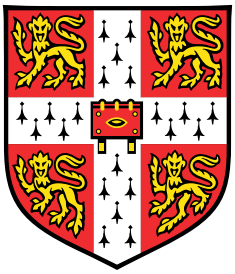
Inferring Genomic Histories of Structured Populations: Lessons from the Hominids



Tariq Desai

Supervisor: Aylwyn Scally

Advisor: Frank Jiggins

Department of Genetics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Magdalene College

September 2019

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee

Tariq Desai
September 2019

Abstract

Inferring Genomic Histories of Structured Populations: Lessons from the Hominids

Tariq Desai

Demography is a major determinant of the variation of species. Understanding how it effects the genetic composition of populations allows us to infer history from genetic data. I contribute to this project in three ways. First, I estimate the historical trajectories of effective population size in *Pan* and *Pongo*. These analyses are based on past rates of coalescence, estimated within and between populations, and they use the largest known samples of whole genomes for each species. From these, I infer new histories of migration and separation in each genus.

Second, the interpretation of past rates of coalescence is known to depend on whether a population is structured. I analyse a novel model of non-equilibrium island structure and propose an approach through which it can be used to aid the interpretation of estimated historical changes in effective population size. This is based on comparisons of the effects on theoretical distributions of coalescent times. The approach is then applied to examples arising in chimpanzee and human population history, in both cases suggesting effective population size changes are more likely due to changes in census population size, rather than island structure.

The proliferation of ancient DNA sequencing presents a rich new resource. I develop a haplotype-based approach to help test more directly for ancestral structure. It is based on reconstructions of the ancestral recombination graph of a sample. The method incorporates an explicit coalescent analysis of locus ancestry under idealised demographic models. It jointly analyses samples of ancient and modern whole genomes in which modern sequences trace some recent part of their ancestry to populations from which the ancient sequences are drawn. I illustrate some of the possibilities and limitations of this tentative approach and apply it to questions about the peopling of North America.

To my mother, always

Faieza Desai
1959 - 2018

Acknowledgements

I should first thank my supervisor, Aylwyn Scally. This research would not have happened had Aylwyn not taken the chance and invited me to join his group. I suspect he will always stand in my mind as an exemplar of scientific independence and scepticism, not least for his routine reminders of our collective ignorance! I may have needed more guidance and patience from him than he was expecting, but his support all the way through was unwavering. Likewise, I must thank my advisor, Frank Jiggins, for his overseeing this process, and for allowing me to be something of a brood parasite among his lab members on the first floor.

I am thankful to the Genetics Department for providing a stimulating and supportive environment. I learnt a great amount in my time there, especially from the evolutionary genetics groups in the Department and elsewhere in the University. Our weekly journal clubs and joint lab meetings were invaluable. For this, I thank Aylwyn, Frank, Chris Illingworth, John Welch, Chris Jiggins, Toomas Kivisild, Richard Durbin, as well as members of their groups who took part in those sessions. I must also thank Joel, Julien, Dany, Sam, Arun, Lucia, Jon, Anood, Ruoyun, and Alex, for stimulating conversation, companionship, and sometimes even a little biology.

It is virtually impossible to do research today without being part of a large network of collaborators. I owe a great deal to those with whom I collaborated on various projects and extend my gratitude to each of them. For stimulating discussion, and for leading demanding and rigorous projects, I would especially like to thank Maja Mattle-Greminger, Alexander Nater, Michael Krutzen, Marc de Manuel, Martin Kuhlwilm, Peter Frandsen, Tomas Marques-Bonet, Freddi Scheib, and Toomas Kivisild.

I am grateful to the Gates Cambridge Trust for funding my studies and for bringing me into the Gates community. For memories of snow on First Court, candlelit carols, and daffodils along the Cam, I am glad to have been a part of Magdalene College.

I am infinitely grateful to my parents: to my father, who taught me the value of words and sentences, and who provided our home with purpose and security; and to my mother, who gave everything for us, and whom I miss, and will miss, every day. I hope to live up to their love and sacrifices.

To Aniqah, thanks for making sure things never got too quiet around the house, and for making me feel guilty about not exercising. To Azhar, thanks for keeping the ship afloat while I was away. I admire and love you both.

To the Khan family, thanks for being my home away from home, and to Aunty Saabira for being our rock in trying times. It all started in Salt River, so to Aunty Abie, Aunty Bieda, and everyone at Kingsley Road, thank you for the encouragement, and the constant inspiration.

To the friends who supported me through it all, I could hardly have asked for more. I won't be able to name everyone, but to Amandla and Arif, and especially to Greg and Taskeen, I cannot imagine what Cambridge would have been like without you. And to Vinayak, Tara, and Laura, I couldn't hope to show you how much I've drawn from our friendship these last few years. I look forward to the years ahead.

Finally, to Mahvish: I agree. A better world isn't going to build itself, and I hope we'll get the chance to do our bit, together.

Contents

Chapter 1

Introduction

1.1 Population history and hominid genomics

Demography is a major determinant of the genetic variation of species. Changes in population size, as well as migration and admixture, interact with fundamental evolutionary processes to alter a population's genetic composition [17]. Although these effects can be complex, there is considerable interest in trying to understand how they shape observable variation, since doing so would enable us to draw inferences about population history from genetic data. This has been a longstanding project in population genetics and it has met with some success. As least since the pioneering work of Cavalli-Sforza [15], and increasingly since the development of high-throughput and ancient DNA sequencing [81, 104], genetic methods have become significant tools in history-telling (see Section 1.1.1).

The growing influence of population genetics is not only a consequence of rapid advances in data collection. It has been supported by significant developments in theory, statistical techniques, and computing. With regards to theory, what may loosely be thought of as the classical tradition of population genetics, beginning with the foundational work of Fisher, Haldane, and Wright, retains considerable importance in historical inference [122, 18]. However, it has been complemented by several more recent advances, perhaps most significantly by the development of the Kingman coalescent [62], and its extensions [e.g. 180]. The empirical value of the coalescent rests to a large degree on its computational efficiency over previous approaches [171]. This underpins the widespread adoption of several efficient tools for simulation and inference [e.g. 51, 70]. Developments in computing have also allowed several other statistical approaches to become central to the work of inferring population history, such as the landmark STRUCTURE method and approaches inspired by it [121, 1]. Several new statistical techniques have also been developed to address questions arising from new sources of data, such as the widely used family of f-statistics which infer historical relationships between potentially admixing populationsusing ancient DNA [114].

Nonetheless, many important questions remain open. Key features of hominid evolutionary history are still obscure or contested [120], requiring more data as well as the development of new methods, or the refinement of old ones, to resolve. Methodologically, the genetic effects of some key demographic processes are still hard to predict, especially when they interact with each other. This limits our ability to draw inferences about population history, and to assess the accuracy of those drawn from some popular methods. Using new data, mathematical analysis, and original applications of established computational tools, this dissertation addresses several aspects of these problems.

I have organised this work into three studies, summarised in more detail in Section 1.2. The first is an attempt to clarify aspects of the evolutionary histories of the genii *Pan* and *Pongo*. New samples of whole-genome sequences from these hominids are used to draw inferences about their population size histories and of gene flow between their populations. The approaches here are based on estimates of historic rates of coalescent events. This source of information about demography is known to be subtle to interpret in the presence of persistent non-random mating, or *population substructure* [87]. In the second study I look more closely at the effects of substructure and what they imply for the interpretation of coalescent time distributions. In the final part of the thesis, I incorporate ancient DNA in an attempt to develop an approach to directly test for ancestral population substructure, using a rich data structure expressing the relationships between the coalescent lineages of a sample. Applications in the last two studies are drawn from chimpanzee and human population history.

1.1.1 A brief history of hominid demographic inference

The earliest uses of genetics in hominid population history focused on human evolution. Although constrained to the study of single genetic markers, such as mitochondrial DNA (mtDNA) or the non-recombining region of Y chromosomes, starting from the 1980s, geneticists contributed influentially to early debates about the human region of origin. One notable instance was the reconstruction of a worldwide mtDNA phylogeny in 1987 which led to the widespread adoption of the out-of-Africa model [13]. Being studies of single locus ancestry, early efforts had limited statistical power, however. Later investigations of large sets of genome-wide single nucleotide polymorphisms (SNPs) and short tandem repeats supported the out-of-Africa model by demonstrating, for example, consistently greater levels of diversity in Africa [165]. Other studies also showed correlations between current population structure and geography [e.g. 136, 128]. After the development of efficient genotyping and high-throughput whole-genome sequencing technologies [for a historical review, see 44], more fine-grained analyses of extant population structure were made possible [e.g. 69]. This data has also facillitated the improved characterisation of the historic demography of populations, allowing us to detect paths of migration and past changes in population size [e.g. 134, 70].

A significant shift in our understanding of the past occured after the development of methods for the recovery of ancient DNA (aDNA). With the sequencing of neanderthal and denisovan genomes, the out-of-Africa model has had to be revised to accommodate archaic introgression into individuals with ancestry from outside Africa [38, 132]. Significantly, ancient DNA of humans from the last 50000 years has underscored the previously underestimated difficulties of inferring key aspects of population history without DNA from the past: the widely accepted histories of several populations have undergone major changes after the sequencing of DNA from relevant archeological sites [for reviews of the contributions of aDNA, see 104, 131].

It has since become common to combine modern and ancient DNA when reconstructing demographic histories of populations. Through this approach we have learnt that people who trace their ancestry to the out-of-Africa migration are likely to descend from a single founding Eurasian population [75], despite the probable existence of several waves of outward migration from Africa [47]. This is supported by several lines of evidence, including the roughly similar proportions of neanderthal ancestry observed in non-Africans [139], and the common timing of a major ancestral population bottleneck widely thought to be caused by the shared cross-continental migration [81]. We are also beginning to understand the composite and recent descent of modern Europeans from ancient farmers and several apparently distinct hunter-gatherer populations [68]. A similar process has likely occured in the South Asian subcontinent [100].

The influx of new data and analyses has also raised striking new questions, and drawn our attention to several longstanding ones. Hitherto unsuspected ancestry has been detected in several modern populations, one recent example being the Australo-Melanesian ancestry found in several Native American populations [127, 148]. The migratory route of this apparently distinct ancestral population is still unknown, nor is there much clarity on its precise relationship to other ancestral Native Americans. Increasingly, attention has also turned to the understudied prehistory of Africa, especially to questions around the timing and geography of the recent southward migration by people of West African descent [97], and to population structure during the early evolution of humans. We observe the deepest divisions between extant human populations in Africa [7], but we do not know the extent to which ancestral human populations predating this division were also structured. There is some interest in a new multiregional model of human evolution, consisting of several populations spread across Africa, and also in the possibility of archaic introgression from extinct hominids into the ancestors of modern Africans [150].

There is less data available for the study of non-human hominids. Initial sequencing of primate genomes prioritised sampling a large number of species and, especially, obtaining sequences of model organisms used in medical research [83]. Nonetheless, in the last decade significant demographic studies have been published on each of the non-human great apes:

chimpanzees and bonobos [123, 24], gorillas [183], orangutans [77, 102], and including at least one major comparative study of the entire family [120]. Ancient DNA from non-hominin great apes has not yet been published. Nonetheless, the research already undertaken has revealed complex histories of migration and gene flow between populations, as well as demographic changes due to changing environmental conditions, including the effects of hunting by humans. In Chapter 3 I review in more detail the population histories of chimpanzees, bonobos, and orangutan species.

1.1.2 The value of demographic history

While there is significant intrinsic interest in determining ancestral demography, an understanding of population history is also needed to address questions of broader value to biology and conservation. I consider three of these below: detecting the effects of selection, understanding speciation, and adequately controlling for structure in genome-wide association studies.

Isolating the effects of selection One of the oldest projects of population genetics is to determine the effects of forms of selection on variation within and between populations [122]. This understanding helps determine the causes of differentiation and adaptation, and allows us to isolate the functional consequences of genomic variants [e.g. 33]. Models of demographic history play several key roles in such studies, of which I highlight four: their use in the statistical tests which isolate loci under selection, their relationship to the effectiveness of selection, their use in locating the source and timing of variants, and the way they allows us to relate population history to ancestral environments.

Several statistical tests have been used to scan genomes and locate sites under selection [105]. One consistent difficulty with these approaches is the high rate of false positives, induced in part by demographic phenomena replicating the effects of selection: various statistics have been shown to produce spuriously significant values as a result of changes in population size [22] and due to the presence of population substructure [155, 141]. In order to limit false discoveries, it is useful to have a reasonable model of a population's demographic history. Ostrander et al. (2017) outline a representative approach they used to study selection in the canines [112]. They began by inferring demographic models of several dog breeds. These models included estimates of the times at which ancestral populations merged, and of various population bottlenecks. They compute the values of their test statistic (F_{ST}, discussed in Chapter 2) in windows across the genomes of their samples, and simulate many such regions under the demographic model they infer. These simulations are used to determine an approximate neutral distribution for the statistic, and with this a significance threshold for regions potentially under selection. Further analyses of the process of domestication homed in on these regions. Similar approaches have been applied to hominids [e.g. 86].

One of the longstanding theoretical predictions of population genetics is that selection acts more effectively in populations which have historically higher effective populations sizes (for more on effective population size, see Chapter 2). This prediction has been supported empirically in several great ape species, including humans [120, 45]. The observation implies that demographic history provides a useful prediction of the global effectiveness of forms of selection, and their influence relative to neutral processes like drift.

The migratory and gene flow history of a population allows us to determine the source of alleles under selection. Notable examples of this have been proposed in human evolution, where it has been suggested that several archaically-derived variants have experienced positive selection [e.g. 52, 125]. In such cases population history also allows us to obtain estimates of the time at which alleles entered a population and thus to estimate the strength of the selective force. Timing and migration are relevant in another way: they allow us to determine the regions in which an ancestral population might have been found, and thus to clarify the environmental conditions and possible selective forces acting on a population. This possibility is discussed in the case of orangutans in Chapter 3, although it is also a consideration in trying to determine the process through which human populations adapted to regional differences in diet and immunity [e.g. 2]

Understanding speciation Understanding how intrinsic reproductive barriers emerge between diverging populations is a problem at the centre of evolutionary biology. Some take it to be the defining characteristic of species [21], though various alternative species concepts are in use [48]. Regardless, reproductive isolation is generally thought to be an important component of speciation, and disagreements about classification tend to exist where populations fall in an intermediate region of divergence [137]. In recent years we have accumulated examples of gene flow between formally designated species, either ongoing or during secondary contact subsequent to a major initial divergence. In the hominids we have evidence of gene flow between bonobos and at least one chimpanzee subspecies [24], and also between humans, neanderthals and denisovans [104]. The phenomenon is particularly well-documented in *Heliconious* butterflies [93] and cichlid fishes [80]. Since gene flow tends to limit the differential accumulation of mutations which could reduce hybrid fitness, the mechanism through which reproductive isolation, or semi-isolation, can occur in interbreeding populations needs to be understood in the light of the history of gene flow between populations.

Reducing spurious signals in genome-wide association studies (GWAS) The increasing prominence of genome-wide association studies (GWAS) has turned attention to the problem of cryptic population structure. A rapidly developing form of research, GWAS attempt to explain the heritable components of phenotypic traits, such as height or risk of heart-disease, by the contributions of (typically) many genetic markers [168]. Several models

have been developed showing relationships between GWAS results and evolutionary processes like selection and drift [e.g. 146]. There is, however, a more direct role played by demography in conducting GWAS. Most such studies assume individuals in samples are unrelated to each other, but as samples get larger this gets harder to guarantee. When a sample is drawn from a structured population or when it contains individuals which are significantly related to each other, the related individuals will typically share many variants, only some of which will be causally related to traits of interest. This can result in spurious associations between alleles and study traits [160]. GWAS commonly reduce the rate of false positives by controlling for structure using a technique like principal components analysis (PCA), discussed in Chapter 2. Nonetheless, a fine-grained knowledge of the causes of stratification can aid in the correct design of studies. This can require understanding the admixture history of populations from which samples are drawn [108]. Last, in the development of polygenic scores, accounting for complex genetic correlations with quantitative traits, we have learnt that there are limitations in applying the results of association studies between spatially separated populations [28]. It might be useful in such cases to have accurate models of population history to better predict how well results generalise across populations and to identify populations which would need to be included in studies to make risk score models more generalisable [85].

1.2 Overview of thesis

Methodological review I begin by reviewing some population genetic theory which is relevant to arguments put forward in this thesis. Special attention is placed on outlining coalescent arguments, and on central concepts such as effective population size and population substructure. Ambiguity sometimes arises in the literature around the use of such concepts. I will primarily aim to clarify the way they are employed in this text, and will make no attempt to settle any foundational disputes. Methods which apply the sequential Markovian coalescent are frequently used and referred to throughout this work, so their theoretical underpinnings are developed in greater detail. I discuss some of their known difficulties as well. In Chapter 5, it will be important to have an understanding of the ancestral recombination graph (ARG) of a sample. This data structure is rich, but it is known to be difficult infer and applications are thus relatively rare. I try to convey as much of an intuition for the structure as is needed to follow the arguments set out in Chapter 5.

Following this will be a review of the key methods currently used to determine structure and relatedness in modern and ancient populations. These are organised by their underlying assumptions and statistical approaches. I outline in greater detail a few of the more popular tools, along with some of their limitations, placing special attention on those which I use or refer to in the text.

The demographic history of *Pongo* and *Pan* In Chapter 3 I present original studies of the population histories of orangutans, and of chimpanzees and bonobos. These investigations are based on the largest collected samples of whole genome sequences for each species. They aim to determine the historical changes in population size and cross-population migration rates which have shaped present-day genetic variation. Each genus is closely related to our own, and the chapter provides insight into the evolution of hominids in markedly different geographic and demographic conditions.

Orangutans are found on two islands, Sumatra and Borneo, corresponding to the two historically recognised *Pongo* species. Sufficiently accurate demographic histories of populations in the genus will help clarify the process by which they came to be separated. In particular, they will help us relate species history to past environmental conditions of the region. Recently, it has been proposed that one population on the island of Sumatra should be recognised as a species distinct from other Sumatran orangutans (and from Borneans) [102]. An accurate history of gene flow between Sumatran island populations will inform our understanding of the conditions under which these putatively separate populations evolved. Bornean orangutans are spread widely over their native island, and neither the ancestral relationships between current populations nor their migratory histories have been completely resolved. As mentioned above, such information is useful in studies which aim to detect the effects of natural selection in the species [86].

In contrast to the orangutans, chimpanzee subspecies and bonobos have evolved in relative proximity to each other. The degree to which populations have remained distinct in the past is unclear, though recent data has provided evidence for gene flow between bonobos and at least one chimpanzee subspecies subsequent to their major divergence period [24]. As with the orangutans, the environmental conditions which have maintained observable *Pan* structure are not well understood. Many studies favour explanations for population divergence involving varying rainforest cover, and prominent changes to the position and volume of equatorial rivers [161]. Information on the history of gene flow and population size will provide some insight into these questions.

The primary approach in this chapter is to use methods which involve estimating rates of coalescence within and between populations. This allows us to characterise the effective population size of ancestral populations, and to infer historical levels of gene flow between extant populations. I also use the *Pongo* and *Pan* studies to comment on some of the limitations and opportunities of these methods. This will have implications for their utility in studying other species, and in particular, other hominids.

The genomic effects of population substructure In the course of the previous chapter it emerges that population substructure can complicate our interpretations of ancestral rates of coalescence. This observation follows a recent theoretical study [87], and also an argument

made in the first publication in this lineage of methods [70]. It stems from the recognition that changes in census population size and in population substructure can produce effects on historical rates of coalescence which closely resemble each other. These histories are thus difficult to distinguish when exclusively using this source of information. In order to explore the parameter space of structured histories which can explain some past change in inferred coalescence rates, we need to determine the effects of ancestral population structure on the time distribution of coalescent events.

In Chapter 4 I present a formal model of a general demographic history featuring transient n-island structure, and I analytically determine its relevant coalescent distribution. For comparison sake I undertake a similar analysis of a single-pulse population size change, or "hump", model. Following this, I propose an approach which can be used to narrow the search space of explanatory demographic histories. The technique is based on the idea of specifying a hump model, using the ancestral effective population size history of some individual or group, and matching it with a set of transient n-island models. The matching is based on finding models with similar divergence from the coalescent distribution of an appropriately chosen panmictic, constant-sized history. The theoretical analyses of these models makes this process easier than would be the case if we used simulation. Once a range of plausible structured models is determined, up to the constraints of the models being compared here, we can appeal to external evidence, genetic or otherwise, to assess their plausibility.

