recent times, starting around 1,000 years ago, compared to around 20,000 years ago for PSMC.

5.5 Discussion

In this thesis we have presented a novel method to estimate both the time of divergence of populations and their variable sizes using multiple cold sequenced regions of the genome. Understanding when different populations arose and how their sizes have changed through time are central questions about populations histories. Moreover, genomic data is probably our only way to learn about the size changes a population might have experienced in the past. This can also have an impact on results of genetic analyses that assume a constant



(a) YRI / GBR



(b) YRI / JPT

Figure 5.8: Comparison of: (a) our estimates of the population sizes for YRI and GBR and (b) the estimates presented by Li and Durbin(38). The axis on both Figures are on the same scale.

population size. Many methods have been developed over the years to shed light on these questions. Due to the complexity of the problem, every method either simplifies the model with strong assumptions or uses an approximation. Moreover, they need to down sample their data either using less individuals or only portions of the genome.

Our method is the only one, to our knowledge, that estimates both the time of divergence and variable population sizes. Our assumptions are that no migrations occurred between the descendant populations and that we are able to find regions of the genome with approximately no recombination. We have shown through many simulation studies

that the method performs well in many situations. Overestimating the time of divergence only occurs when there is a strong bottleneck, and population size over-smoothing only occurs for drastic, instantaneous changes. Admixture events have a small effect on our estimates of the population sizes, resulting in a slight underestimation of the ancient ancestral population sizes and the appearance of a bump in the estimates when the populations are more divergent. Migrations, as one might expect, have a stronger effect on our estimates, mainly when migrants are exchanged for long time intervals at high rates. In these situations, the estimated time of divergence will be between the real split time and the time at which the populations become isolated (when going forward in time).

We have also explained the necessity of filtering the real data to ensure that the remaining regions pass the three gametes test. We have tested the effect of filtering the data when recombination was included in our simulated data. The only effect was the underestimation of the population sizes for more ancient times, around a million years ago. For real datasets, we also needed to know the type of the ancestral allele. This assumption could have been easily avoided by modifying our importance sampler, which we could add to our method and test to see if it has an effect on our estimates.

We applied the method to five different populations of the 1000 Genomes Project using one African, two Asian and two European populations. The main challenge with real data is how to rescale the estimates into years. This is related to the uncertainty in the mutation rate. The results we obtained are comparable to some of the previous published results. We estimate the time of the out of Africa migration to be around 80 thousand years ago, which agrees with fossils records but is more recent than some of the previous estimates. Our estimates of the time of the European/Asian split are around 47 thousand years ago which is close to previous estimates if they are rescaled accordingly. Our estimates of the variable population sizes demonstrate the importance of avoiding strong constraints when modelling population sizes, since they tend to both expand and decrease. Therefore, allowing the population size to vary freely between each epoch of time might be a good strategy. We

made direct comparisons between our estimates and Li and Durbin's estimates by adopting their scaling. Both methods gave similar estimates of how the populations have changed through time, though our method has better resolution for more recent times.

There is still work to be done; we need to find better confidence intervals for our estimates. We are clearly underestimating the standard deviation of our likelihood estimates. A better bootstrap method might help. We also plan to extend our model to include the possibility of a migration band to allow a smoother split instead of the instantaneous split we are currently modelling.

Bibliography

[1] Alves, I., Hanulová, A. Š., Foll, M., and Excoffier, L. Genomic Data Reveal a Complex Making of Humans. *PLoS genetics 8*, 7 (July 2012), e1002837. 127

[2] Amyot, D. Pyarg. `https://launchpad.net/pyarg/`, 2011. 56

[3] Armitage, S. J., Jasim, S. A., Marks, A. E., Parker, A. G., Usik, V. I., and Uerpmann, H. P. The Southern Route "Out of Africa": Evidence for an Early Expansion of Modern Humans into Arabia. *Science (New York, NY) 331*, 6016 (Jan. 2011), 453–456. 140

[4] Beaumont, M. A. Joint determination of topology, divergence time and immigration in population trees. In *Simulations, Genetics and Human Prehistory*, S. Matsura, Ed. McDonald Institute for Archaeological Research, Cambridge, 2008, pp. 135–154. 42

[5] Becquet, C., and Przeworski, M. A new approach to estimate parameters of speciation models with application to apes. *Genome Research 17* (2007), 1505–1519. 40, 43, 46

[6] Becquet, C., and Przeworski, M. Learning about modes of speciation by computational approaches. *Evolution 63*, 10 (2009), 2547–2562. 41

[7] Collins, F. S., and Mansoura, M. K. The Human Genome Project. Revealing the shared inheritance of all humankind. *Cancer 91*, 1 Suppl (2001), 221–225. 4

[8] Cox, M. P., Woerner, A. E., Wall, J. D., and Hammer, M. F. Intergenic DNA sequences from the human X chromosome reveal high rates of global gene flow. *BMC genetics 9*, 1 (2008), 76. 30, 156, 157

[9] Davison, D., Pritchard, J., and Coop, G. An approximate likelihood for genetic data under a model with recombination and population splitting. *Theoretical population biology* (2009). 34, 46

[10] Delaneau, O., Zagury, J.-F., and Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods 10*, 1 (Jan. 2013), 5–6. 140

[11] Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B. Methodological 39*, 1 (1977), 1–38. 89

[12] Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and ...* (2012). 31, 46

[13] Ethier, S., and Griffiths, R. The infinitely-many-sites model as a measure-valued diffusion. *The Annals of Probability 15*, 2 (1987), 515–545. 16, 19, 23

[14] Fagundes, N. J. R., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F. M., Bonatto, S. L., and Excoffier, L. Statistical evaluation of alternative models of human evolution. *PNAS 104*, 45 (2007), 17614–17619. 157

[15] Fearnhead, P., and Donnelly, P. Estimating recombination rates from population genetic data. *Genetics 159*, 3 (Nov. 2001), 1299–1318. 33

[16] Fenner, J. N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *American Journal of Physical Anthropology 128*, 2 (Oct. 2005), 415–423. 143

[17] Garrigan, D., Kingan, S. B., Pilkington, M. M., Wilder, J. A., Cox, M. P., Soodyall, H., Strassmann, B., Destro-Bisol, G., de Knijff, P., Novelletto, A., Friedlaender, J., and Hammer, M. F. Inferring Human Population Sizes, Divergence Times and Rates of Gene Flow From Mitochondrial, X and Y Chromosome Resequencing Data. *Genetics 177*, 4 (Dec. 2007), 2195–2207. 30, 156, 157

[18] Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., Gibbs, R. A., The 1000 Genomes Project Consortium, and Bustamante, C. D. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences 108* (2011), 11987. 157

[19] Griffiths, R., and Tavare, S. Sampling Theory for Neutral Alleles in a Varying Environment. *Philosophical Transactions: Biological Sciences 344*, 1310 (June 1994), 403–410. 14, 57, 58, 63

[20] Griffiths, R. C., and Tavaré, S. Simulating probability distributions in the coalescent. *Theoretical population biology 46* (1994), 131–159. 19, 23, 28, 45

[21] Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., and Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics 43*, 10 (Oct. 2011), 1031–1034. 32, 46, 156, 157

[22] Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS genetics 5*, 10 (2009), 1–11. 42, 46, 156, 157

[23] Hein, J., H Schierup, M., and Wiuf, C. *Gene Genealogies, Variation and Evolution*. Oxford University Press, June 2005. 7, 13, 14

[24] Heled, J., and Drummond, A. J. Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology 8*, 1 (2008), 289. 30, 46

[25] Hey, J. On the Number of New World Founders: A Population Genetic Portrait of the Peopling of the Americas. *PLoS Biology 3*, 6 (May 2005), e193. 30

[26] Hey, J. Isolation with Migration Models for More Than Two Populations. *Molecular biology and evolution 27*, 4 (Mar. 2010), 905–920. 30, 46

[27] Hey, J., and Nielsen, R. Multilocus Methods for Estimating Population Sizes, Migration Rates and Divergence Time, With Applications to the Divergence of Drosophila pseudoobscura and D. persimilis. *Genetics 167*, 2 (June 2004), 747. 28, 46, 55

[28] Hey, J., and Nielsen, R. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences of the United States of America 104*, 8 (2007), 2785. 28, 46