This is followed by two empirical applications of the technique. The first is to a situation drawn from the previous chapter, where we were concerned with the interpretation of one specific change in the inferred population size of a chimpanzee subspecies. The second is taken from human population history. We determine whether one of the characteristic signatures of population size change in out-of-Africa populations can alternatively be explained by changes in population structure. In both cases I suggest that structure is less likely as an explanation than a genuine change in census population size, though the conclusion is stronger in the case of the chimpanzees, where the relevant timescale excludes most forms of structure. This empirical discussion is followed by a brief examination of the effect of relaxing one assumption related to coalescent lineage sorting at the start of structured periods. I observe that while it is generally accepted that population structure will tend to depress coalescent rates relative to comparable panmictic conditions, under certain admixture regimes coalescent rates can be inflated. This situation may not be uncommon in human history. We conclude by noting the intrinsic difficulty of detecting periods of ancestral population structure using modern sequence data alone. The approach developed in this chapter can help narrow down the range of plausible histories, but recent experience has taught us that even striking examples of ancient substructure can sometimes only be detected using ancient DNA [68]. This insight is the starting point for the following chapter.

Inference of cryptic ancestral substructure using the ARG In Chapter 5 I develop a haplotype-based approach to more directly test for ancestral structure. The method uses an analysis of the ancestral recombination graph (ARG) of a modern sample, and compares haplotype segments isolated using the ARG with variation found in relevant ancient sequences. I construct an explicit demographic model for the sharing of variants between modern and ancient DNA, and use this to explore hypotheses about ancient population structure. The model attempts to detect the times during which hypothetical ancient populations merged, and through this, infer aspects of population structure.

Some of the implications of this method are shown using simulations. Empirically, I use the approach in an attempt to address a question around population structure relevant to resolving the sequence of demographic changes which occurred during the peopling of North America. Ultimately, the approach will be constrained by the availability of high-coverage ancient DNA, though it indicates some of the potential information gained from inferred ancestral recombination graphs.

Conclusions I end by discussing the findings of each chapter, drawing out their relations to each other and to the established literature. I also highlight some important avenues for future research based on the limitations of the work in this thesis, indicating some of the opportunities provided by the new approaches developed here.

Chapter 2

Methodological Background

I have divided this review into two sections. In the first, I outline key parts of the theoretical background of this thesis. This will also allow me to clarify the terminology I use and to state the simplifying assumptions I make. In the second section, I examine several of the statistical and computational tools commonly used to infer population history, with an emphasis on those used to study structure. I also discuss some of the limitations of these methods, and their relationships to those developed in Chapters 4 and 5. In both sections, special attention is paid to the sequential Markovian coalescent and its applications, as this body of work plays a prominent role in the thesis.

2.1 Theoretical background

The coalescent and its extensions The coalescent is a stochastic process modelling the genealogy of a sample of individuals at one locus. For a generic locus (such as a gene) and a sample of n individuals drawn from a Wright-Fisher population of size N, the coalescent is standardly derived by tracing the ancestry of individuals going backwards in time from the present, allowing them to "choose" parents at random in each generation [171]. The probability that any pair of lineages in the sampled generation do not choose the same parent one generation back is $1 - (2N)^{-1}$, on the assumption that the population is diploid. In the limit as $N \to \infty$, and applying the standard time scaling of t in units of $2N$ generations, we obtain the fact that the waiting time to choosing the same parent, or *coalescence*, of any pair of samples (technically, their ancestral lineages) is exponentially distributed, with probability density e^{-t}. Analogously, when starting with a sample of size n we would wait some random time T_n for the first coalescence, where T_n converges to an exponential distribution as $N \to \infty$, with mean given by $1/\binom{n}{2}$, since in large N Wright-Fisher populations it is a sufficiently good approximation to consider only the probability that at most two lineages can coalesce in a given generation. This observation underpins the continuous time Markov chain known as the Kingman coalescent, or sometimes the n-coalescent [62]. (Formally, the states of the

Markov chain are equivalence relations that identify lineages which share ancestry at a given point in the evolution of the process.)

The convenience of modelling genealogies using this approach arises from two key properties: the mutation process is handled independently of genealogy, and a large class of models, with varying breeding structure, can be approximated by a coalescent, provided the correct time-scaling is chosen. It has been pointed out that while this latter "robustness" property of the coalescent justifies its use as a model to help understand many natural populations, it also implies that sample variation will to some degree be insensitive to changes in breeding structure [147]. Typically, variation is modelled by placing mutations randomly on the branches of a coalescent-generated genealogy according to a Poisson process with intensity given by the some scaling of the mutation parameter θ (usually $\theta/2$). The mathematical advantages of this approach were demonstrated early through simplified derivations of several classical results, like Ewen's sampling formula [63].

Since the initial development of the coalescent model, it has been generalised or extended in several ways. One strong assumption underlying the standard coalescent is that of the exchangeability of lineages. This arose in the Wright-Fisher derivation as a consequence of the assumptions of neutrality and absence of population structure. Some generalisations of the coalescent involve relaxing this assumption, to allow some pairs of lineages to have a greater probability of coalescing with each other than with others. This has allowed models of population structure to be developed [107], as well as models featuring various kinds of selection [56]. Another form of generalisation involves embedding coalescent processes in richer random graph structures which model the genealogical relationships between linked loci, such as is needed to model the complex ancestral relationships of recombining sequences [50]. This is described further in the section on the ancestral recombination graph (ARG) below. Finally, other generalisations, so-called Λ-coalescents, allow multiple coalescent events in a single generation [96]. Although I will not be referring to these, they are useful when modelling populations in which offspring number among individuals has much higher expected variance than in a standard Wright-Fisher population.

Effective population size The original motivation behind defining an effective population size, N_e, is to generalise the value of N in a Wright-Fisher population. It allows us to rescale more complex population models so that they can be seen to behave in some salient respect like Wright-Fisher populations with size given by the values of N_e. Typically, natural populations have N_e much smaller than their census population sizes [17], though precise values depend on the behaviour being considered. These have included such factors as the increase in variance of allele frequency, extent of inbreeding, or several other characteristics of drift [30], and they can be inconsistent with each other even if closely related [175]. The effective population size has been formulated in terms of coalescent processes too, in which

the N_e of a complex population model is derived from the linear timescaling required to convert the genealogical process underlying it to a standard coalescent, although in some cases this so-called coalescent effective size has been shown not to exist [147, 171]. The concept plays a central role in population genetics, as the product of N_e and mutation rate determines the level of variability at neutral sites, while the product of N_e and the selection coefficient determines the probability that a non-neutral allele will be driven by purifying or positive selection to fixation [16].

A slightly different notion of effective population size has become more common in recent literature, and I shall be using this related concept. Instead of attempting to characterise the genealogical behaviour of a population model using a single number (which the model might only approach asymptotically), it is common now to treat N_e as a function of time, as the parameter derived from the inverse historical (and usually varying) rates of coalescence. This can be seen as coming from the genealogical process underlying a Wright-Fisher population with time-varying N. The usage has been implicit in several recent methods which attempt to infer historical population size changes [e.g. 70], though it was pointed out explicitly by Mazet et al (2015) [87], who propose using a new vocabulary for this concept, calling it the inverse instantaneous coalescent rate (IICR). I have not adopted their terminology in this thesis for the sake of consistency with the usage of N_e elsewhere. While its common use is relatively new, the notion of a time-dependent N_e has been implicitly adopted from at least as early as Slatkin and Hudson (1991) [153], who show that the scalar version of N_e can be derived from a harmonic mean of population sizes over a population's history, provided the sizes are sufficiently large. Consistency between these definitions arises from equating the inverse of the expected times to coalescence [17].

Modelling population structure Models of population structure have been studied since the early years of modern population genetics, and several have become key tools of analysis. The simplest, which I use most often in this thesis, is the island, or n-island, model [181]. In these models a population is divided into several subpopulations, or demes, with migration determined by a single symmetric rate M, which, from a coalescent perspective, determines the distribution of waiting times between inter-deme migrations of ancestral lineages. The stepping-stone model relaxes the assumption of symmetric migration probabilities but retains the island structure [60]. In one or two dimensions, this structure only allows migration between adjacent islands and models the effects of decreasing genetic relatedness by distance. It is also possible to arrange islands in hierarchies, so that migration is greater between some clusters of populations, and migration across cluster boundaries is relatively rare [154]. Coalescent modelling of island-type models demonstrates a separation of timescales, in which lineages in the same, or closely-related, demes tend to coalesce quicker than lineages in distinct demes [107, 106]. The first phase has been described as the "scattering" phase, and is

followed by the much longer "collecting" phase [170]. In the limit, the genealogical distribution underlying these models has been shown to approach the *structured coalescent*. Note that these three island-type models, the basic n-island, stepping stone, and hierachical model, tend to be analysed in stable form, in which it is assumed that no changes in demographic structure occur through the course of the population's history. This assumption is discussed further in Chapter 4.

Another way of understanding the genetic effects of population structure is to try and model the historical relationships between populations observed today. Classically studied by attempting to resolve populations into graphical relationships similar to phylogenetic trees, modern approaches now study them as so-called admixture graphs, which allow populations to merge and split in the past [e.g. 114]. Still more flexible approaches can accommodate both admixture relations and periods of migration between populations, though at the expense of computational efficiency [e.g. 31]. Both these sorts of models are described in more detail in the following section. Regardless, there is a strong trade-off in each of the modelling approaches mentioned above between mathematical tractability and real-life complexity, and strong generative models of such phenomena as fine local structure or population gradients are still lacking [109].

Finally, the degree to which a real population departs from panmixia has a classical quantitative measure, called F_{ST} [182]. I refer to this several times, usually using its modern coalescent form demonstrated first by Slatkin (1991) [151]. The theoretical form of the measure is given by $F_{ST} = (E(T_D) - E(T_S))/E(T_D)$, where $E(T_D)$ is the expected time to coalescence of loci chosen from different demes, and $E(T_S)$ is the expected time of loci from the same deme. In humans, this value can range from 0-15% [109]. One limitation of using F_{ST} to understand population history is that it requires prior information about spatial structure in a population in order to make sense of the concepts of "same" and "different" demes. As such, it has less value in work, such as this, where we are more interested in detecting forms of structure than in quantifying known subdivision.

Ancestral recombination graphs (ARG) Briefly, recombination was first incorporated into coalescent theory by Hudson (1983) who showed how to model recombination in two- and four-locus models [50]. Since it was proposed by Griffiths (1991) [39], we think of recombination and coalescent processes together as generating the ancestral recombination graph (ARG) of a sample. In this graph, nodes represent either recombination events in the sample or coalescence events, while edges can be thought of as representing ancestral sequences. Going backwards in time, recombination events cause sequence segments to split, and coalescent events cause them to merge. The process terminates when the most recent common ancestor (MRCA) of the entire sequence has been attained. I sometimes refer to the regions in sequences which correspond to single ancestral trees as segments, or even where

no ambiguity is likely, as loci. In other words, these are the sequence intervals on which no recombination event causes a split before the MRCA is reached. Most relevant for what follows, the genealogy of any individual locus, its "marginal genealogy", can be modelled using a single-locus coalescent process, although coalescent trees generated at linked sites are of course not independent. A pioneering algorithm to simulate sequences using the ARG, called `ms`, was developed and implemented by Hudson (2002) [51], though a greatly more efficient process, `msprime`, has recently been developed by Kelleher, Etheridge and McVean (2016) [58]. Before the exact algorithm underlying `msprime` was implemented, less efficient simulation (still quicker than `ms`) had been conducted using an approximation to the exact coalescent with recombination, called the sequential Markovian coalescent, summarised below. Some work in this thesis uses the simulation tool `scrm` which is based on this approximation [157], and in Chapter 5, I use ARGweaver, a Markov Chain Monte Carlo method which samples the ARG given data under similar assumptions [130].

The sequentially Markovian coalescent (SMC) The sequentially Markovian coalescent (SMC) model was developed by McVean and Cardin (2005) [91] and Marjoram and Wall (2006) [82]. Like the coalescent with recombination that it approximates, described above, the SMC traces the ancestry of present day sequences backwards in time till the MRCA of the every locus is obtained. Motivated by the computational difficulties of likelihood inference (and until msprime, simulation [58]) under the full model, SMC restricts the class of permitted coalescence events by limiting long-range linkage.

Without this approximation, the state space of possible ARGs is large, in part because the ARG can contain a large amount of uninformative genealogical data. Many coalescent events, for example, occur between sequence segments which do not contain overlapping ancestral material, intervals of the sample sequences which have not yet reached their MRCA. The SMC simplifies the full model by preventing coalescent events between lineages which have no overlapping ancestral material. This has the effect of greatly reducing the number of possible ARGs, while leaving the marginal genealogies of ancestral material unaffected. Inference of SMC graph structure is dependent solely on the marginal genealogies of sequences of ancestral material, and while these are embedded in the full ARG, they do not uniquely determine it. Efficient inference schemes for the full model have not yet been developed.

The model is named for the way it generates linked genealogies sequentially along sequences. These methods are based on the "spatial", rather than "temporal", formulation of the coalescent with recombination. A preceding algorithm originally proposed by Wiuf and Hein [180] uses the full infinite-sites (defined in the following section) coalescent with recombination. It generates genealogies by moving along the sequences and updating a generated history whenever a recombination point is reached. Crucially, the step in which the history is updated requires knowledge of the states of the histories at all previous locations

along the sequence. This requirement is relaxed in the SMC model. SMC ensures that the distribution over potential histories at each step along the sequence depends only on the previous configuration of the history. The process thus possesses the Markov property and opens itself up to more efficient analysis with Markov chain techniques.

Studying several summary statistics derived from genealogical structure, certain linkage disequilibrium properties, and the distribution of expected times to most recent common ancestor (TMRCA), McVean and Cardin demonstrated that while significantly reducing computation costs, this assumption does not greatly affect inferences about population history [91]. Though it has since been shown that the SMC can be a bad approximation when modelling the length of shared tracts in admixing populations [74].

Key modelling assumptions Throughout this thesis I shall be making several simplifying assumptions when modelling population history. I shall assume that breeding structure in the hominids is well modelled by a Wright-Fisher constraint on offspring variance, or equivalently, that genealogies at independent loci are best modelled by the standard coalescent with time-varying N_e. This entails the assumption that we are drawing our samples from relatively large populations, which seems appropriate in the case of the hominids [16, 120]. I also assume that most of the variation shaped by the recent population history has evolved neutrally. This is partly a constraint imposed by the methods that I employ, and although still contested, recent research is beginning to show much more of the genome than was expected (possibly as much as 95%) has been shaped by selection [e.g. 119]. Where relevant I shall mention any potential effects that this might have on inferences. Together, these assumptions imply conformity with the exchangeability assumption of the classical coalescent. This assumption will only be broken when modelling population structure, which will be clear from context.

Several assumptions will be made regarding the modelling of mutation and recombination. I shall be applying the *infinite-sites* constraints when modelling mutation [61]. These assume that mutations never occur at the same place more than once and as a result that segregating sites in a sample of sequences can simply be described by binary states, usually denoted 0 for ancestral type and 1 for derived (where such information is known). Indels, structural variants, and short tandem repeats will not be looked at. I will also be assuming constant mutation and recombination rates along the genome. This is a strong simplifying assumption which we know not to be true [49]. In the relevant sections where I draw on constant values I will discuss where they come from. I do not expect the aggregate effects of this assumption to bias any of the major inferences drawn about population history. Similarly, I will be assuming that over the timescales of interest, mutation rates are constant. Since chimpanzees and humans, for example, are thought to have different mutation rates today, there must have been some change in this rate in the time since their divergence [142]. The same can be said for generation length, viewed as the average zygote-to-zygote time, which will also

be assumed to be constant, even though we have evidence that this parameter has changed during evolution [64]. It is unclear over which time periods (and on which phylogenetic branches) these shifts in mutation rate and generation length have occurred, although I expect that since we focus on recent population history the assumption of constant values is not significantly inaccurate.

2.2 Methods of inferring population history

Broadly speaking, studies of population history use genetic data aim to learn two things regarding the population structure of their sample: (a) the extent to which variation between individuals reflects historical and extant substructure, and (b) the history of admixture and migration between inferred or presupposed populations. Typically, each study uses several methods to address these questions. The methods exploit different correlations between individual sequences, use different statistical techniques, or are based on different modelling assumptions. As such, they shed light on complementary aspects of population history, although their conclusions can sometimes be difficult to reconcile with each other or with plausible demographic models. In this light, it is especially important to understand the limitations and appropriateness of each approach. I focus on the ability of these methods to infer past population structure. The first approach, PSMC, is used more broadly to estimate past rates of coalescence, and uses this as a proxy for history of population size changes.

2.2.1 Inferring population history with PSMC

Li and Durbin [70] developed a hidden Markov model (HMM) which implements a coalescent time inference scheme based on the SMC in the special case of diploid sequences. It is called the pairwise SMC (PSMC), and is based on the insight that variation in the local density of heterozygotes is informative of past recombination and coalescent times along the sequences inherited from an individual's parents. Intuitively, this arises due to the fact that segments of the diploid sequence at which parent haplotypes have older TMRCAs have had more time to accumulate segregating mutations. Thus, for example, adjacent regions with significantly different densities of heterozygotes indicate different ages of first common ancestry between their respective parent haplotypes, and also suggests that recombination occured between their ancestors. PSMC seeks to estimate the distribution of the TMRCA of segments along the unphased sequence of an individual and uses these times to estimate the rate of coalescence at various times in the past. These rates are inverted to determine a piecewise constant estimate of historical N_e.

In more detail, unphased sequences of individuals are segmented into bins of some chosen length. Moving sequentially along the pair of sequences, the model observes a binary string of 0s and 1s, corresponding respectively to the presence or absence of at least one heterozygote

in a bin. Bins which have a great proportion of loci absent or masked are observed as missing data points. The hidden states of the HMM consist of the discretised TMRCA, with transitions between the states corresponding to ancestral recombination events. The SMC approximation guarantees the Markov property of the path through hidden states.

The free parameters of the model are the scaled mutation rate θ, the scaled recombination rate ρ and the piecewise constant effective population size, $N_e(t)$. The emission probabilities from a (hidden) state t are $e(1|t) = e^{-\theta t}$, $e(0|t) = 1 - e^{-\theta t}$. When there is a data point missing $e(.|t) = 1$. The probability of transitioning to state t from another state s, is given by

$$p(t|s) = (1 - e^{-\rho t})q(t|s) + e^{-\rho s}\delta(t - s). \tag{2.1}$$

Here, $q(t|s)$ is the transition probability conditioned on the occurrence of a recombination event, and δ refers to the Dirac delta function. Assuming that the neutral mutation rate μ is known, we set $N_0 = \theta/4\mu$. If $\lambda(t) = N_e(t)/N_0$ is the relative population size during state t, then $q(t|s)$ is given by

$$q(t|s) = \frac{1}{\lambda(t)} \int_0^{\min\{s,t\}} \frac{1}{s} \times \exp^{-\int_u^t \frac{dv}{\lambda(v)}} du. \tag{2.2}$$

These expressions are derived by constructing a continuous-time Markov chain according to the SMC model and integrating transition and emission probability densities over discretised time intervals. The resulting discrete-time process forms the basis of the HMM. The discrete intervals increase exponentially in length the further back in time the model runs. To reduce the complexity of the search space, the user inputs a specific pattern of time intervals over which the inferred population remains constant. The estimated distribution of TMRCA is here determined in units of coalescent time $2N_0$ and is scaled to real time using μ (mutations per base pair per generation) and generation time g (years). Model parameters are estimated using the expectation-maximisation (EM) algorithm, with Powell's direction set method used to numerically minimise the Q function in the maximisation step.

The multiple sequentially Markovian coalescent 2 (MSMC2) [79], is named after MSMC, an earlier extension of PSMC to multiple sequences developed by the Schiffels and Durbin (2014) [144]. Given a set of haplotypes from two different population groups, MSMC2 does a pairwise comparison of sequences from the two groups similar to PSMC with the SMC' correction [82]. Instead of trying to infer the overall gene flow history on an *ad hoc* basis using these curves, MSMC2 infers a composite likelihood of the data [172]. I use this in Chapter 3 to infer histories of gene flow between populations and compare the results with a similar analysis using PSMC.

Limitations Li and Durbin show that using the inferred TMRCA distribution, PSMC performs well in recovering certain population size histories from data simulated using ms.

However, since relatively few coalescent events between pairs of loci in humans occur before about 20kya and after 3 Mya, PSMC is limited in its ability to make inferences about the TMRCA distribution, and thus population size, outside of those boundaries. A greater number of independent genealogies, and thus a greater number of sequences, are needed to make inferences about more recent histories. In the analysis of sequence data, this limitation is seen in the sometimes excessive degree of variation in curves derived from individuals of a single population group in both very recent and very ancient times. PSMC is also limited in its ability to detect sudden changes in population size, such as historical bottlenecks. It tends to smear out these changes over significantly longer time periods. Li and Durbin run their analysis on `ms` simulated data derived from a population which 100kya undergoes an instantaneous collapse in N_e. PSMC infers that this change occurred at an even rate between 100kya and 200kya. Also, the inferred population histories scale linearly in estimates of the generation time and mutation rate, which are not inferred by the model. As such, any uncertainty in those parameters affects the interpretation of the results in real-time. A twofold increase in the mutation rate, for example, would halve the estimate of N_e, while generation length shifts curves in time. These effects are discussed in Chapter 3.

As noted above, the SMC limits long range linkage and this can bias the inference of ancestry tract lengths when populations are admixing. PSMC may fail to account for this. A different kind of difficulty with complex demography arises in the interpretation of coalescent rates. It was pointed out by Mazet et al (2015) that it can be difficult to interpret the coalescent rates of structured populations. This is discussed further in Chapter 4. Finally, as noted above, the assumption of selective neutrality across the genome might not be appropriate, and there is specific evidence in the case of PSMC that linked selective sweeps can bias results in complex ways [145].

Model identifiability Important theoretical questions underlie the attempt to infer population histories. One set of questions relates to the degree to which sequences contain information about past demographic events. In other words, we do not know how distinct from each other two histories need to be before we can distinguish sets of sequences from either population, nor how much sequence data would be required to make the distinction at a desired level of confidence.

We need not be confined to SMC approaches either. We might ask if a distinctive enough signal of past demography persists at all in the genomes of living populations, and how strong that signal is if it does. This question, of *model identifiability*, has received some attention over the last few years, though the first significant result was by Myers and others in 2008 [99]. Using a different approach to the models described thus far, they asked whether it would be possible to find two population histories which produced the same expected allele frequency spectrum today. They assumed an infinite sites model with panmictic mating

and approximated the effects of drift using a diffusion process. Under these conditions they showed that indeed it is possible to find such population histories. Even a perfect knowledge of the expected allele frequency could not distinguish these scenarios. In 2014, Bhaskar and Song relooked at this question, starting from a similar set of assumptions to the earlier paper [8]. They observed that although the sets of histories offered as counter-examples were mathematically interesting, they were biologically unlikely: they required large oscillations in population size on a time scale much shorter than the length of a generation. In their paper, Bhaskar and Song asked if the result still held if you imposed meaningful constraints on possible histories. They assumed that the population shape could be resolved into time intervals on which the number of individuals was either constant or expressible as an exponential function. The result in this case is positive. A perfect knowledge of the population allele frequency spectrum would guarantee the uniqueness of the inferred history. The paper went further and determined that complete knowledge was not required: provided you had a large enough sample, its allele frequency distribution is sufficient to distinguish between underlying demographic models. Moreover, they supplied a general lower bound on the size of the sample required.

This significant result formed the starting point of a paper by Kim et al (2015) whose inference scheme is closer to that of the PSMC [59]. Assuming populations histories can be expressed as piecewise constant functions, they posed the question as a hypothesis testing problem, attempting to distinguish between two population histories which differ only on a single time period, between T and $T + S$. Beginning with a different idealised data set, a collection of L independent coalescent times observed from one of the populations, they manage to place a lower bound on the uncertainty of our ability to decide from which population the data originated. Intuitively, the bound grows when S is small, T is large, and the extent of the difference in population size histories over the period is small.

This result is relevant because it is the first analytical indication of what the absolute limitations to SMC-based inference might look like. In fact, the limitations apply more generally to those methods which infer demography via coalescence times. Kim et al. use their result to argue, for instance, that the difficulty PSMC has with inferring sudden changes in effective population size is a problem from which all methods based on inferring coalescence times must suffer. This is due to the increased difficulty of deciding the history as the value of S decreases, for some fixed value of L. Since the lower bound is placed on idealised data points, usually only inferred during analysis, the lower bound on real-life data is expected to be considerably higher.

2.2.2 Methods for the inference of historical population structure

Principal component analysis (PCA) PCA is a non-parametric method widely used by applied statisticians to summarise and reveal structure in high-dimensional datasets. Its

use in population genetics was pioneered by Menozzi, Piazza and Cavalli-Sforza (1978) [92]. Visual appeal and ease of use account for its enduring popularity. Researchers present low-dimensional projections of their samples onto subspaces spanned by the first few "principal components". These can be understood in several ways, most simply by observing that it is possible to orthogonally project study individuals onto any line in the feature space and calculate the variance of the resulting set of points. The first principal component is the line with greatest such variance, and the second is the line with greatest variance which is also orthogonal to the first. Higher components are defined analogously. (Finding unique axes is not in practise a problem.) Principal components were classically determined by calculating the eigenvectors of the covariance matrix of a zero-centred matrix representation of the data. Newer approaches use a singular value decomposition of the data in order to avoid constructing the potentially very large covariance matrix. In the resulting low-dimensional projections, more genetically similar individuals, under a similarity measure related to shared SNPs and their respective frequencies, end up closer to each other [135, 113, 66].

While the method makes no demographic assumptions, observed patterns in projections are frequently interpreted in the light of simple historical processes. For example, a clustering of individuals into several clear groups is often taken as evidence of long-standing population substructure [e.g. 81], and it is possible to formally test this clustering [115]. However, several simple historical processes can be superimposed, and high dimensional projections are hard to visualise, so the inheritance of multiple ancestry components by one population can be difficult to detect. Moreover, markedly different historical processes can produce similar projection patterns. In the original applications of the method by Menozzi et al., gradients were read as evidence of outward migration from a source population. It has since been shown that this pattern can be produced through isolation-by-distance in equilibrium demographic models [110]. McVean (2009) showed that the projective distance between individuals increases with their mean coalescent time [90]. This provides a systematic way to generate and assess plausible histories of a sample provided sequences are close in age (McVean's implicit assumption). While it is still a useful summary of genetic similarity, no similarly straightforward theoretical interpretation of the PCA of a mixed ancient and modern sample has been proposed.

STRUCTURE-like population identification Populations are commonly identified using methods which assume sequences are drawn from an admixture of several discrete populations. Usually, the number of populations K is preset. These clustering methods include STRUCTURE [121], Frappe [163], and ADMIXTURE [1]. The approach underlying the earliest of these, STRUCTURE, is still influential. It assumes that each population is in Hardy-Weinberg equilibrium, and that loci are unlinked and so in complete linkage equilibrium. Populations are characterised by their allele frequency distributions at sample

loci, and individual sequences are supposed to have been generated by sampling randomly from the distribution of the appropriate population at each locus. Individuals might draw from multiple populations according to their "ancestry components". The novelty of STRUCTURE lay in its ability to jointly infer the allele frequency distributions of the K populations and the ancestry components of individuals. Other methods take a similar approach, but differ in their choice of inferential framework. For example, where STRUCTURE uses MCMC to estimate model parameters, ADMIXTURE achieves considerable performance gains in computational time using maximum likelihood techniques based on the same underlying likelihood function. There is no consensus on the best way to choose K, but since many values of the parameter are typically informative, the methods tend to be run on several. If a single optimal K is desired, STRUCTURE chooses the number with greatest estimated model evidence, while ADMIXTURE chooses the number which produces the most robust output when random portions of the data are masked.

These methods have succesfully recapitulated known historical patterns of migration. Famously, STRUCTURE was used by Rosenberg et al. (2002) to detect genetic clusters corresponding to continental groups as well as smaller-scale subpopulations [136]. However, they are liable to misinterpretation when populations have complex migratory histories [32]. In addition, assuming a discrete population substructure where variation between samples is continuous, is likely to bias inferences. PCA can help to decide the appropriateness of the discreteness assumption, although recent work by Bradburd et al (2017) more directly addressess this problem [11]. Another concern with these approaches is that little validation has been performed on the effects of staggered sampling times on their output and interpretation. Evidence that does exist suggests differences are likely to be significant [55]. Without further study, it is difficult to say how best to interpret these methods on datasets featuring both modern and ancient sequences.

Ancestral admixture analysis using f-statistics Although they incorporate no explicit historical models, the output of methods like ADMIXTURE are often informally interpreted in the light of ancestral relationships between populations. Methods which model ancestral relationships are usually based on admixture graphs which modify population-level generalisations of phylogenetic trees by allowing populations to descend from combinations of several ancestral sources. An influential statistical framework for fitting admixture graphs uses the recently developed family of f-statistics [134, 114].

Their key observation is that it is possible to additively partition drift along branches of an admixture graph. Shared branches between populations are revealed by the covariance of allele frequencies. This insight underpins the definition of the first quantity $F_2(P_1, P_2) = \mathbb{E}[(p_1 - p_2)^2]$. In this notation, P_1 and P_2 refer to populations, and p_1 and p_2 refer to allele frequencies of biallelic loci in the respective populations. This quantity can be used

to derive a distance metric between populations, measuring the amount of shared drift between them and thereby, for example, allowing us to construct a similarity matrix of a sample of populations. Using classical results from phylogenetics, we could use this matrix to test the treeness of a sample. However, F_2 is more important as a theoretical quantity underpinning the definition and analysis of the quantities $F_3(P_0; P_1, P_2) = \mathbb{E}[(p_0 - p_1)(p_0 - p_2)]$ and $F_4(P_1, P_2; P_3, P_4) = \mathbb{E}[(p_1 - p_2)(p_3 - p_4)]$. (The notation is analogous to that of the definition of F_2.) These are theoretical quantities; the corresponding sample statistics are referred to as f_2, f_3 and f_4. The most important application of f_3 is as an admixture test, while f_4 is used most often to determine admixture proportions under an admixture graph model. Related to this framework is the D-statistic, described in more detail in Chapter 5.

Peter (2016) has shown intuitive connections between f-statistics, classical phylogenetics, and the expected branch lengths of coalescent trees [117]. An inference from this analysis is that f_3 has greatest power as a test for admixture when admixture proportions (from P_1 and P_2 into P_0) are equal, when the target population (P_0) has a large effective population size, and when the original split between populations is much earlier than the secondary contact. This illustrates the general fact that the expected values of these statistics depend on certain features of demographic history. Less pertinent when the statistics are used as a binary test for admixture or treeness, demography should be considered in applications where the statistics are used to fit model parameters or compare likelihoods of admixture graphs. Patterson et al [114] warn that the methods should be used cautiously in situations where populations have experienced a greater than expected amount of drift due, for instance, to serial founder effects.

Some other admixture graph methods, such as TreeMix [118], based on F_3 and F_4 statistics, automatically optimise graph choice. Although, as generally used, the ADMIXTOOLS framework requires user specified models and it is unclear in applications if any possible graphs are missed from their model selection. One deeper simplifying assumption the framework makes is that the underlying relationships between populations can be modelled with the graph structure. However these graphs cannot explicitly model periods of migration which do not amount to merging or splitting of populations. It is uncertain how these other forms of demographic relationships between populations can affect inferences, although the theoretical results of Peter, mentioned above, provide a starting point.

Haplotype-based approaches Sets of haplotype segments which share a long ancestral history are particularly useful in demographic analysis. They can be identified in several ways. In principle they can be derived from the ancestral recombination graph of a sample, but since this is computationally difficult to infer, approximate methods have been developed. One powerful model is the the Li and Stephens copying model [73], upon which popular tools like ChromoPainter are built [67]. ChromoPainter determines the local ancestry of

segments according to their similarity to haplotypes in a reference, or "donor", set. The "painting" of some locus of a modern haplotype by a donor is equivalent to identifying the donor which has the most recent common ancestor with the modern individual at that locus, relative to the other donors. Sites of historical recombination correspond to endpoints of the segment intervals, also called "chunks". ChromoPainter infers these using the Li and Stephens hidden Markov model. One application of ChromoPainter is to study the local structure of populations, and the tool most commonly employed to do so is fineSTRUCTURE [67]. These approaches are limited by the requirement that data is well-phased.

Skoglund et al (2015), working with a sample of modern and ancient haplotypes, used ChromoPainter to obtain segments of shared ancestry [148]. They use a pair of ancient sequences as the donor set with which they paint a sample of modern sequences. For each modern haplotype, they compare the respective counts of chunks painted by the ancient sequences. This analysis is conducted on a geographically widespread modern sample and complemented with global admixture tests based on D and f statistics. However, chunk-counts can be difficult to interpret in isolation. The difference in donor-matched chunks depends on the relative window of opportunities for coalescence with the ancestors of the donor haplotypes. Human population coalescent times inferred by methods like PSMC, suggest that many coalescent events will be old. Thus many chunks will coalesce in the common ancestors of all study populations. A random portion of the matches with donors will then be independent of recent population history. The extent to which this will bias chunk-counts in any individual study should be accounted for with demographic models which incorporate parameters known to affect opportunities for coalescence, such as effective population size, migration conditions, and population substructure. The age of ancient sequences, if they are significantly different, will also affect the opportunities that the ancestors of the samples have to coalesce and should likewise be modelled. This point is expanded on in Chapter 5.

Flexible generative methods Another broad framework for fitting models of population history involve what can be described as generative methods [65]. These are based on using the results of simulations of potential demographic histories to fit model parameters to observed data. In principle they can model any form of demographic process that can easily be simulated. The most influential framework for optimising parameter choice is the approximate Bayesian computation (ABC) approach [5]. ABC arguments involve using descriptive statistics to summarise properties of the genetic composition of a sample, then simulating genetic data under various historical models and optimising the choice of simulation parameters which produce comparable summary statistics in simulated data. Parameters so chosen are assumed to approximate the true population history. While these methods have had some success, especially when used in limited contexts [137], they depend on choosing informative summary statistics which can be computationally infeasible to approximate with

simulated data. Recent approaches involve using regression methods to reduce the difficulty of finding matching simulation parameters [23].

Chapter 3

Chimpanzee, Bonobo, and Orangutan Population History

Demographic processes which have shaped the genetic diversity of hominins have had a similar influence on non-human great apes. The family is phenotypically diverse and extant species have survived markedly different environmental and demographic conditions. These differences allow us to compare the effects of demographic processes on closely related species under varying conditions, and to better isolate their consequences. Given the unusual amount of data we have on great apes, this exercise provides insight into genomic evolution with relevance beyond the family.

In this chapter we look at historical effective population size trajectories and cross-population gene flow in the orangutans, and in chimpanzees and bonobos. Inferences of these histories are made using the sequential Markovian coalescent (SMC) methods described in Chapter 2 and draw on the largest known whole genome datasets available for these species. The chapter is divided into two sections. Except where indicated the material is original. The first focuses on orangutans and incorporates material I produced for Nater et al (2017) [102] and Mattle-Greminger et al (2018) [86]. The second section looks at both chimpanzees and bonobos, and is based on analyses I produced for de Manuel et al (2016) [24].

The comparison between these two genera is particularly interesting given that speciation in the orangutans occurred across now-isolated islands (Borneo and Sumatra) which experienced different climatic and environmental histories. (According to recent evidence, speciation also occurred within Sumatra, though as I will describe below we have less salient information about this event.) On the other hand, chimpanzee and bonobo speciation is thought to have occurred due to the formation of the Congo River which, it has been argued, separated their common ancestors from each other and facilitated genetic differentiation. Either side of the river, however, environmental and climatic conditions are unlikely to have been very different. Thus a comparison between the effects of speciation and migration on

the species has some interest for questions about the role of environment in speciation and demographic change. Both genera have also experienced periods of range contraction and expansion, and provide case-studies in the effects of environmental change on demography.

I further address questions related to the accuracy and interpretation of SMC-based methods, and their appropriateness in answering specific demographic questions. The particular histories of orangutans and chimpanzees, and their close relationship to humans, make them good species to empirically test the applicability of these methods. I end with some description and analysis of questions related to population substructure which arise as a result of trying to interpret the effective population size trajectories, though this question is looked at more systematically in the next chapter.

3.1 Orangutan demographic history

3.1.1 Background

Orangutans are the only non-human great apes found outside Africa. Of surviving hominid lineages, they are descendants of the branch which diverged earliest from the others. They are currently native only to Borneo and Sumatra, South East Asian islands which give their names to the two historically recognised species: the Bornean orangutan (*P. pygmaeus*) and the Sumatran (*P. abelii*) [41]. Recently, evidence has been put forward supporting the existence of a third species on the island of Sumatra, the Tapanuli orangutan (*P. tapanuliensis*) [102]. This evidence was largely morphological and behavioural. As will be seen below, genetic support for the classification can be ambiguous. In this chapter I most often refer to the various populations of orangutan by geographic location, and for the sake of clarity when modelling demography will largely set aside recognised species or subspecies designations. Where relevant, I shall indicate how the geographic labels relate to taxonomic classifications. Note that the designation of the new species leaves us with slightly misleading names for the previous species. I shall occasionally have to refer to "the orangutans on Sumatra" to mean collectively the Sumatran *and* Tapanuli orangutans.

The paleogeography of Sundaland Borneo and Sumatra are part of the Sunda shelf. Varying sea levels during the Pleistocene (about 2.6 Mya to 12 kya) periodically exposed currently submerged parts of the shelf [42]. Relative to other Equatorial regions there is a great difference between the area of land exposed today and the area exposed during glacial periods [169]. During the last glacial maximum (24-18 kya), for example, both islands formed part of the Sundaland landmass which connected many islands in the vicinity to mainland Asia and formed a subcontinent approximately as large as Western Europe [9]. It is unknown when the periodic passages of land connecting the islands supported rainforests rich enough to allow the spread of the largely arboreal species. Paleo-geographic evidence suggests that a

Savanna corridor ran through the middle of Sundaland during glacial periods, separating modern-day Sumatra, Borneo and Java, and possibly obstructing the migration of orangutans while facilitating the spread of *Homo* species [43, 9].

The fossil record currently supports a model in which orangutans gradually spread southwards from southern China, where they lived during the early Pleistocene (about 2.6 Mya to 780 kya), to the South East Asian islands, including Java, by the late Pleistocene (about 126 kya to 12 kya) [156]. Setting lower bounds by the fossil record, they reached Borneo at least as early as 50 kya, Sumatra 80 kya, and Java 120 kya, after which environmental changes greatly reduced forest cover in the region, leaving those islands the only remaining orangutan habitats [156]. The arrival on Sumatra precedes the earliest evidence for humans on the island by at least 10000 years [174]. Humans are thought to have played a major role in reducing orangutan population sizes in more recent times, and especially closer to the onset of the Holocene (12 kya) after the emergence of technology to hunt tree-dwelling animals and after forests were degraded for the purposes of agriculture [156].

Environmental conditions have at times varied between Borneo and Sumatra (including during periods when the islands were connected) [179]. Research into the correlations of orangutan phenotype with environment has focused on the current habitats of various populations [e.g. 86]. These conditions may not provide a representative view of the recent conditions under which orangutans evolved given the species' drastic reduction in range and the extreme environmental changes the region experienced during the late Pleistocene [156]. Nonetheless, it has been argued that the regions in Sumatra where orangutans are currently found have been historically more hospitable. For example, soil conditions have supported increased fruit production, the orangutan's primary source of food [179]. These conditions currently allow orangutans on Sumatra to live at higher population densities than those on Borneo [84].

Sumatran and Tapanuli orangutan habitats are confined to small regions in northern Sumatra. In this study, we have samples from Sumatran orangutans found in Langkat, North Aceh, and West Alas, and Tapanuli orangutans found in the Batang Toru region, their only known habitat. For a summary of the geographic locations of all the individuals in our sample, see Figure 3.2. These species are phenotypically different from Borneans in terms of their morphology, metabolism, cognition and social behaviour, and it has been suggested that some of these differences can be accounted for by adaptation to distinct environments [86]. Differences between Tapanuli and Sumatran orangutans in these aspects are smaller, but have been argued to be distinctive enough to justify different species classifications [102]. Since their current ranges are separated by around 100km [102] it is harder to explain these as environmental adaptations unless at least one of the current populations is a remnant of a population whose range extended further away. It has been hypothesised that the relatively high altitude of the rainforests inhabited by Tapanuli orangutans might have caused

adaptations not found in Sumatrans [177], but we do not know how typical this condition is in the recent history of the Tapanulis. In addition, Lake Toba, formed after the Toba volcano supereruption about 70 kya, lies between the Tapanuli and Sumatran orangutan ranges and it has been speculated that this limits gene flow between the populations [102].

Bornean orangutan habitats are spread across the island. In this study we have obtained samples from each of the three recognised subspecies: *P. p. pygmaeus* from Sarawak, *P. p. wurmbii* from Central and West Kalimantan, and *P. p. morio* from North Kinabatangan, South Kinabatangan and East Kalimantan (see Figure 3.2). These territories span a much larger total area than orangutan regions in Sumatra, and were estimated in 2008 to support roughly an order of magnitude more individual animals (approximately 54000 compared to 6500), although Borneans are thought to have been much more affected by the degradation of forests in recent years [176]. The current regions in Borneo also span a far greater range of latitudes than the orangutan territories on Sumatra.