[29] HO, S. Y. W., and Shapiro, B. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Molecular Ecology Resources 11*, 3 (Feb. 2011), 423–434. 30

[30] Hoffecker, J. F., Holliday, V. T., and Anikovich, M. V. From the Bay of Naples to the River Don: the Campanian Ignimbrite eruption and the Middle to Upper Paleolithic transition in Eastern Europe. *Journal of Human Evolution 55* (Nov. 2008), 858–870. 139

[31] Howie, B., and Marchini, J. Impute. `http://mathgen.stats.ox.ac.uk/impute/`, 2013. 140

[32] Hudson, R. R. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology 23*, 2 (Apr. 1983), 183–201. 7, 14

[33] Hudson, R. R. Gene genealogies and the coalescent. *Evolutionary Biology 7* (July 1991), 1–44. 7

[34] Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics (Oxford, England) 18*, 2 (Feb. 2002), 337–338. 63, 65

[35] Hunter, D. R., and Lange, K. A Tutorial on MM Algorithms. *The American Statistician 58*, 1 (2004), 30–37. 93

[36] Kingman, J. F. C. The Coalescent. *Stochastic Processes and their Applications 13* (Apr. 1982), 235–248. 7

[37] Laval, G., Patin, E., and Barreiro, L. Formulating a Historical and Demographic Model of Recent Human Evolution Based on Resequencing Data from Noncoding Regions. *PLoS ONE* (2010). 157

[38] Li, H., and Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature 475*, 7357 (July 2011), 493–496. 37, 46, 94, 155, 157, 158, 159

[39] Li, N., and Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics 165*, 4 (Dec. 2003), 2213–2233. 33, 35

[40] Lopes, J. S., Balding, D., and Beaumont, M. A. PopABC: a program to infer historical demographic parameters. *Bioinformatics (Oxford, England) 25*, 20 (2009), 2747–2749. 41, 46

[41] Mailund, T., Dutheil, J. Y., Hobolth, A., Lunter, G., and Schierup, M. H. Estimating Divergence Time and Ancestral Effective Population Size of Bornean and

Sumatran Orangutan Subspecies Using a Coalescent Hidden Markov Model. *PLoS genetics 7*, 3 (Mar. 2011), e1001319. 16, 35, 46

[42] Mailund, T., Halager, A. E., Westergaard, M., and Dutheil, J. Y. A New Isolation with Migration Model along Complete Genomes Infers Very Different Divergence Processes among Closely Related Great Ape Species. *PLoS genetics 8*, 12 (2012). 36

[43] Matsumura, S., and Forster, P. Generation time and effective population size in Polar Eskimos. *Proceedings. Biological sciences / The Royal Society* (Mar. 2008). 143

[44] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature 467*, 7319 (Oct. 2010), 1061–1073. 139, 142

[45] McDougall, I., Brown, F. H., and Fleagle, J. G. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature 433*, 7027 (Feb. 2005), 733–736. 139

[46] McLachlan, G., and Krishnan, T. *The EM algorithm and extensions*, second ed. Wiley series in probability and statistics. Wiley, Hoboken, NJ, 2008. 89

[47] McVean, G. A. T., and Cardin, N. J. Approximating the coalescent with recombination. *Philosophical transactions of the Royal Society of London Series B, Biological sciences 360*, 1459 (July 2005), 1387–1393. 16, 34

[48] Millard, A. R. A critique of the chronometric evidence for hominid fossils: I. Africa and the Near East 500–50ka. *Journal of Human Evolution 54* (June 2008), 848–874. 139

[49] Motoo Kimura, G. H. W. The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. *Genetics 49*, 4 (Apr. 1964), 561. 14

[50] Nielsen, R. Maximum Likelihood Estimation of Population Divergence Times and Population Phylogenies under the Infinite Sites Model. *Theoretical population biology 53* (1998), 143–151. 28, 46

[51] Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics 12*, 6 (May 2011), 443–451. 127

[52] Nielsen, R., and Wakeley, J. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics 158*, 2 (June 2001), 885–896. 28