Current knowledge of orangutan genomic evolution Research based on mitochondrial and microsatellite variation have suggested deep divisions between orangutans on Sumatra and Bornea, and have also indicated population structure on both islands correlating with regional geography and natural barriers to migration [e.g. 4, 103, 101]. The draft assembly of an orangutan genome was first published in 2011, by Locke et al [77]. They sequenced an additional 10 genomes, five each of the recognised species at the time. Their findings pointed to a greater historical effective population size among the Sumatrans, despite their lower census population size today. Reanalysis of this data has highlighted apparent discrepancies in the TMRCA of mitochondrial and autosomal sequences which some studies have suggested are significant enough to require an explanation based on sex-biased migration between regions across islands [e.g. 78]. However, there has been little formal modelling of this scenario. The relatively low sequence coverage and geographic range of this study prompted Prado-Martinez et al (2013) to collect a larger and higher quality sample in their comparative study of great ape evolution [120]. Their sample extended the geographic range of publically available orangutan sequences although it included no sequences from the Batang Toru region, in which the Tapanuli are found, and neither from the Kinabatangan and East Kalimantan regions of Borneo. This meant, for example, that only a single low coverage (6.03x) sequence of the *P. p. morio* subspecies was available. There were also too few high-coverage samples in these studies to undertake a thorough SMC-based analysis of cross-coalescence between populations, as done in this study. Despite these limitations, Prado-Martinez et al. showed interesting historical variation between the available populations and corroborated the finding of historically greater population sizes on Sumatra.

3.1.2 Results

Data In order to overcome these and related limitations, Nater et al. (2017) sampled the 36 orangutan sequences used in this study [102]. They pooled 16 new sequences with the 20 previously collected by Locke et al. [77], and Prado-Martinez et al. [120] (10 specimens from each study). Sequence coverage of the new sequences is on average greater than the coverage in the study by Locke et al. The final set of sequences, along with coverage information, is summarised in Figure 3.1. All but 5 of the sequenced genomes derive from wild-born individuals. The animals bred in captivity are nonetheless known to be first-generation offspring of wild-born animals of the same species. The ancestral locations of the remaining samples were inferred by Nater et al. using mitochondrial haplotypes, which, given the great degree of philopatry in orangutans and previous research on mitochondrial variation, is thought to strongly correlate with geographic location. The total collection consists of 16 individuals from Sumatra (including 2 from the Tapanuli species which was first classified in this study) and 20 from Borneo, spanning almost the entire present-day range of orangutans. Relevant for the analysis of ancestral gene flow between populations, the sample has 14 males. The geographic extent of the sample is shown in Figure 3.2, which was produced by Nater et al.

PCA As a summary of autosomal genetic distance between samples, Nater et al. produced the principal component analysis shown in Figure 3.3. (Further information about this technique is presented in Chapter 2.) There are three striking clusters in this analysis. These correspond to the newly proposed taxonomy of the genus, illustrated on the figure. The strongest signal of divergence, along the first principal component (which accounts for 34.2% of the variation), is between the orangutans on Borneo and those on Sumatra. The other striking difference, seen largely along the second principal component (which explains 3.6% of the variation), is between the orangutans either side of Lake Toba, namely the Tapanuli (Batang Toru) and the Sumatrans. There are also noteworthy differences between the recognised subspecies of Bornean orangutans.

Autosomal PSMC analysis I estimated historical changes in effective population size using PSMC [70]. This approach is discussed in Chapter 2, with further information provided about the application in this study in Section 3.3. The results are shown in Figure 3.4. First, observe that the plots show drastic variation between time steps and individuals in the most recent and oldest periods. These effects are likely to be artefacts, as discussed below. In the most recent time periods, the fossil record suggests that orangutan populations on both islands declined, especially since the end of the Pleistocene (from about 20 kya) [156]. This is consistent with the hypothesis that a recent decline in orangutan populations was due to human intervention. In the oldest periods the variance between individuals is also unlikely to

Pongo Species	Location	ID	Sex	Coverage	Source
P. abelii	West Alas	PA_KB4361	F	5.66	Locke et al. 2011
P. abelii	Langkat	PA_KB4661	M	4.76	Locke et al. 2011
P. abelii	Langkat	PA_KB5883	M	4.99	Locke et al. 2011
P. abelii	West Alas	PA_SB550	F	4.86	Locke et al. 2011
P. abelii	Langkat	PA_A947	F	27.39	Prado-Martinez et al. 2013
P. abelii	Langkat	PA_A948	F	23.71	Prado-Martinez et al. 2013
P. abelii	North Aceh	PA_A949*	F	27.39	Prado-Martinez et al. 2013
P. abelii	Langkat	PA_A950	F	26.28	Prado-Martinez et al. 2013
P. abelii	Langkat	PA_A952	M	21.03	Prado-Martinez et al. 2013
P. abelii	West Alas	PA_B017	F	13.74	Nater et al. 2017
P. abelii	North Aceh	PA_B018	M	16.31	Nater et al. 2017
P. abelii	West Alas	PA_B020	F	16.3	Nater et al. 2017
P. abelii	West Alas	PA_A953	F	17.78	unpubl. Prado-Martinez et al. 2013
P. abelii	West Alas	PA_A955	F	25.27	unpubl. Prado-Martinez et al. 2013
P. tapanuliensis	Batang Toru	PA_KB9528	F	5.79	Locke et al. 2011
P. tapanuliensis	Batang Toru	PA_B019	M	16.92	Nater et al. 2017
P. pygmaeus	Central Kalimantan	PP_KB4204	M	5.61	Locke et al. 2011
P. pygmaeus	Central Kalimantan	PP_KB5404	F	12.24	Locke et al. 2011
P. pygmaeus	Central Kalimantan	PP_KB5405	M	5.61	Locke et al. 2011
P. pygmaeus	Sarawak	PP_KB5406	F	4.9	Locke et al. 2011
P. pygmaeus	East Kalimantan	PP_KB5543	M	6.03	Locke et al. 2011
P. pygmaeus	Sarawak	PP_A939*	F	20.48	Prado-Martinez et al. 2013
P. pygmaeus	Central Kalimantan	PP_A940*	F	21.8	Prado-Martinez et al. 2013
P. pygmaeus	Central Kalimantan	PP_A941*	F	23.17	Prado-Martinez et al. 2013
P. pygmaeus	Central Kalimantan	PP_A943	F	24.17	Prado-Martinez et al. 2013
P. pygmaeus	Central Kalimantan	PP_A944	M	23.32	Prado-Martinez et al. 2013
P. pygmaeus	South Kinabatangan	PP_5062	M	13.81	Nater et al. 2017
P. pygmaeus	West Kalimantan	PP_A983	M	29.71	Nater et al. 2017
P. pygmaeus	East Kalimantan	PP_A984	F	29.89	Nater et al. 2017
P. pygmaeus	East Kalimantan	PP_A985	M	30.13	Nater et al. 2017
P. pygmaeus	North Kinabatangan	PP_A987	F	30.65	Nater et al. 2017
P. pygmaeus	North Kinabatangan	PP_A988	M	31.06	Nater et al. 2017
P. pygmaeus	South Kinabatangan	PP_A989	F	27.3	Nater et al. 2017
P. pygmaeus	Central Kalimantan	PP_A938*	F	18.62	unpubl. Prado-Martinez et al. 2013
P. pygmaeus	Sarawak	PP_A942*	F	23.12	unpubl. Prado-Martinez et al. 2013
P. pygmaeus	Sarawak	PP_A946	M	22.39	unpubl. Prado-Martinez et al. 2013

Fig. 3.1 Summary of orangutan sequences used. Asterisks indicate animals which were not wild-born and had their locations inferred using mitochondrial haplotype.

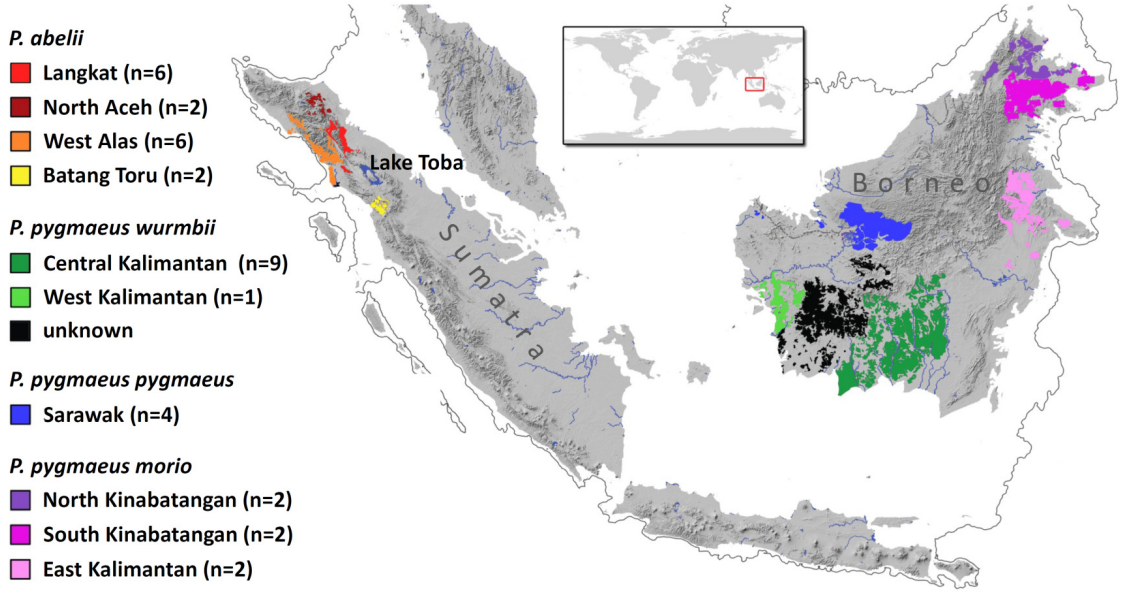


Fig. 3.2 Location and population designation of orangutans used in study. Grey outline indicates the exposed region of the Sunda shelf (the Sundaland continent) during the last glacial maximum (LGM, 24-18 kya). At its greatest extent during the Pleistocene, including during the LGM, Sundaland had a total area comparable to the area of Western Europe today [9]. Figure taken from Nater et al. (2017) [102]. Note that this version predates the proposal of the new species, so the Batang Toru are here still considered Sumatran orangutan.

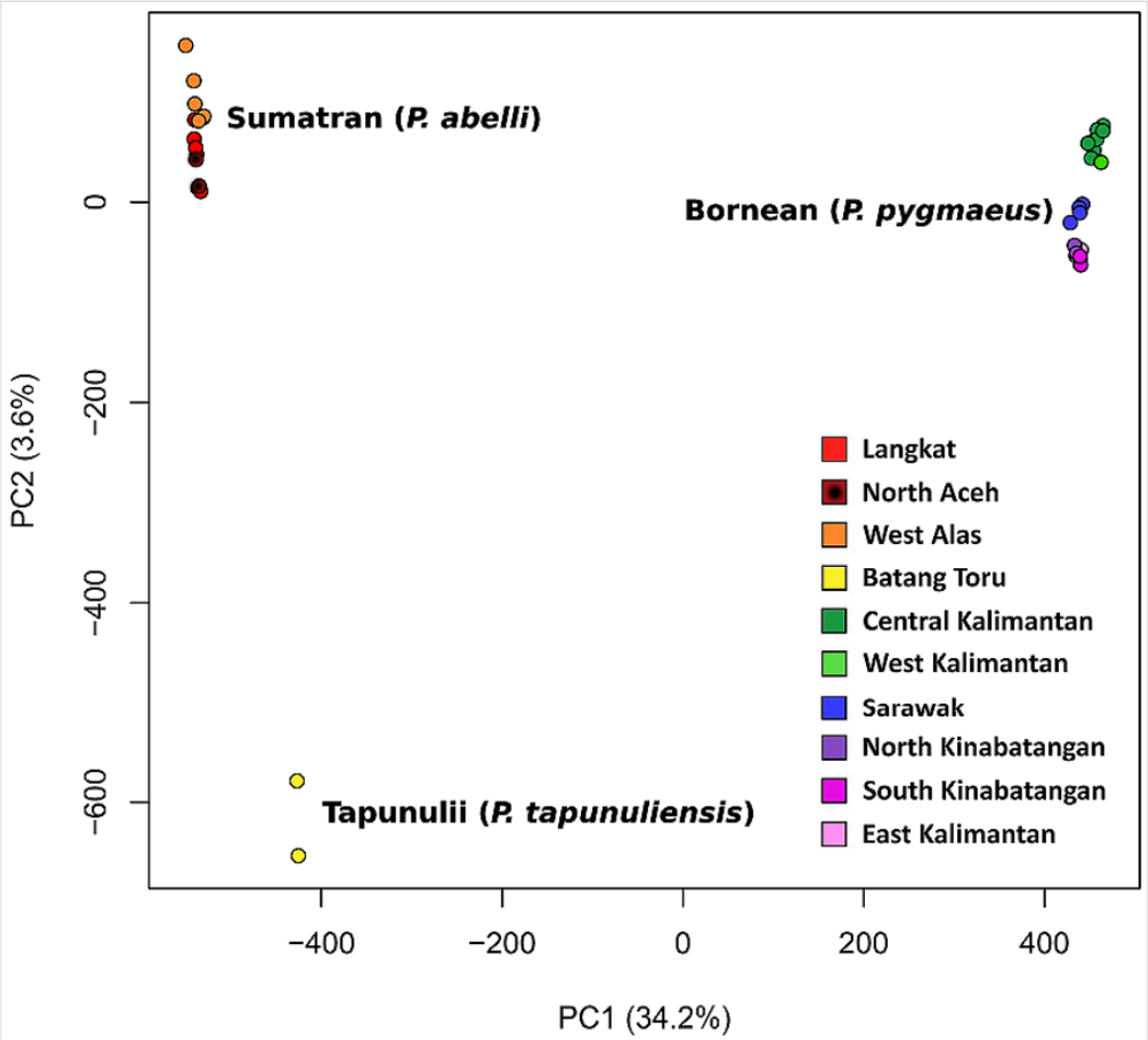


Fig. 3.3 Principal components analysis of orangutans. Image based on analysis by Nater et al (2017) [102]

be a result of genuine changes in N_e as all individuals are likely to trace the same proportion of their ancestry to the same populations, but the inferred values shown in the plots at this time period vary considerably by individual. The increased variance between individuals is thus most likely caused by the sparsity of coalescent events in the very recent and very distant pasts. This leaves PSMC with too few data points to draw reasonable inferences during these periods. Similar effects have been seen in human populations and other great apes [e.g. 70, 24].

From the present until 500 kya, the inferred N_e trajectories show clear differences in demographic history of the orangutans found on different islands. The populations on Sumatra appear to have historically higher effective population sizes, and also undergo fewer bottlenecks. The only noticeable decline occurs at around 80-90 kya and lasts until about 20 kya. This decline appears to affect all the individuals on Sumatra, regardless of population location or species designation. It is difficult to distinguish the Tapanuli (Batang Toru) individuals from the Sumatrans on this analysis. The Bornean populations undergo two marked bottlenecks. The first occurs at around 300-500 kya, and the second occurs 50-150 kya. These bottlenecks, and the lack of noticeable recovery, leave the Bornean orangutans with historically lower population sizes than the Sumatrans for most of the period in which their differences are distinct. The fact that the Borneans never recover their previous population size after the bottlenecks, suggests that the cause of the loss of population might be range restriction. The decline in the Sumatran population in the late Pleistocene (about 12-126 kya) is not as sudden as previous PSMC analyses have shown [120, 102]. It was suggested that a sudden decline in population size around 70 kya might have been caused by the Toba eruption. Since PSMC has been shown to smear out drastic sudden changes in population size [70], the gradual change in this particular analysis is not necessarily inconsistent with previous plots. Differences in filtering and variant calling in other studies may have produced sufficient differences to cause this discrepancy. I return to a discussion about sequence quality and the effect of variant calling methods below. Observe also that the Sumatran island individuals exhibit greater variation in their effective size trajectories in this recent period. This might be caused by natural variation in N_e , due possibly to population structure [87], or methodological variation caused by the low density of coalescent events, as mentioned above.

In Figure 3.5, all populations are plotted together. Here I have selected only the high-coverage (>20x) sequences. Observe that variation between individuals within respective islands is significantly reduced relative to the previous plots, suggesting that poor sequence quality acted to decrease the accuracy of inference. However, note that a single Sumatran individual's curve follows a significantly different path to the others, in that between 30 kya and 100 kya is has a much lower inferred effective population size than the others. It is also follows the atypical trajectory in the most recent displaye time interval, shooting upwards

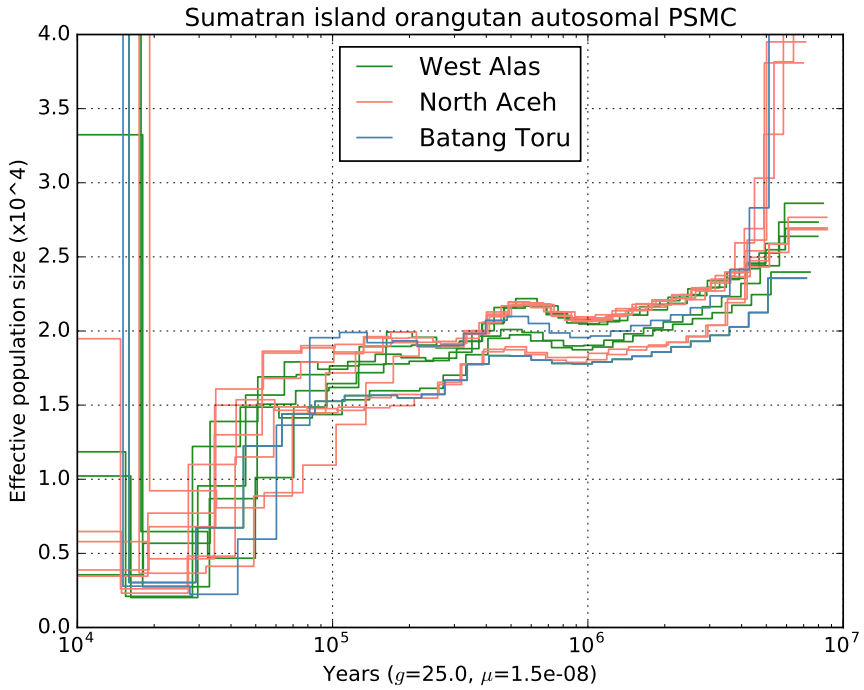
before the other Sumatran curves. I cannot explain this discrepancy, since while the individual (PA-A952) does have lower sequence coverage than the others in this plot (21x), it is not very much lower than the next lowest coverage (PA-A948 with 24x coverage). In addition, it was not born in captivity, and was sequenced as part of the original Prado-Martinez et al (2013) study, so there is no more reason to be sceptical of its provenance or recent ancestry.

Without filtering out the lower coverage sequences, it appears that the period during which effective population sizes on Borneo and Sumatra are distinct lasts from the most recent observable times until approximately 500 kya. Restricting ourselves to the higher coverage samples, we see that the inferred population sizes are distinct earlier than this, with the Borneans having higher estimated values. Given the wide geographic range over which ancestral populations might have lived, this distinction would be consistent with old (earlier than 500 kya) restrictions in gene flow between these populations. Some discussion of this possibility is provided in the following section, but I do not explore this further.

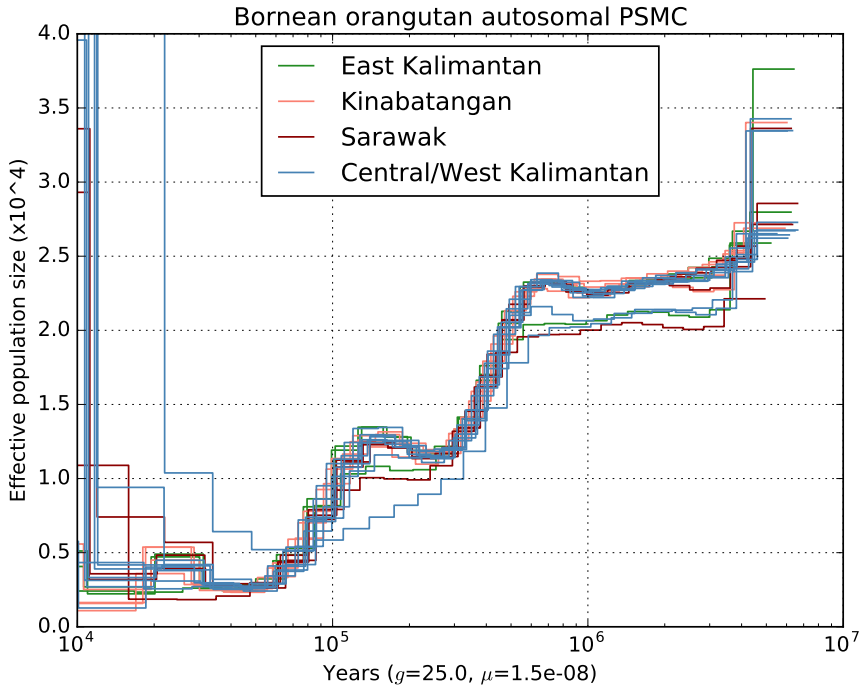
In addition, even if we assume that their ancestors had the same effective population size earlier tha 500 kya, it would not be possible to determine the separation time between ancestral Borneo and Sumatran populations using this information alone. Under this approximation, reading from the past to the present (right to left), the period during which the trajectories start to differentiate can be interpreted as the time during which their ancestral populations began to diverge. In the simplest model, in which a single population bisected and the two subpopulations gradually separated, this can be read as a lower bound on the time at which divergence would have occurred. This is the case because the times at which N_e is distinct reflects periods during which the ancestral populations were exchanging few migrants. When N_e is the same, however, it might merely be the case that their ancestral populations are distinct but have the same population size. Thus we cannot use this analysis alone to produce an upper bound on the time when the populations would have started their divergence. The time when the trajectories appear to start their divergence, however, coincides with the earliest of the two major bottlenecks seen in the Bornean orangutans. This would suggest that divergence between ancestral orangutan populations, and eventually speciation, occured due to isolation of smaller populations of orangutans on Borneo.

For a discussion on bootstrapping and uncertainty, see Methods (Section 3.3).

Ancestral gene flow analysis with MSMC2 In order to explore ancestral gene flow between populations we apply a strategy first used by Li and Durbin (2011) [70]. As discussed above and in Chapter 2, PSMC infers the distribution of pairwise coalescent times from a single diploid sequence. Li and Durbin realised that if you paired up one haplotype each from phased sequences drawn from two distinct populations, then you could run PSMC on this so-called pseudodiploid sequence. What is inferred in this case is not the N_e trajectory of either population (nor that of the meta-population), but a measure of the rates of cross-



(a) Sumatran historical N_e



(b) Bornean historical N_e

Fig. 3.4 Autosomal PSMC analysis of historical effective population sizes (N_e) of (a) orangutans on Sumatra (*Pongo abelii* and *Pongo tapanuliensis*) and (b) orangutans on Bornea (*Pongo pygmaeus*). Each line corresponds to a single individual, and colours correspond to regions in which the individuals were found or in the case of captive-bred animals, the regions where their parents were inferred to originate from. The Tapanuli orangutans are referred to by their geographic location "Batang Toru".

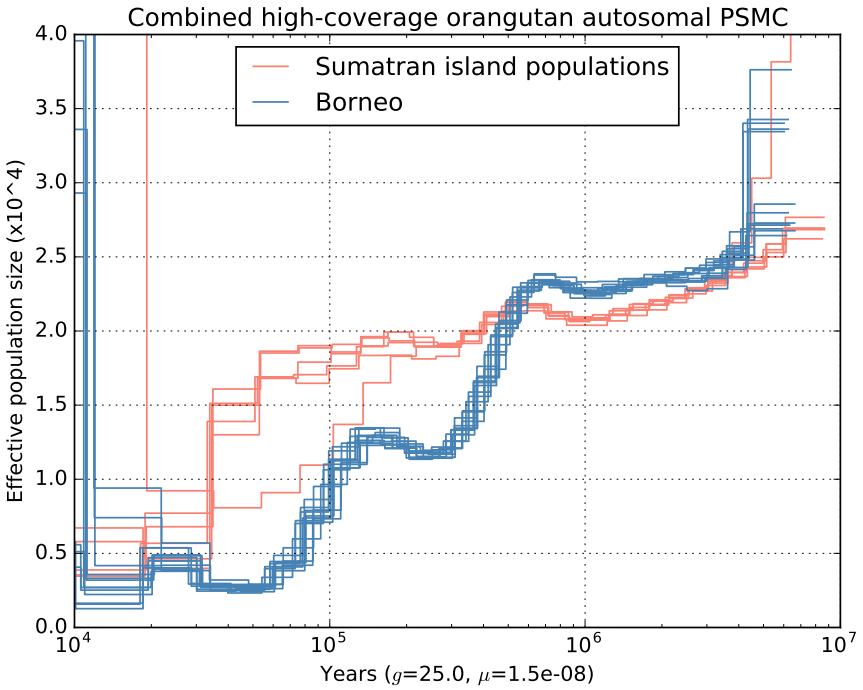


Fig. 3.5 PSMC analyses of individuals with high sequence coverage. "Sumatra" refers to high-coverage sequences from all populations on Sumatra, including the Tapanuli orangutans.

coalescence between the ancestral populations. To see how this allows us to infer temporal patterns of gene flow between populations, observe that if no interbreeding occurs between two populations at some time T, then the rate of coalescence at that point should be zero, and the inferred N_e effectively infinite. On the other hand, when two ancestral populations at T are effectively a single unit, then N_e should be the same as that inferred on a usual PSMC analysis of any individual in either population. Intermediate levels of "cross-coalescence" should show up as intermediate values of N_e . This technique is a powerful look into the timing and process of separation of populations, and even, in relevant cases, of speciation. It bears mentioning however, that when this technique is applied, PSMC is not implementing an explicit demographic model of diverging populations. The historical implications of the technique should be thought of as a *post hoc* interpretation of the result (see Chapter 2).

To avoid bias induced by errors resulting from the phasing of low coverage sequences, we focused only on male X chromosomes. Outside of the pseudoautosomal regions, these can be seen as naturally occuring haplotype sequences. Although they may introduce bias related to sex-specific demographic process, such as sex-biased migration, these effects are likely not great enough to affect the broad inferences we draw in this section. As discussed in the Methods below, this will require an adjustment to a sex-chromosome specific mutation rate when scaling the curves to non-coalescent units.

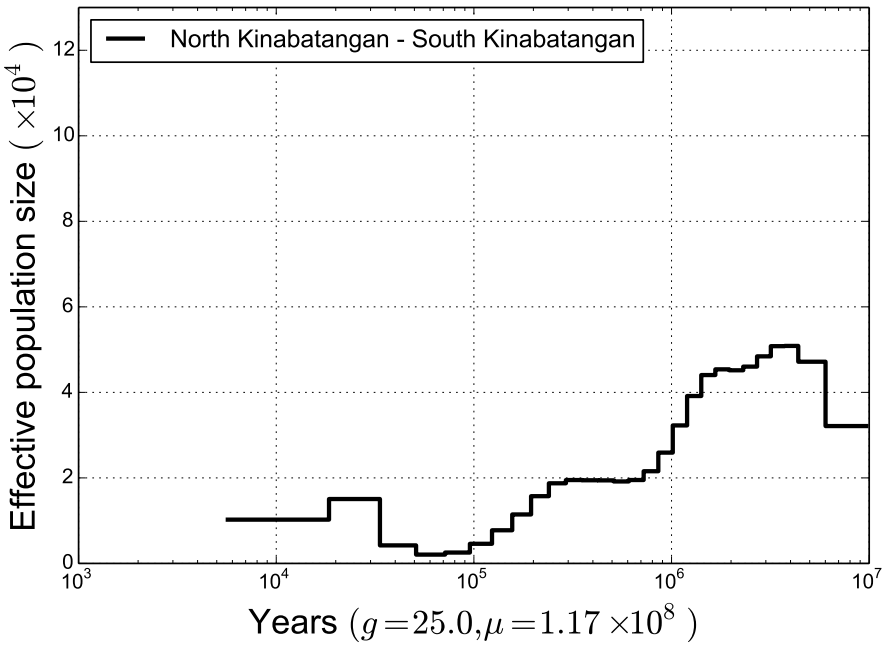


Fig. 3.6 Cross-coalescence analysis of Kinabatangan populations.

Inference of ancestral gene flow was carried out using the multiple sequential Markovian coalescent (MSMC2) model [172]. The method is described in Chapter 2. It allows us to jointly analyse male X chromosomes from any pair of populations and estimate ancestral gene flow between the two. In Section 3.2 I compare MSMC2 and PSMC inferences and illustrate the relationship between them.

To illustrate how the approach works, we show what is effectively a "null" run using data from the males in North and South Kinabatangan. These are two regions in Borneo containing the same subspecies (*P. p. morio*, see Figure 3.2). The orangutans here are genetically similar relative to other Bornean orangutans, as shown in the PCA in Figure 3.3, and also geographically close. Thus we might guess that they have shared a history of interbreeding. Setting aside the scaling of the curves, this is what is shown in Figure 3.6. Here, the N_e curve shows the two distinctive bottlenecks that we saw in all the Bornean autosomal PSMC analyses. This suggests that the sampling of haplotypes was chosen from a Bornean population without significant recent substructure. However, the observation of similarity in these curves is only true at a qualitative level. There is considerable uncertainty in the estimate of the X chromosome mutation rate and the current value would need to double in order to produce similar quantitative values as the autosomal estimates (see Figure 3.24 and the related discussion).

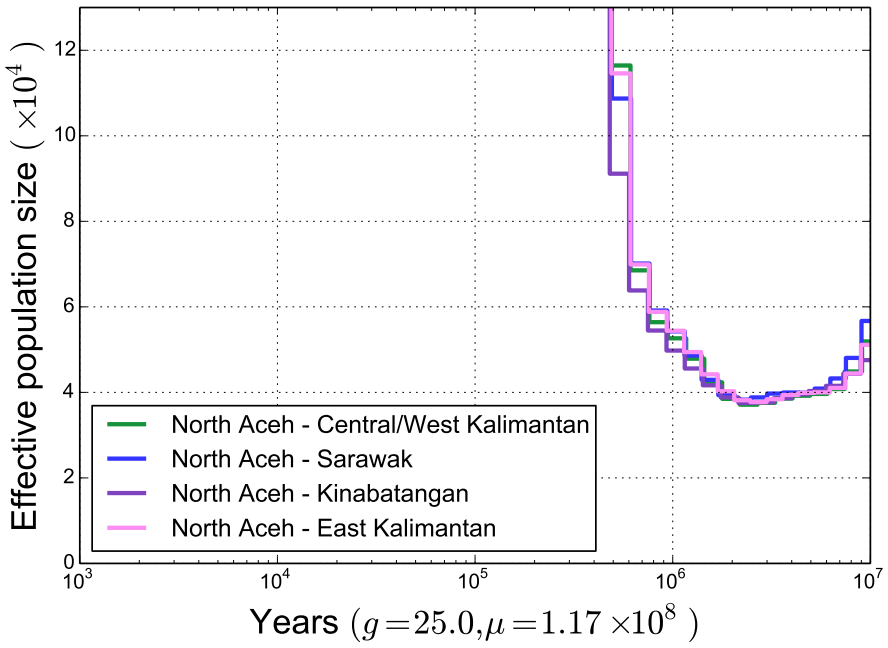


Fig. 3.7 Cross-coalescence analysis of North Alas and Bornean populations.

In contrast, however, the following two plots, Figures 3.8 and 3.7, illustrate the cross-coalescence curves when one group of males is from a Bornean population and the other is either from North Aceh or Batang Toru, regions where the two different species on Sumatra live. No other males from Sumatra were in our sample. Here we see that the inferred "effective population size" is very different to that inferred in the previous analyses. It shows a signal of divergence between Bornean and Sumatran ancestral populations by the gradual increase to an effectively infinite N_e between 1.5 Ma and 500 ka. This is seen in all analyses pairing the Northeast Alas and Batang Toru populations with the various Bornean populations. As noted above in the autosomal PSMC analyses, a sudden cessation in gene flow between the populations may be smeared out over a larger time span, so the actual divergence between populations might have occurred relatively rapidly at some point between these bounds. We detected no gene flow between the two orangutan species at any more recent time. Note that this pattern is very different to that seen in the Kinabatangan populations and is not comparable, even qualitatively, with any of the autosomal estimates.

The similarities between the various Bornean populations in each plot suggest either of two possibilities: at the time of divergence the Borneans ancestors were a single interbreeding population; or, if there were multiple subpopulations exchanging few migrants, then the extant populations derive similar proportions of their ancestry from each of them. It suggests that differentiation between the subspecies and regional populations occurred more recently than

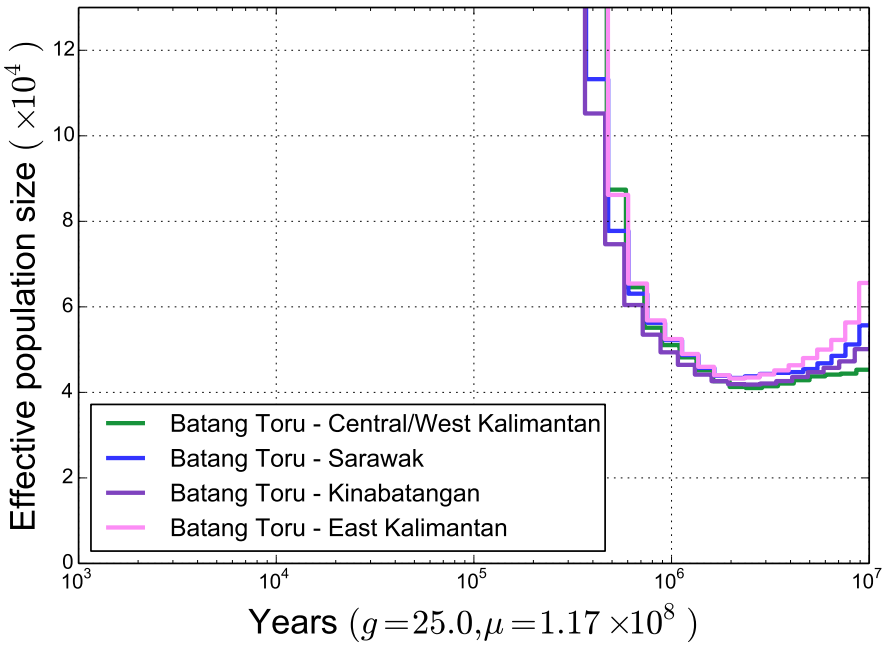


Fig. 3.8 Cross-coalescence analysis of Batang Toru and Bornean populations.

500 kya. In Figure 3.9 we show the cross-population coalescence rates in Borneo. Between 20 kya and 40 kya, we observe signals of divergence between most of the Bornean populations. It appears that Sarawak and East Kalimantan orangutans diverged more recently from each other than the East Kalimantan diverged from the Kinabatangan (20 kya compared with 30 kya). This is remarkable given that the latter pair of populations are considered to constitute the same subspecies, whereas the former are not. However, due to the relative paucity of coalescent events during this period, it is difficult to resolve accurately the order in which the populations diverged from each other. Complex divergence processes would also tend to obscure times of initial separation. This might be the case if populations, for example, diverged very gradually in the presence of ongoing gene flow, or if gene flow was mediated via a third population. As with the Kinabatangan, we observe gene flow between the Sarawak and Central/West Kalimantan populations until recent times. While less surprising from a geographic point of view, it is nonetheless unexpected given their different subspecies classification.

Observe that the two Borneo-Sumatra divergence plots are very similar. The cessation of gene flow between the Northeast Alas and all Bornean populations, and between the Batang Toru and all Bornean populations, occurred at similar times. It appears as though the ancestors of the Batang Toru may have diverged slightly more recently from the ancestors of the Borneans, though that difference may not be significant, especially since this analysis

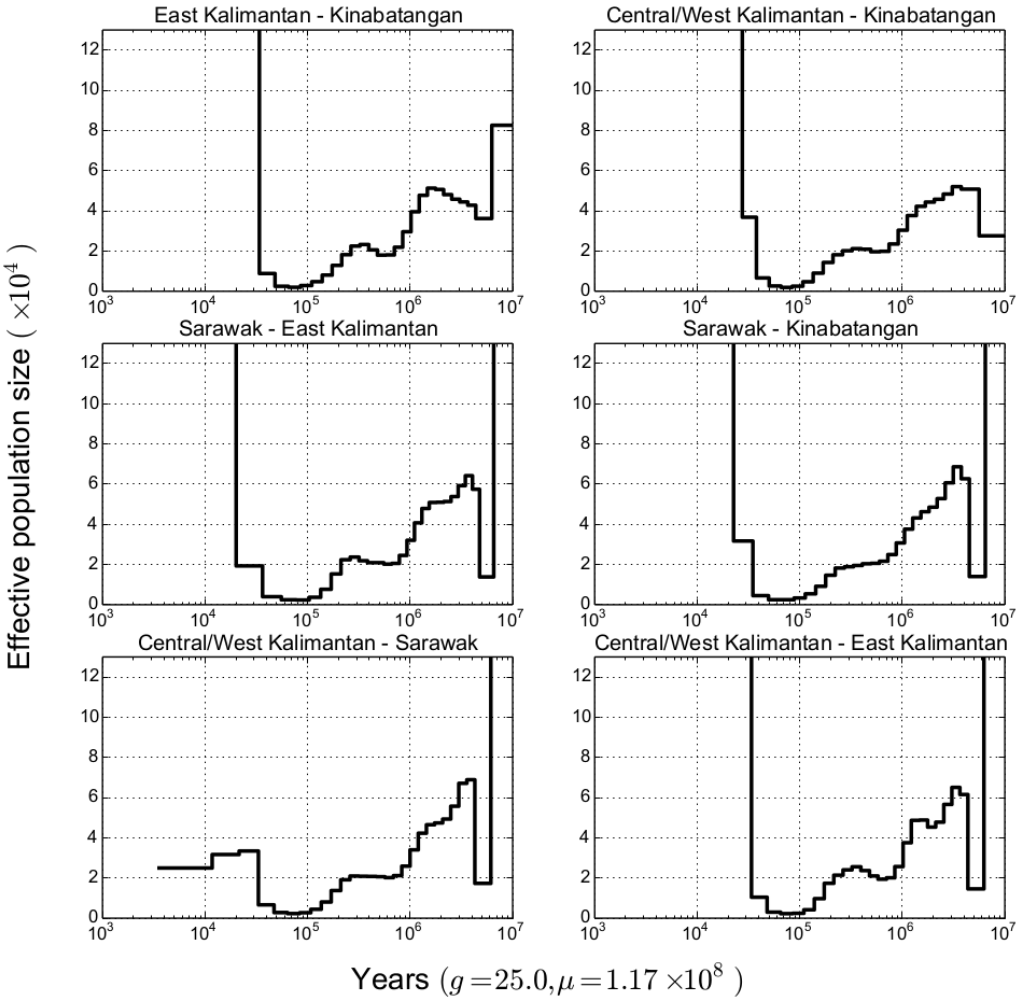


Fig. 3.9 Cross-coalescence analysis of Bornean populations.

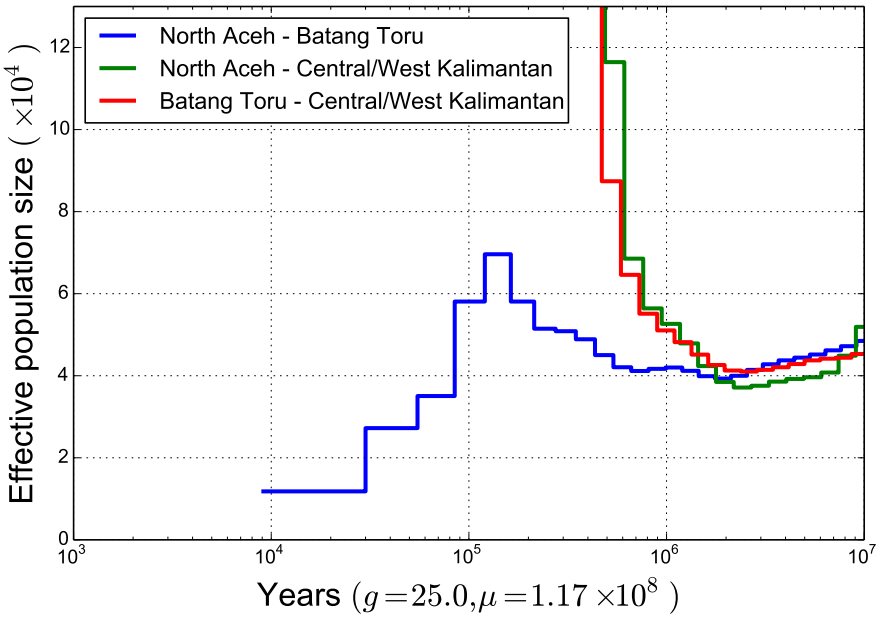


Fig. 3.10 Cross-coalescence analysis of Sumatran populations.

was run on only a single Batang Toru male. The fact that the times are similar in the two plots suggests that the Batang Toru and Northeast Alas populations have close histories. Most likely there was substantial gene flow between ancestral Batang Toru and Northeast Alas populations during the period of divergence from Bornean populations. This would be consistent with their PCA projections, since the variance explained by PC1 is 34.2% and the Sumatran island populations are distinctly separated from the Bornean populations along this principal component. Further evidence of this gene flow is provided by a cross-population analysis of the Northeast Alas and Batang Toru, Figure 3.10, indicating continuous gene flow at least until the most recent point in time the method can detect. It does however show a gradual reduction in gene flow between the Batang Toru and North Aceh populations from around 1 Mya to 100 kya, followed by a subsequent increase of gene flow to more recent times. It is uncertain what degree of gene flow over this length of time might induce the degree of genetic differentiation associated with an apparent speciation event. However, it is not clear that we would be able to disintinguish the curve observed here from a single autosomal PSMC analysis of a Sumatran individual. In other words, this analysis does not support a scenario in which the Batang Toru and Northeast Alas were historically distinct populations. In Chapter 2 there is a brief discussion on the implications of gene flow for various concepts of species and the process of speciation. Under the species concept used by Nater et al this signal of gene flow did not disqualify the populations from consideration as distinct species.