[53] Nordborg, M. Coalescent theory. In *Handbook of Statistical Genetics*, M. J. Bishop and C. Cannings, Eds. John Willey & Sons, Inc., Feb. 2001, pp. 179–212. 7, 13, 14

[54] Oakes, D. Direct calculation of the information matrix via the EM. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 61*, 2 (Apr. 1999), 479–482. 93

[55] O'Connell, J. F., and Allen, J. Dating the colonization of Sahul (Pleistocene Australia–New Guinea): a review of recent research. *Journal of Archaeological Science 31*, 6 (June 2004), 835–853. 139

[56] Paul, J. S., and Song, Y. S. A principled approach to deriving approximate conditional sampling distributions in population genetics models with recombination. *Genetics 186*, 1 (Sept. 2010), 321–338. 38

[57] Paul, J. S., Steinrucken, M., and Song, Y. S. An Accurate Sequentially Markov Conditional Sampling Distribution for the Coalescent With Recombination. *Genetics 187*, 4 (Apr. 2011), 1115–1128. 38

[58] Peters, J. L., Roberts, T. E., Winker, K., and McCracken, K. G. Heterogeneity in Genetic Diversity among Non-Coding Loci Fails to Fit Neutral Coalescent Models of Population History. *PLoS ONE* (2012). 30

[59] Petraglia, M., Korisettar, R., Boivin, N., and Clarkson, C. Middle Paleolithic Assemblages from the Indian Subcontinent Before and After the Toba Super-Eruption. *Science (New York, NY) 317* (July 2007), 114–116. 139

[60] Rannala, B., and Yang, Z. Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using DNA Sequences From Multiple Loci. *Genetics 164*, 4 (Aug. 2003), 1645. 32

[61] Scally, A., and Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics 13*, 10 (Sept. 2012), 745–753. 56, 142

[62] Sheehan, S., Harris, K., and Song, Y. S. Estimating Variable Effective Population Sizes From Multiple Genomes: A Sequentially Markov Conditional Sampling Distribution Approach. *Genetics* (2013). 38, 46

[63] Steinrücken, M., Paul, J. S., and Song, Y. S. A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. *arXiv.org* (Aug. 2012). 38

[64] Stephens, M., and Donnelly, P. Inference in Molecular Population Genetics. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 62*, 4 (2000), 605–655. 1, 19, 23, 25, 33, 45, 47, 58, 63

[65] Strasburg, J. L., and Rieseberg, L. H. How Robust Are "Isolation with Migration" Analyses to Violations of the IM Model? A Simulation Study. *Molecular biology and evolution 27*, 2 (Jan. 2010), 297–310. 30

[66] Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics 105*, 2 (Oct. 1983), 437–460. 7

[67] Takahata, N., and Satta, Y. Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from dna sequences. *Proceedings of the National Academy of Sciences 94*, 9 (1997), 4811–4815. 142

[68] Tellier, A., Pfaffelhuber, P., Haubold, B., Naduvilezhath, L., Rose, L. E., Städler, T., Stephan, W., and Metzler, D. Estimating Parameters of Speciation Models Based on Refined Summaries of the Joint Site-Frequency Spectrum. *PLoS ONE 6*, 5 (May 2011), e18155. 43, 46

[69] Wakeley, J. *Coalescent theory: An introduction.* Roberts & Company Publishers, 2009. 7, 13

[70] Wakeley, J., and Hey, J. Estimating Ancestral Population Parameters. *Genetics 145* (1997), 847–855. 39, 43, 46

[71] Wall, J. D., Lohmueller, K. E., and Plagnol, V. Detecting Ancient Admixture and Estimating Demographic Parameters in Multiple Human Populations. *Molecular biology and evolution 26*, 8 (July 2009), 1823–1827. 157

[72] Wang, Y., and Hey, J. Estimating Divergence Parameters With Small Samples From a Large Number of Loci. *Genetics 184*, 2 (Feb. 2010), 363–379. 31, 46

[73] Wei, G. C. G., and Tanner, M. A. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association 85*, 411 (Sept. 1990), 699–704. 90

[74] Wiuf, C., and Hein, J. Recombination as a point process along sequences. *Theoretical Population Biology 55*, 3 (June 1999), 248–259. 16, 33