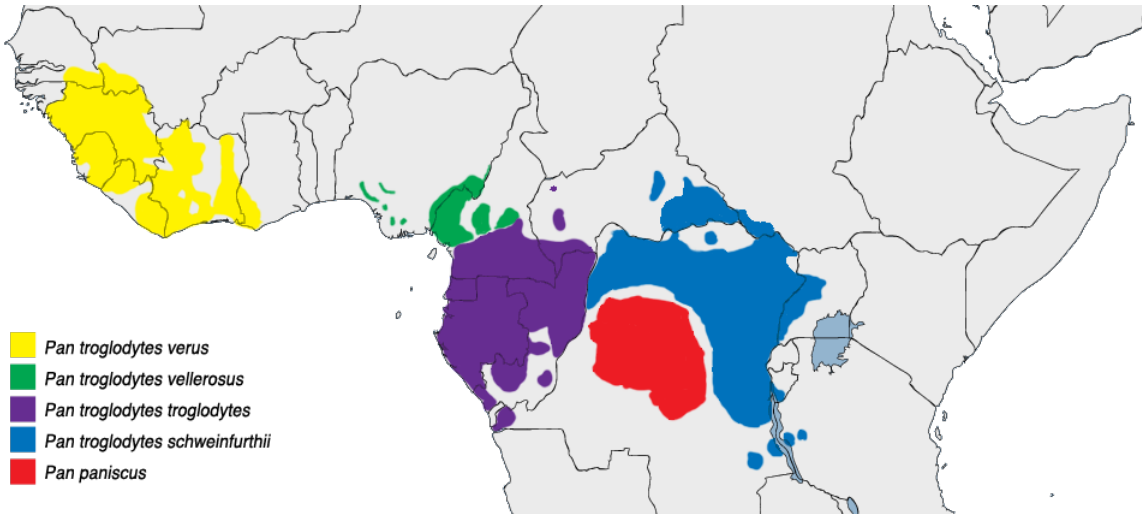


Fig. 3.11 Map of *Pan* species and subspecies ranges. Note subspecies name *vellorosus* is now designated *ellioti*. Image credit: Wikimedia Commons.

3.2 Chimpanzee and bonobo demographic history

3.2.1 Background

The *Pan* genus is phylogenetically closer to hominins than to orangutans [37, 64, 120]. Two extant *Pan* species are recognised, the common chimpanzee (*P. troglodytes*, also just "chimpanzee") and the bonobo (*P. paniscus*, historically the "pygmy chimpanzee"), and the chimpanzees consist of four recognised subspecies: Eastern (*P. t. schweinfurthii*), Central (*P. t. troglodytes*), Nigeria-Cameroon (*P. t. ellioti*), and Western (*P. t. verus*) chimpanzees [41, 111]. Living in dense tropical rainforests, members of the genus are currently spread across several Central and West African countries. Deforestation and hunting have led to a drastic reduction in their population sizes and they are both currently classified as endangered [54].

Biogeography and *Pan* evolution The Congo River, the second largest river by volume after the Amazon [138], forms a natural boundary today between the habitats of chimpanzees and bonobos [20, 164]. Neither species is known to swim. The right bank of the river marks the outer limit of the territories of Eastern and Central chimpanzees, while bonobo habitats reach only as far as the left bank. (In the relevant regions the right bank is largely north and west of the river.) The ranges of species and subspecies are shown in Figure 3.11. This division in the geographic distribution of the genus is thought to play a key role in its evolutionary history. It has been widely speculated, for example, that the formation of the Congo River was the main cause of bonobo-chimpanzee speciation, since the newly-forming

river is believed to have bisected the range of an ancestral *Pan* species and limited gene flow between the resulting populations on either side[e.g. 14, 164].

However, there has been little discussion of the geographic and evolutionary conditions under which this scenario could be supported. Takemoto et al. (2015) raised four major questions about the role that the Congo River played in bonobo-chimpanzee speciation [161]: (1) whether ancestral *Pan* was found on both sides of the river; (2) whether *the formation* of the Congo River was the proximate cause of the division of the bonobo and chimpanzee ancestral populations; (3) whether, or during which periods, the river formed an imsurmountable barrier to gene flow; and (4) whether ancestral *Pan* was adapted to savannah regions and could have had a much larger range than its extant descendents. Depending on the answers to these questions they lay out several possible hypotheses about the process of speciation, including: that the newly-forming river bisected the ancestral *Pan* range into discrete territories causing speciation; that the river is old but that a migration of proto-*Pan*, either from the North or the West, divided the ancestral populations either side of it, after which forests contracted and migration decreased; that the ancestors lived on the right bank of the river and that during various arid periods, water levels were low enough that the river could be crossed, but that for extended periods this was rare and allowed speciation to occur.

Although there is little evidence in the fossil record, some findings from paleogeography weigh in against several of these hypotheses. The first recognised ancestral chimpanzee fossil was reported in 2005 [88], at the same time as the draft assembly of a chimpanzee genome was first published [173]. The fossil was dated to the Middle Pleistocene ($\approx$ 500 kya), and was found in the Rift Valley, far east of the current chimpanzee range. This provides evidence of extensive migration by ancestral members of the genus, but since the fossil is more recent than the hypothesised speciation time (see below) it is largely uninformative about any migrations which could have led to speciation. The ancient range might, for example, be the result of a post-speciation eastward migration of ancestral chimpanzees. There is also little evidence that common ancestors of chimpanzees and bonobos were adapted for savannah habitats. As a result, there is little reason to think that their habitats were much different than in recent millenia, although a limited fossil record leaves this uncertain. The changing distributions of forests is also poorly understood.

In addition to proposing the hypotheses above, Takemoto et al. review evidence on the age and evolution of the Congo River: in the last twenty years, hydrological estimates have put the age of the current form of the river as at least 5 My [e.g. 138, 158], while sedimentary evidence suggests the river formed at least 16 Mya [e.g. 3]. Both these estimates are much older than the apparent chimp-bonobo speciation time (see below). Takemoto et al. also suggest that aridity during glacial periods made crossings possible, although the locations of the possible crossings are hard to determine with current evidence [161]. As a result, they

favour the last of the hypotheses, and rule out the possibility that gene flow stopped as a result of the formation of the Congo.

The non-overlapping regions over which chimpanzee subspecies range are spread across several countries in Central and Western Africa (see Figure 3.11). North of bonobo territory is the region in which Eastern chimpanzees are found. Mimicking the local border between the two Congo nations, the territory is separated from that of Central chimpanzees, to their west, by the Ubangi River. This additional example of populations being separated by a river is further evidence of water systems in the region affecting animal demography, though there is as yet little we can say about gene flow across this barrier. Further west and to the north are the regions of the most recently recognised subspecies, the Nigeria-Cameroon chimpanzee, and west of that the region of the Western chimpanzee. These populations are found relatively far from the other chimpanzees. It appears from their current distributions that their ancestral populations were confined to refugia, and separated into from other populations during drier periods, when forest ranges contracted. Though the history of changing forest cover, as mentioned above, is poorly understood, as is the history of gene flow between subspecies. No ancient DNA from the ancestral *Pan* genus has been published. As such, it is difficult to know how far current geographic distributions reflect historical territories, though chimpanzees and bonobos are not as migratory as humans.

Current knowledge of *Pan* genomic evolution As with the orangutans, several mitochondrial DNA (mtDNA) and microsatellite studies of chimpanzees and bonobos have been undertaken. Phylogenies constructed on the basis of mtDNA show chimpanzees and bonobos forming respective monophyletic clades, and thus support the major division of species [53]. While most of the studies which focus on chimpanzees support an early division within the species between the Eastern and Central populations, and the Western and Nigeria-Cameroon populations, it was not initially clear that phylogenies in the latter clade supported the subspecies designation of the Nigeria-Cameroon population [6]. Larger samples in recent mtDNA and microsatellite studies have however shown this secondary division within the major chimpanzee clades [76, 10, 36], with results largely coinciding with taxonomy. Recent work has also allowed some demographic modelling of the species history, showing distinct demographic histories of chimpanzees and bonobos [53]. Within bonobos, mtDNA-based analyses have been interpreted as showing smaller rivers acting as barriers to gene flow, resulting in detectable population structure [57, 162].

As mentioned in Chapter 2, and previously in this chapter, there is greater uncertainty in demographic conclusions drawn from small numbers of loci. The draft assembly of the chimpanzee genome was published in 2005 and allowed the study of variation at the genomic level [173]. The bonobo genome was first published in 2012 [123]. Prado-Martinez et al. (2013) looked at a sample of chimpanzee and bonobo genomes from across the species' ranges [120].

They constructed a neighbour-joining phylogenetic tree and ran FRAPPE and PCA analyses on their sample, which supported long-standing divisions between chimpanzee subspecies. Indeed, their results support a deeper historical division between Nigeria-Cameroon and Western chimpanzees than between Eastern and Central chimpanzees. This is striking given initial difficulties in showing the significance of the division in the former clade using mtDNA based on larger numbers of individuals. Nonetheless, Prado-Martinez et al. also provide evidence of recent gene flow between Nigeria-Cameroon and Eastern chimpanzees. They ran PSMC on their samples, showing comparatively high historical N_e in the chimpanzees relative to the other great apes, and distinct historical N_e trajectories within the subspecies. However, they had too few male sequences to perform a population-level cross-coalescence analysis of the sort done below.

Most recently, de Manuel et al. (2016) collected a larger sample of chimpanzee sequences and produced a more thorough analysis of *Pan* demography which this study formed part of [24]. Their central finding has to do with gene flow between the species. Chimpanzees and bonobos have been reported to hybridise in captivity [167], but no example of interbreeding has been observed in the wild. Initial tests for interbreeding between bonobos and chimpanzees, reported when the bonobo genome was first published, supported the scenario of no inter-species gene flow after their ancestral divergence [123]. Several lines of evidence put forward by de Manuel et al. did suggest however that gene flow occured between chimpanzees and bonobos at some ancient time.

3.2.2 Results

Data de Manuel et al. sequenced 40 new chimpanzee whole genome sequences. Pooled with *Pan* samples sequenced in previous studies (see Methods), the total dataset consisted of sequences from 69 individuals, including 10 bonobos. The individuals were found in sanctuaries in Europe and Africa. The pair of chimpanzees which contributed the samples out of which the reference genome was assembled were not wildborn, and one was excluded in several analyses because it was a hybrid of two different subspecies (see below). There were 24 male chimpanzees and 2 male bonobos in the sample. The dataset includes new sequences from each subspecies of chimpanzee. There exists considerable geographic information about most of the newly sequenced chimpanzee individuals, though that information will not be used here. It is not known where the bonobos were recovered from, but they are known to be wildborn. Mean sequence coverage across the sample was 25x. Sequence information is summarised in Figures 3.12, 3.13, and 3.14.

PSMC analysis of historical N_e In Figure 3.15 I show the output of autosomal PSMC analyses run across the genus and on each individual in our sample. Broadly, all the analyses tend to show an aggregate decline in effective population size from about 8 Mya until at

	Nigeria-Cameroon	Central	Western	Eastern	All
Number of samples	10	18	12	19	59
Number of males	4	6	6	6	22
Average coverage	17.25	23.08	26.5	30.86	24.42

Fig. 3.12 Summary of chimpanzee sample distribution and average sequence coverage

least 1.5 Mya. After this, N_e in each population recovers, but eventually, after one or two further periods of decrease and increase, all decline to their lowest values in the periods closest to the present. If the oldest, genus-wide, decline in N_e entirely reflects a decrease in census population size, it might be due to a regional contraction in forest cover or pandemic disease, followed either by expansions in forest range or recovery and resistance to the disease. This could also be the result of mergers in ancestral populations which had previously been sharing minimal migrants. This second possibility is discussed more below and in Chapter 4.

It is striking that individuals drawn from the same recognised species or subspecies, tend to exhibit historical N_e curves more like each other than individuals drawn from different populations. (A few exceptions are mentioned below.) Historical N_e trajectories correlate with the current genus taxonomy, even supporting the previously contentious hypothesis that the Nigeria-Cameroon chimpanzees have a distinct population history. Indeed, a naive construction of population phylogeny based on this analysis, in which it is assumed that inferrred N_e divergence corresponds straightforwardly with population divergence, would produce a phylogeny close to that based on direct phylogenetic methods [eg. 24]: the earliest separation is between the ancestral bonobo and the ancestral chimpanzee population; within chimpanzees, the Western and Nigeria-Cameroon subspecies (the "Western clade") split earliest from the others, after which the pair soon began to diverge from each other; Eastern and Central chimpanzees (the "Eastern clade") track each other closely for several more tens of thousands of years before also separating.

Recall the discussion in Section 3.1.2 on the implications of these analyses for population divergence times. From about 2.5-3 Mya the bonobo N_e curves first tend to be distinguishable from the chimpanzee curves. Divergence between the two ancestral populations would thus have occured as least as early as that. Immediately after this, it is the ancestors of chimpanzees which decline more rapidly in effective population size. They possess a lower N_e than bonobos for as much as the first half of their histories as distinct populations (the time axis is in log scale). This is surprising given the relative genetic diversity and geographic range of chimpanzees today. For example, the hypotheses proposed by Takemoto et al. (2015), discussed above, are more consistent with a small bonobo founding population than a small chimpanzee one [161]. The bonobo N_e recovery after the genus-wide initial decrease peaks at around 150 kya, although for most of the second half of their history after divergence they have a lower N_e than chimpanzees. This more recent history is consistent with the

ID	Subspecies	Origin	Sex	Phase
Akwaya_Jean	Pan_troglodytes_ellioti	Western Cameroon	M	1
Banyo	Pan_troglodytes_ellioti	Western Cameroon	F	1
Basho	Pan_troglodytes_ellioti	Western Cameroon	M	1
Damian	Pan_troglodytes_ellioti	Western Cameroon	M	1
Julie	Pan_troglodytes_ellioti	Western Cameroon	F	1
Kopongo	Pan_troglodytes_ellioti	Western Cameroon	F	1
Koto	Pan_troglodytes_ellioti	Western Cameroon	M	1
Paquita	Pan_troglodytes_ellioti	Western Cameroon	F	1
Taweh	Pan_troglodytes_ellioti	Western Cameroon	F	1
Tobi	Pan_troglodytes_ellioti	Western Cameroon	F	1
100037_Vincent	Pan_troglodytes_schweinfurthii	Tanzania-Gombe National Park	M	1
100040_Andromeda	Pan_troglodytes_schweinfurthii	Tanzania-Gombe National Park	F	1
9729_Harriet	Pan_troglodytes_schweinfurthii	Uganda-West	F	1
A910_Bwambale	Pan_troglodytes_schweinfurthii	Uganda-West	M	1
A911_Kidongo	Pan_troglodytes_schweinfurthii	DRC	F	1
A912_Nakuu	Pan_troglodytes_schweinfurthii	DRC	F	1
A996_Diana	Pan_troglodytes_schweinfurthii	DRC-South	F	2
B002_Padda	Pan_troglodytes_schweinfurthii	DRC-Central	M	2
B007_Cindy	Pan_troglodytes_schweinfurthii	Uganda-West	F	2
B010_Ikuru	Pan_troglodytes_schweinfurthii	DRC	F	2
N013_Tongo	Pan_troglodytes_schweinfurthii	Rwanda	M	2
N015_Cleo	Pan_troglodytes_schweinfurthii	Zambia	F	2
N017_Bihati	Pan_troglodytes_schweinfurthii	DRC	F	2
N018_Trixie	Pan_troglodytes_schweinfurthii	DRC-east	F	2
N019_Maya	Pan_troglodytes_schweinfurthii	DRC-South	F	2
A957_Vaillant	Pan_troglodytes_troglodytes	Gabon-East	M	1
A958_Doris	Pan_troglodytes_troglodytes	Gabon-West	F	1
A959_Julie	Pan_troglodytes_troglodytes	Gabon-East	F	1
A960_Clara	Pan_troglodytes_troglodytes	Gabon	F	1
9668_Bosco	Pan_troglodytes_verus	NA	M	1
A956_Jimmie	Pan_troglodytes_verus	NA	F	1
A991_Berta	Pan_troglodytes_verus	Ivory coast	F	2
A992_Annie	Pan_troglodytes_verus	Guinea	F	2
A993_Mike	Pan_troglodytes_verus	Guinea	M	2
B005_SeppToni	Pan_troglodytes_verus	Liberia	M	2
B006_Linda	Pan_troglodytes_verus	Liberia	F	2
Clint	Pan_troglodytes_verus	Captive Born	M	1
N014_Cindy	Pan_troglodytes_verus	Ivory coast	F	2
N016_Alice	Pan_troglodytes_verus	Ivory coast	F	2
X00100_Koby	Pan_troglodytes_verus	NA	M	1
B025_Marlin	Pan_troglodytes_troglodytes	NA	F	2
B024_Negrita	Pan_troglodytes_troglodytes	Equatorial Guinea	F	2
B023_Blanquita	Pan_troglodytes_troglodytes	Equatorial Guinea	F	2
B022_Tibe	Pan_troglodytes_troglodytes	Equatorial Guinea	M	2
B021_Yogui	Pan_troglodytes_troglodytes	Equatorial Guinea	M	2
A990_Noemie	Pan_troglodytes_troglodytes	Equatorial Guinea	F	2
B014_Coco	Pan_troglodytes_schweinfurthii	Zambia	F	2
B013_Athanga	Pan_troglodytes_schweinfurthii	DRC-North	M	2
B012_Washu	Pan_troglodytes_schweinfurthii	DRC-South	M	2
B011_Frederike	Pan_troglodytes_schweinfurthii	Rwanda	F	2
10964_Cindy	Pan_troglodytes_troglodytes	NA	F	2
11352_Mirinda	Pan_troglodytes_troglodytes	NA	F	2
12311_Ula	Pan_troglodytes_troglodytes	Equatorial Guinea	F	2
12320_Lara	Pan_troglodytes_troglodytes	Equatorial Guinea	F	2
12348_Luky	Pan_troglodytes_troglodytes	Equatorial Guinea	M	2
12420_Gamin	Pan_troglodytes_troglodytes	NA	M	2
13656_Brigitta	Pan_troglodytes_troglodytes	Reunion Island	F	2
11528_Alfred	Pan_troglodytes_troglodytes	NA	M	2

Fig. 3.13 Identifiers and regional designations of chimp individuals. Phase 1 individuals were sequenced by Prado-Martinez et al. [120] and phase 2 by de Manuel et al. [24].

Name	Sex	Sequence Coverage
Hortense	F	41.66
Kosana	F	43.06
Dzeeta	F	48.15
Hermien	F	44.87
Desmond	M	46.50
Catherine	F	33.18
Kombote	F	39.40
Chipita	F	29.60
Bono	M	38.80
Natalie	F	39.70

Fig. 3.14 Summary of bonobo sample individuals and sequence coverage. Total number: 10. Average sequence coverage: 40.5.

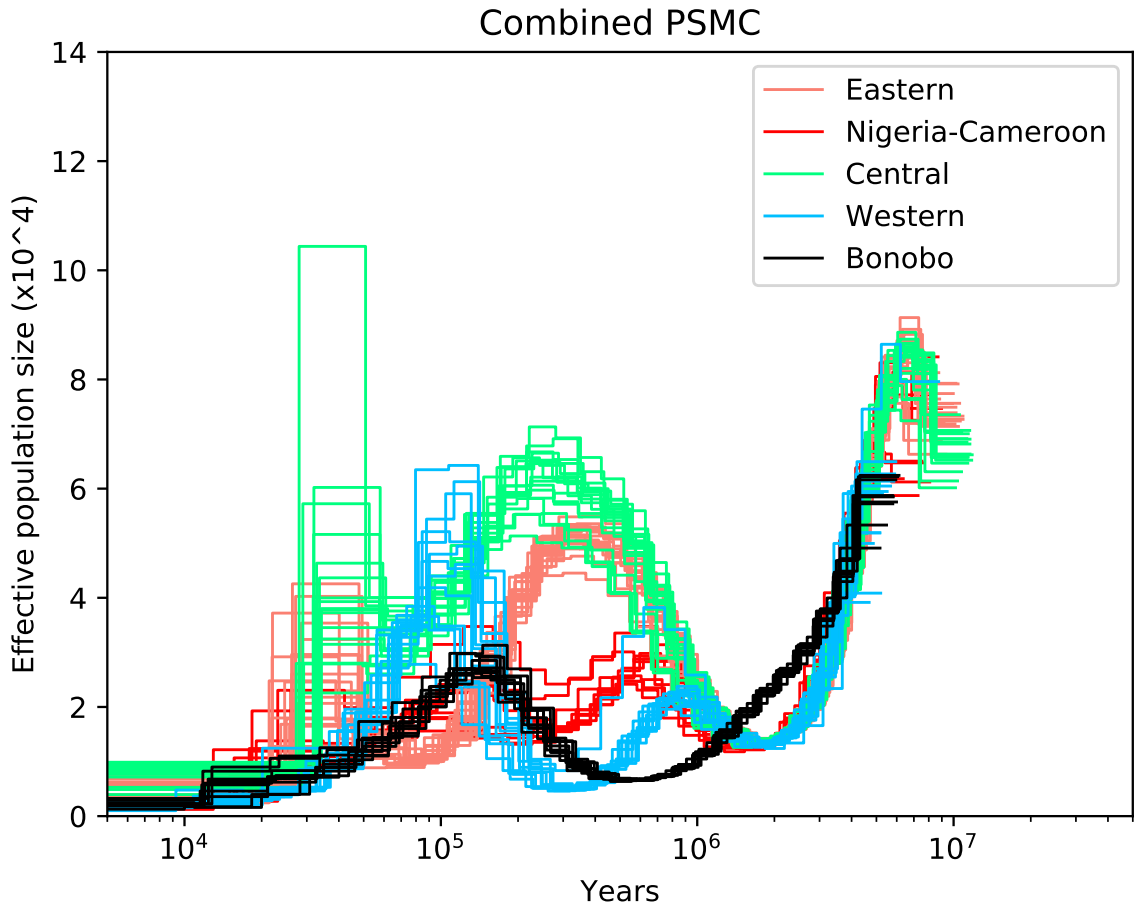


Fig. 3.15 Autosomal PSMC analysis of chimpanzees. Each line corresponds to an individual. Line colours designate either the bonobos or the various chimpanzee subspecies.

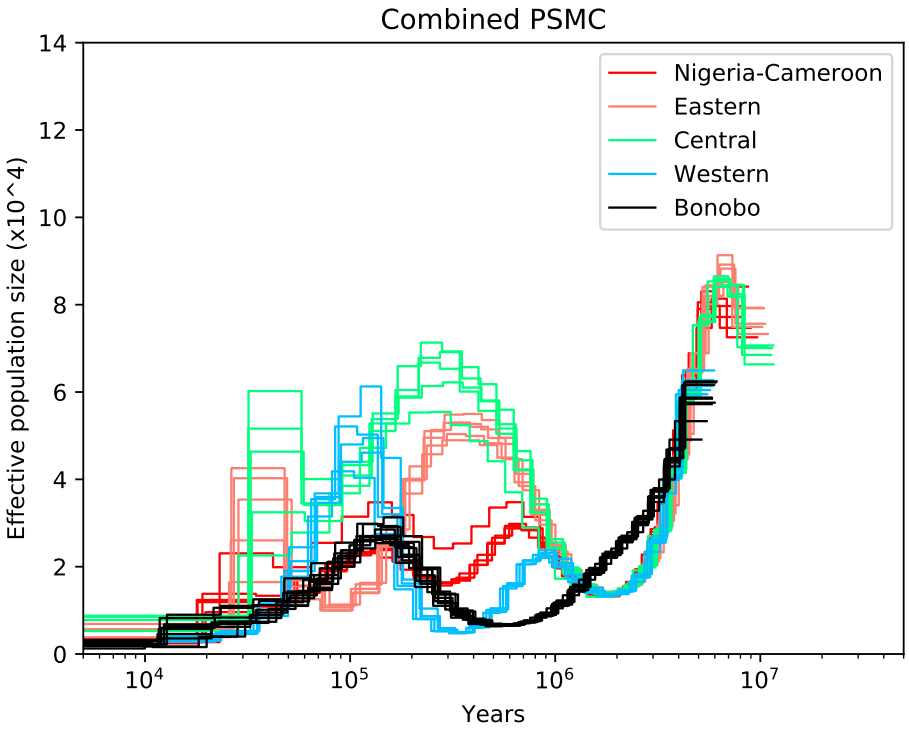
observation that bonobos have a low current heterozygosity relative to other *Pan* populations [24], although heterozygosity cannot straightforwardly be read off these curves (as shown by de Manuel et al., Western chimpanzees and bonobos have similar levels of heterozygosity, but markedly distinct historical N_e curves).

Unlike the bonobos, each chimpanzee N_e trajectory has two peaks after the common ancestral decline. The four subspecies begin to distinguish themselves around 0.8-1 Mya, with the first separation occuring between the major clades recognised in the chimpanzee phylogeny. The Western clade N_e declines over this period while the Eastern clade increases. This suggests either that the process of separation involved smaller founding populations in Western regions around this time, or that there were regional differences in the direction of forest-cover size fluctuations. The latter possibility is supported by the fact that later increases in N_e occur at similar times within the major clades, but at different times between them. If this correlation is a result of continuing gene flow between populations within a clade, we might expect to see this in the cross-population coalescence analysis, which is discussed below.

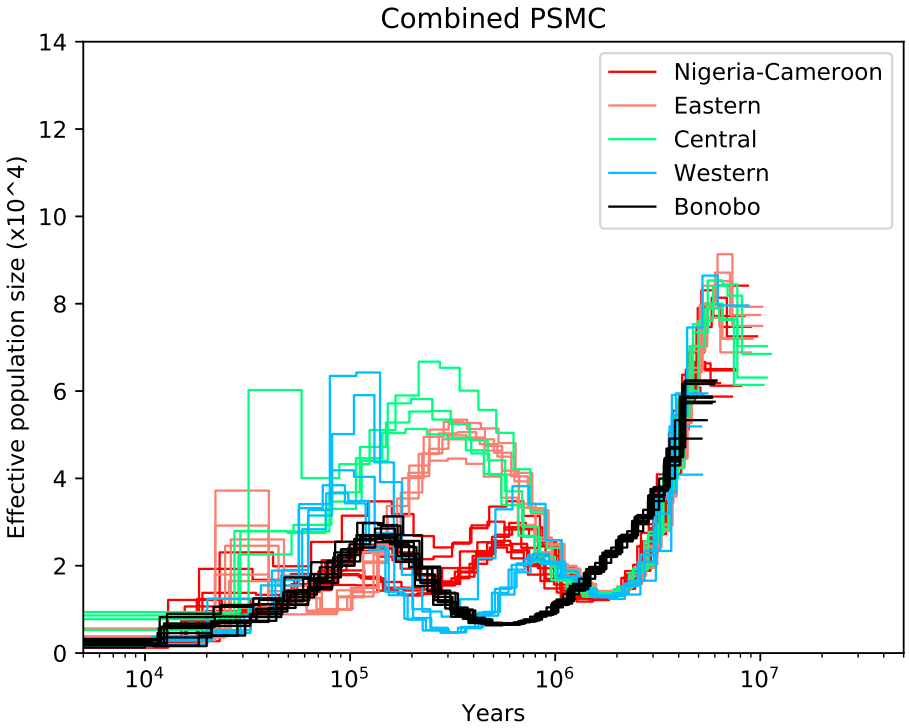
In the Central chimpanzees, the first recovery of N_e after the initial genus-wide decrease gives it momentarily a higher value than seen in any other great ape population N_e trajectory produced by PSMC [120]. Other than for a short period around 100 kya when the Westerns appear to have a larger effective population size, the Central N_e is at least as large as any other chimpanzee population throughout the observed history. One result is that today the Central chimpanzees have the highest genetic diversity in the genus, as measured, for example, by heterozygosity [24].

It is notable that in the first and last time steps of each analysis, the inferred N_e curves do not produce the same scattering effect seen in the orangutans at similar times. This is likely due to the higher average sequence coverage in this analysis, or possibly the use of more stringent variant filtering criteria (see Section 3.3). However, within both sets of Central and the Western chimpanzee inferred N_e curves, there is a notable degree of variance between individuals' maximum N_e values during their more recent peaks. These lie within a time period (30-150 kya) during which PSMC is thought to perform well on great apes [70]. As such, it might be caused by ancient substructure within the populations [87], where individuals draw their ancestry from subpopulations with distinct demographic histories; or it might be an artefact of sequencing error, caused, for example, by low coverage in several individuals. The admixture analysis Figure 3.18, shows little evidence of substructure in the Western chimpanzees, although there is some support for substructure in the Central subspecies. This is further explored in following section. However, the individual with the heighest peak (the Central chimpanzee, Gamin) was sequenced to low-coverage.

Figure 3.16a shows that if we exclude all individuals with lowest sequence coverage then some of the extreme differences between curves within populations are eliminated. However,



(a) High coverage subsample.



(b) Phase 1 subsample.

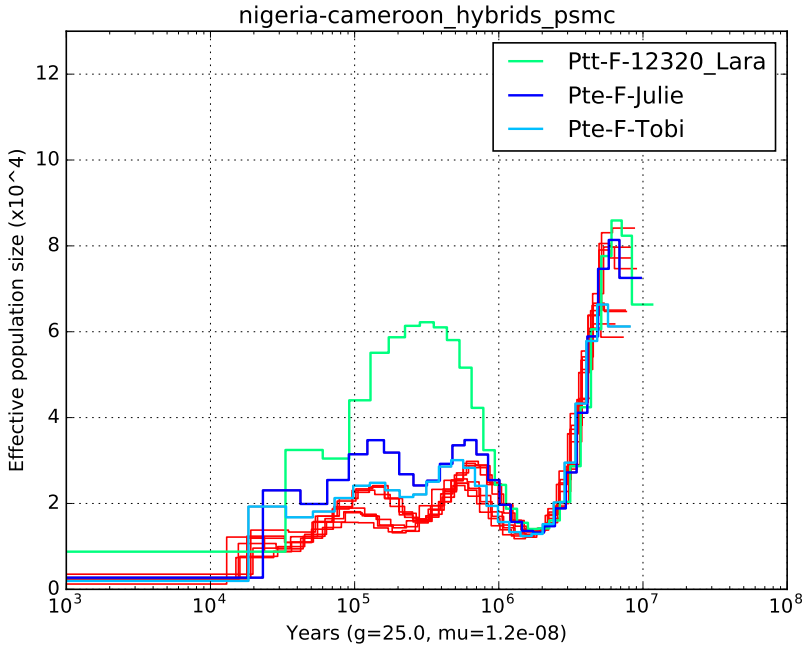
Fig. 3.16 PSMC analysis of high-coverage subsample (see Methods), and subsample included from Prado-Martinez et al. (2013) [120], referred to by de Manuel et al. as "Phase 1" sequences. See Figure 3.13 for sequence IDs.

there is still considerable variability between individuals, and we cannot exclude substructure or sequence quality as explanations for this effect. One reason to think the peaks might be artefactual is that they occur in the second-last time interval. Although this occurs over a time period when we expect PSMC to perform well, it is presumably still drawing an inference from relatively few coalescent events. Figure 3.16b shows PSMC output of individuals sequenced by Prado-Martinez et al., who did not observe the high peaks in the Western and Central curves. This plot is more consistent with their Figure 3b, indicating that the differences were not introduced by any change in the variant-calling pipeline (see Methods).

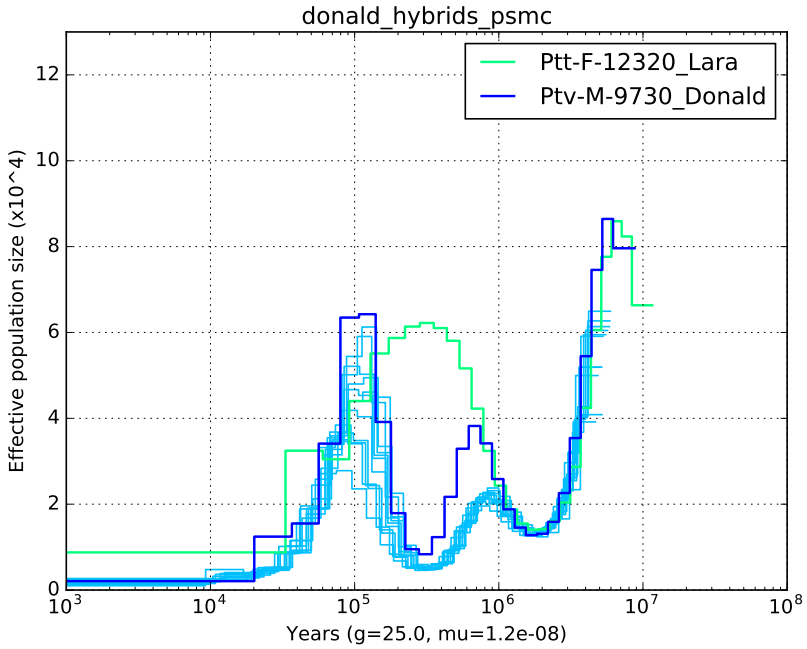
Possible signal of gene flow into Nigeria-Cameroon population From about 800 kya to the present, two Nigeria-Cameroon individuals, Julie and Tobi, appear to have a noticeably higher N_e than other members of the same subspecies (Figure 3.17a). The effect is greater in the N_e curve of Julie. This might be a result of persistent population structure in the population over this time period (Chapter 4), or it might be the result of natural variation in the output of PSMC on this sample. The large variation in the curves of the Central chimpanzees on a similar time period (Figure 3.15) suggests that the positions of the two Nigeria-Cameroon individuals relative to the rest of the population are not unusual at this level of sequence coverage and this sample size.

However, a similar effect is noted with Donald, a Western chimpanzee who has a much higher N_e between 0.3-1 Mya than others in the Western subspecies (Figure 3.17b). Donald was born in captivity and is known to be a second-generation hybrid of Central and Western chimpanzees, having a single Central chimpanzee grandparent. The similar pattern seen in his inferred N_e curve might lead us to predict that Julie and Tobi also have some degree of additional ancestry. Gene flow between the Central and Nigeria-Cameroon populations, for example, is geographically plausible since the ranges of the two subspecies are separated only by the Sanaga River in Cameroon (Figure 3.11). Under some circumstances, an ancestral component not observed in other members of the subspecies will tend to decrease the average coalescent rate relative to members who do not have the component.

While it is not conclusive, this prediction is borne out in the ancestry analysis undertaken by de Manuel et al. Using sNMF, an efficient ADMIXTURE-like method [34], they show that Julie and Tobi share an ancestry component with Central chimpanzees which is comparable to the component that Donald shares with Central chimpanzees (Figure 3.18, taken from de Manuel et al. (2016)). No other non-Central individuals consistently share this component. I would caution against interpreting this simply as meaning that Tobi and Julie are also second-generation hybrids, as the signal might be replicated by an older period of gene flow from the Central into the Nigeria-Cameroon population [32]. Previously, D statistics have shown gene flow between Nigeria-Cameroon and Eastern chimpanzees though have only



(a) High coverage subsample.



(b) Phase 1 subsample.

Fig. 3.17 (a) PSMC analyses highlighting historical N_e of Julie and Tobi, Nigeria-Cameroon chimpanzees. The red lines are the other Nigeria-Cameroon chimpanzees. Lara is a typical (high-coverage) Central chimpanzee. (b) Similar analyses highlighting historical N_e of Donald, a second-generation Central-Western hybrid. Light blue lines are the other Western chimpanzees.

weakly suggested gene flow between Nigeria-Cameroon and Central chimpanzees [36, 120]. Indeed, Prado-Martinez et al. also noticed that Julie and Tobi were unusual and attempted to identify ancestry tracts shared with the Central chimpanzees, but none were found. Since such tracts were identified in Donald, they exclude the possibility of recent gene flow between Central and Nigeria-Cameroon populations. They also did not detect significant gene flow using F_3 statistics and Treemix. Since neither Julie nor Tobi are male, it was not possible to test their cross-coalescence rates with the rest of the Nigeria-Cameroon population. As they were born in the wild, we expect more individuals with this unusual ancestry will be found, and further study should shed light on this history.

Note that while Donald was excluded by de Manuel et al. from further demographic analysis on the grounds that he was a known hybrid, Julie and Tobi were still used. Julie was included as an ordinary Nigeria-Cameroon individual in their ABC-based demographic modelling since she had a high-coverage sequence. This may still be appropriate since neither of these individuals are population-level outliers in several whole genome measures, like PCA clustering, heterozygosity, or D statistic analyses of bonobo introgression [24]. Also, since it is unlikely that the additional ancestry component is a result of a recent Central ancestor, it might be more correct to capture this component of Nigeria-Cameroon variation in demographic modelling.

Cross-population coalescence analysis As with the orangutans, the autosomal PSMC analysis of *Pan* sequences gave us an estimate of the lower bound of the divergence time between the ancestors of bonobos and chimpanzees. In Figure 3.19 we see a more direct estimate of this time using cross-population coalescence rates. Again, we have used the male X chromosomes to conduct this analysis in order to limit errors associated with poor phasing.

First observe that the output for each subspecies is virtually identical. This is consistent with the prediction of Figure 3.15 that the ancestors of each of the four chimpanzee subspecies diverged simultaneously from the ancestral bonobos, most likely forming a common ancestral chimpanzee population soon after separation. This does not mean that the speciation process was simple, only that if it involved multiple intermediate populations or multiple waves of migration, then the ancestors of the extant chimpanzee populations were distributed identically among the involved intermediate or migratory groups.

Until about 3 Mya, and ignoring the very earliest time intervals, the cross-population curves closely track the curves of the autosomal PSMC. Thereafter they begin to increase rapidly, becoming great enough by about 2 Mya that they could not correspond to the N_e of a single *Pan* population. This represents the divergence period between the ancestral populations. The reason they follow the PSMC curves so closely until divergence begins is that the pseudo-diploid sequences are made up of sequences drawn from the same ancestral population over this time period and thus represent almost the same signal as the PSMC

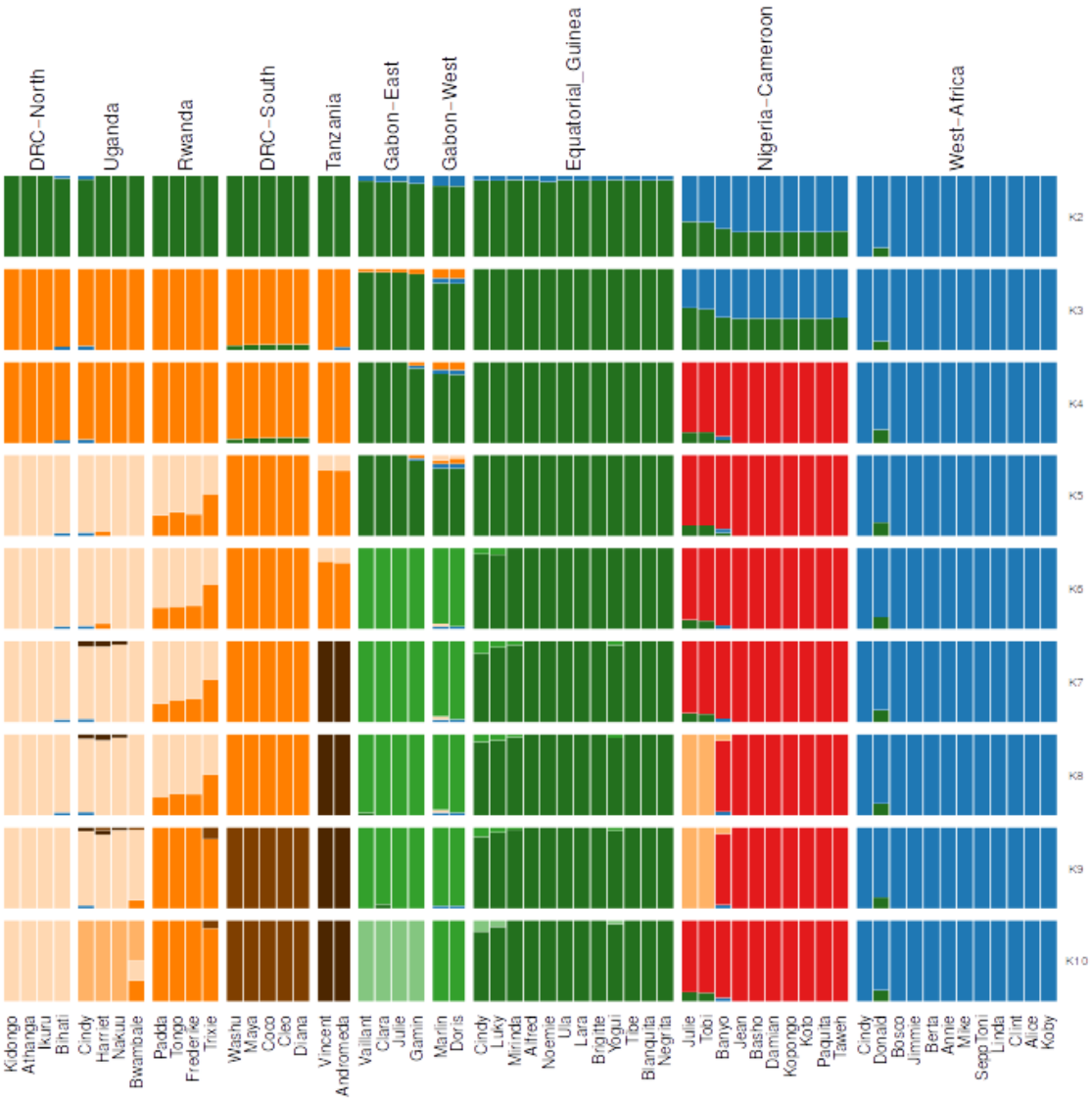


Fig. 3.18 Ancestry component analysis of chimpanzee individuals based on the ADMIXTURE-like method sNMF [34]. From de Manuel et al. (2016) [24]. Note the small additional ancestry components that Julie and Tobi ("Nigeria-Cameroon"), and Donald ("West-Africa") share with Central chimpanzees.

analysis conducted on the autosome. Differences between the curves before divergence might correspond to differences in the effective population size of the X chromosomes. We expect that the average N_e of the X chromosome will be smaller than the average autosomal N_e since males only have a single copy of the locus. Sex-biased demographic phenomena, such as excess male philopatry or different male:female breeding ratios, can also cause differences in effective population sizes on the sex chromosomes. However, the locus is considerably smaller than the autosome and is thus also more likely to be affected by sampling bias and methodological uncertainty. Subtle demography-linked effects may be obscured. In Figure 3.19 the X chromosome cross-population N_e does appear to be smaller than the autosomal values (before divergence). Using the same figure we can support the hypothesised separation time inferred from the autosomal PSMC analysis and say that divergence likely occured between 2-3 Mya.

In their publication, de Manuel et al. produced evidence of old gene flow from chimpanzees into bonobos. This study, some of which formed part of that work, could not detect this signal of gene flow. However, the analyses produced here do not rule out small amounts of gene flow between the ancestral populations of the species. This is discussed below.

Within the chimpanzees, the cross-population coalescence rates are less easy to interpret, most likely because the process of divergence was more gradual, involving the separation of different populations at different times. In Figure 3.20 I have shown the MSMC2 analyses of coalescent rates between each of the six pairs of chimpanzee subspecies. These are superimposed on the autosomal PSMC curves of the same individuals in light grey.

Looking first at the cross-population curves, we notice that most of the curves track each other closely. The two exceptions are notable: both the light blue and the red curve correspond to the only cross-population analyses run within, as opposed to between, the major clades (light blue corresponds to the Eastern clade, red to the Western). From the oldest time periods to the more recent ones, observe that all the curves follow the PSMC curves very closely until about 2.5 Mya. After this period, and excluding the red "Pte-Ptv" (Nigerian-Cameroon and Western) comparison, the cross-population inferred N_e is greater than the autosomal PSMC curves. They monotonically increase at a rate not much greater than the increase in the ancestral Central and Eastern cluster of curves. Then, at around 500-600 kya, all of the curves corresponding to coalescence rates between the major clades increase drastically, most likely marking the final divergence between Eastern and Western clades. Within the clades, the Nigeria-Cameroon and Western chimpanzees separate a short while after this, by about 300-400 kya, and the Eastern and Central chimpanzees separate much later, between 100-200 kya. This supports the general conclusion drawn from the PSMC analysis that inferred population histories correspond to the most likely histories drawn from direct phylogenetic methods.

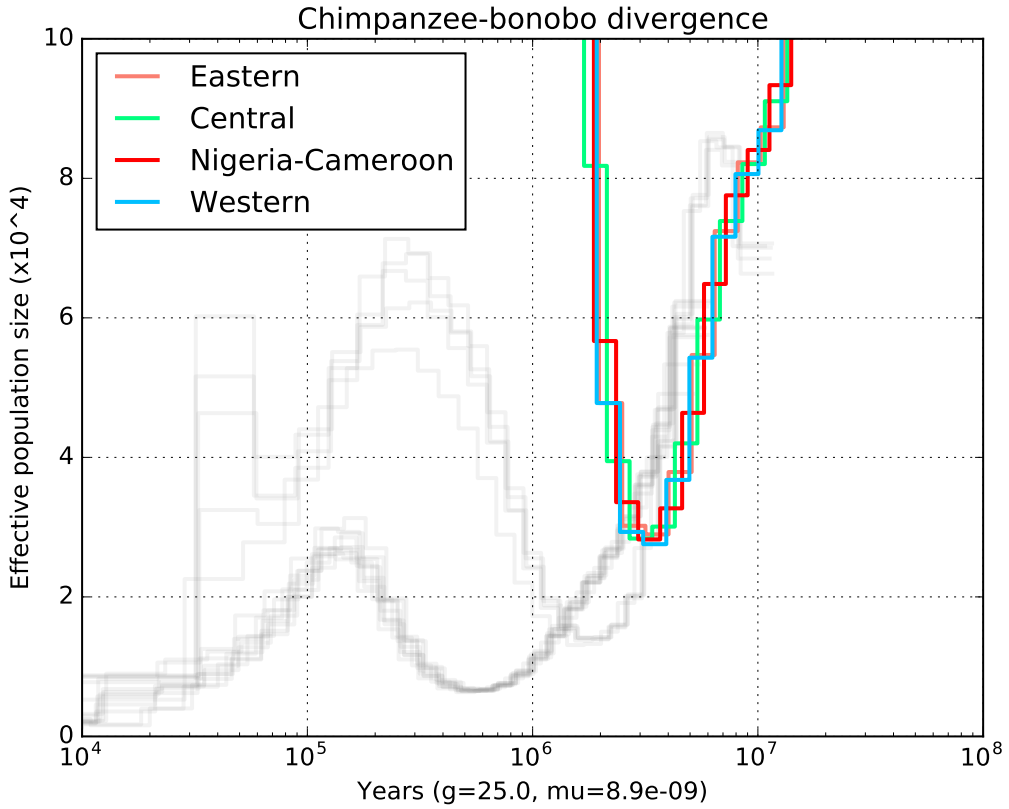


Fig. 3.19 Genus level historical cross-population coalescence rates inferred by MSMC2. Each coloured line represents historical rates of coalesence between a chimpanzee subspecies and the bonobos. Analysis conducted on male X chromosomes. For comparison sake, included in grey are the autosomal PSMC analyses of the high-coverage Central chimpanzees (upper curves) and the bonobos (lower curves). Refer to Figure 3.15 for further clarity.

Note that the population PSMC curves seem to be distinguishable much earlier than the confident lower bound on the divergence time between the major clades (about 700-800 kya, compared to the 500-600 kya mentioned above). Also, the division is not as abrupt and clear as the separation between chimpanzees and bonobos (Figure 3.19). However, the coalescent rates are clearly declining (inferred N_e is increasing) from about 2 Mya. It is striking that the inferred N_e appears greater than that of the ancestral chimpanzee population from 2 Mya, when the ancestral populations are not obviously distinct yet. One interpretation of this observation is that the cross-population coalescent rate could be lower than the within-population rate, an indication that population substructure is beginning to emerge during this period. This reading is supported by the fact that until this point, the cross-population curves very closely track the PSMC curves, and when the increases in inferred N_e occur in both sets of analyses, the cross-population curves have a slightly more concave shape than the PSMC curves. This would seem to be a much earlier sign of structure in the chimpanzee populations than previously thought, starting very soon after the divergence from ancestral bonobos.

An alternative explanation is that the X chromosome mutation rate is too low here, and in the previous figure, so that both curves should be shifted slightly to the left relative to the autosomal PSMC curves. Together with the additional uncertainty inherent in using a smaller locus, this might cause the pattern to disappear. Although this may improve the alignment with the PSMC curves in the previous figure, here, however, it would tend to make the alignment worse on time periods earlier than 3 Mya. After 800-900 kya it is also not clear that the cross-population curves would be expected to closely follow any of the PSMC curves, since the populations begin to appear distinct in the PSMC analysis from this time, and thus another signal is being depicted.

The within-clade curves both have a distinctly different shape to the between-clade curves, even granting the different absolute divergence times. They both decline first, before their drastic increases. The red "Pte-Ptv" curve (corresponding to the Western clade) seems to decline along with the decline in the inferred N_e of the Nigeria-Cameroon and Western chimpanzees. This corresponds to an increase in coalescence rates between the populations, although it might be due to a constant rate of gene flow in the presence of declining populations. A similar explanation can be offered for the decline in the blue Pts-Ptt curve (the Eastern clade).

The detectability of cross-species gene flow A central claim by de Manuel et al. was that there was gene flow between chimpanzees and bonobos after the initial divergence occurred. It would seem from Figure 3.19 that no gene flow existed between the species after about 2 Mya, when divergence process was completed. However, it is still possible that a low level of migration would not be detected using this approach. It is not obvious what threshold

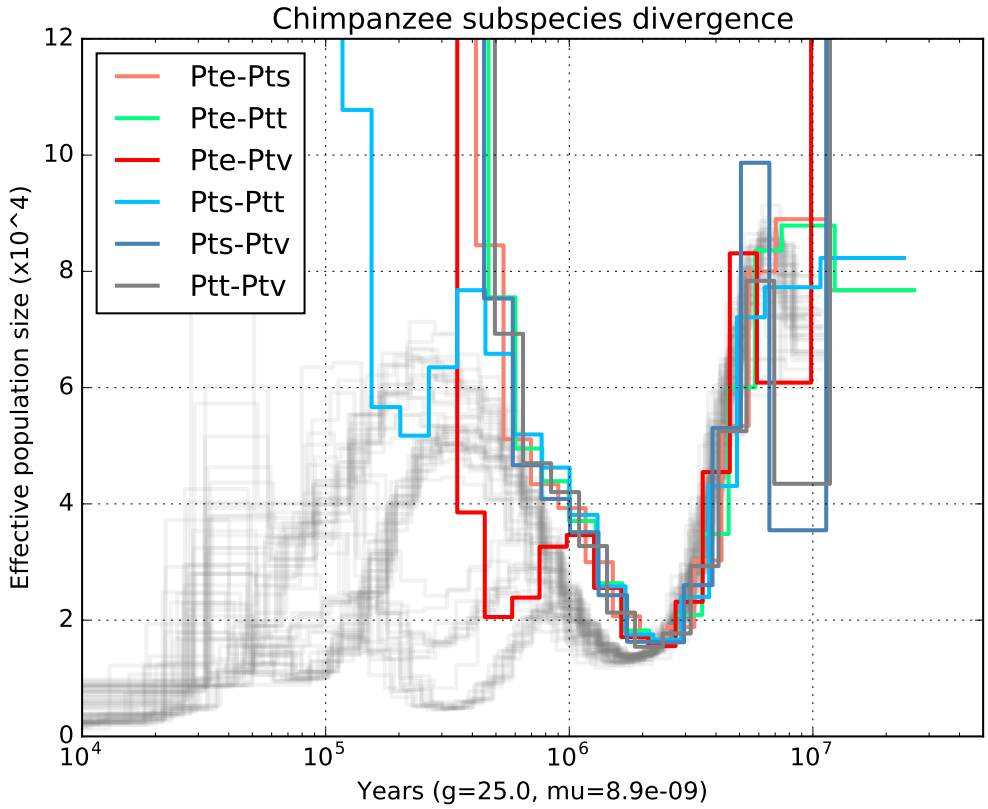


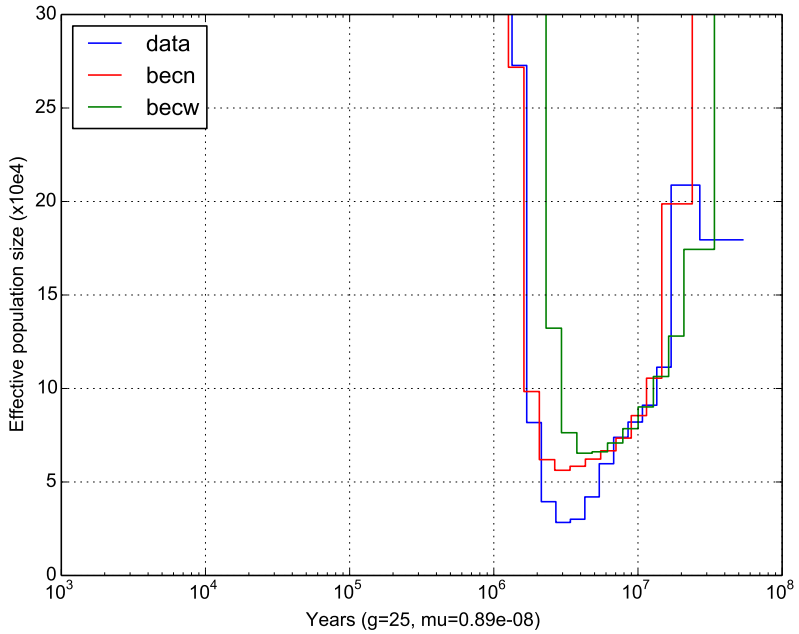
Fig. 3.20 Species level historical cross-population coalescence rates inferred by MSMC2. Each coloured line represents historical rates of coalescence between pairs of chimpanzee subspecies: Central ("Ptt"), Eastern ("Pts"), Western ("Ptv") and Nigeria-Cameroon ("Pte"). Analysis conducted on male X chromosomes. For comparison sake, included in grey are the autosomal PSMC analyses of all the chimpanzees. Figure 3.15 indicates the subspecies represented by each cluster of grey lines.

Source ancestral population	Target ancestral population	Time (kya)	Amount ($2Nm$)
Bonobos	Centrals and Easterns	200 - 500	0.04 - 0.1
Bonobos	Centrals	<200	7×10^{-5} - 0.16
All chimpanzees	Bonobos	200 - 500	1.3×10^{-4} - 0.22

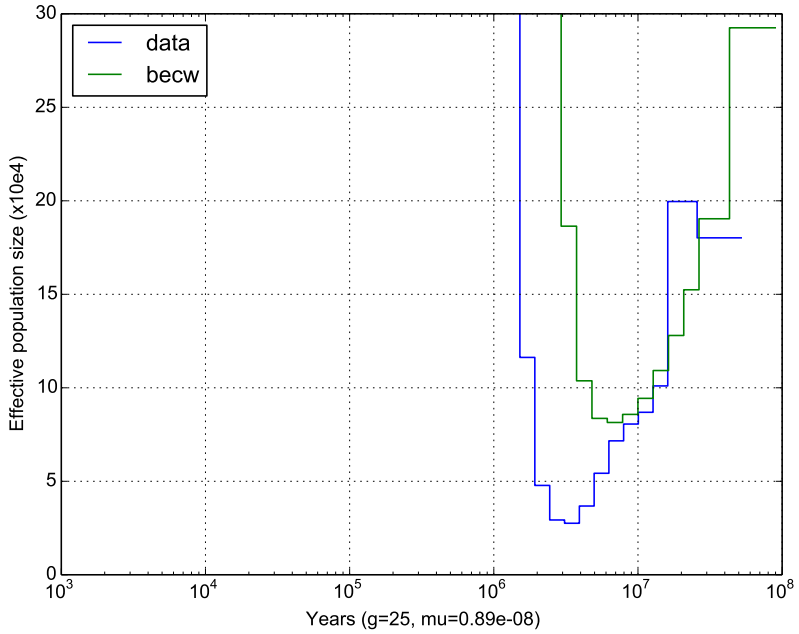
Fig. 3.21 Migration rates inferred by de Manuel et al. [24].

of migration would affect the coalescent rates. As part of the project, Vitor Sousa and Laurent Excoffier conducted an ABC-based analysis using fastsimcoal2 [31], and fit various models of population evolution in the genus. In Figure 3.22, I show cross-population N_e curves inferred from sequences simulated under the two best-fit models produced, by Sousa and Excoffier. The first model, "becn", is best fit for the bonobos, Central, Eastern and Nigeria-Cameroon populations. The second, "becw", is the best fit model in which Nigeria-Cameroon is swapped with the Western population. Figure 3.22a shows the cross-population MSMC2 analyses of the Central chimpanzees and bonobos, using the observed data and sequences simulated under both models (Methods). Figure 3.22b shows the Western chimpanzee and bonobo comparison under the model which features the Western population.

The amount of interspecies gene flow estimated under these models is summarised in Table 3.21. From the figures produced here, we can infer that MSMC2 will not detect the amount of gene flow proposed in either model. It also shows that the "becw" model compares less well to the observed cross-population coalescence rates than the "becn".



(a) Central chimpanzee and bonobo cross-population coalescent rates with simuation



(b) Western chimpanzee and bonobo cross-population coalescent rates

Fig. 3.22 Simulation evidence showing the inability of MSMC2 to detect low levels of gene flow after population divergence. Sequences simulated using two models proposed by de Manuel et al: "becn" consisting of bonobo, Eastern, Central, and Nigeria-Cameroon populations; and "becw" consisting of bonobo, Eastern, Central and Western populations. Blue lines show analyses run on observed sequences, identical to those in Figure 3.19. (a) Inferred cross-coalescence rates between Central and bonobobo populations. (b) Inferred rates between Western and bonobo populations.

3.3 Methods

3.3.1 Data

Orangutans Nater et al. collected blood samples from orangutans found in rehabilation centres across Malaysia and Indonesia. DNA from these samples were sequenced and mapped to the orangutan reference genome ponAbe2 [77] using a standard Burrows-Wheeler Aligner (BWA-MEM) pipeline [72]. Further read filtering and local realignment was undertaken using the Genome Analysis Toolkit (GATK) [89]. Across the entire sample, average effective sequencing coverage was 18.4x, though the range of coverage is considerably lower in data taken from Locke et al (Figure 3.1). A standard GATK pipeline was applied to call and filter variants [25]. Ultimately two VCFs were produced, one for orangutans on Borneo and one for those on Sumatra. Note that the process was also applied to the read data from Locke et al. (2011) and Prado-Martinez et al. (2013) to ensure uniformity in variant calling pipelines. Across the sample, this pipeline identified 30640634 SNPs, currently the largest published catalogue of orangutan single nucleotide variants.

Nater et al. produced a mappability mask, which identified those positions which were in uniquely mappable 100-mer regions of the orangutan reference genome (with fewer than 5 mismatches) in order to exclude regions with a high probability of being ambiguously mapped and thus a high risk of incorrect genotype calls. The mappable regions were identified using the mappability module from the GEM library [26]. This mask is used in both the autosomal PSMC and cross-population MSMC2 analyses.

Chimpanzees and bonobos Blood samples of 40 new chimpanzees were collected and sequenced by de Manuel et al. and pooled with sequences from Prado-Martinez et al (25 chimpanzees and 10 bonobos). Reads (including those from bonobos) were mapped to the chimpanzee reference sequence CHIMP2.1.4 (http://www.ensembl.org/Pan_troglodytes) using a standard BWA-MEM pipeline. SNP variants were called using FreeBayes [35] with standard filtering parameters. In total 22 081 627 high-confidence SNPs were identified, 32% more than in any previous study. The procedure was applied to read data across the sample to ensure uniformity. As described in the orangutan methods, a mappability mask was also inferred for the *Pan* individuals using the mappability module in GEM.

3.3.2 Autosomal PSMC analysis

For the PSMC analyses, variant data in VCF format was extracted and handled using BCFtools [71], the mappability mask was adapted using BEDtools [124], and both were used to produce the PSMC input data in required format using Python scripts adapted from those by Aylwyn Scally (https://github.com/aylwyn/aostools).

PSMC was run using parameter settings which were previously found to be appropriate for great apes [e.g. 70, 120]: I used discrete temporal binning parameters of -p "4+25*2+4+6", an approximation which limits the search space of the method by reducing the time resolution of inferred population size histories over relatively recent and relatively old periods (it signifies in this case that the first inferred populution size spans 4 atomic time steps, the next 25 span 2 such steps, and so on). The ratio of θ to ρ was 5, and the mutation and recombination rates were assumed to be constant. Run on simulations of sequences with varying mutation rates and recombination hotspots, it was shown that PSMC is capable of recovering population histories without losing significant accuracy relative to simulations in which those rates are constant [70]. We assumed bin sizes of 100 bp, also according to recommendations of robustness in previous great ape studies [e.g. 183]. The effects of potential errors in the scaling parameters, mutation rate and generation time, are discussed below. For both the PSMC and cross-population analysis described below, custom Python scripts were written to plot and scale the output, importing Perl and Python modules from Heng Li's PSMC "utility" scripts (https://github.com/lh3/psmc/tree/master/utils) and Stefan Schiffel's helper scripts in msmc-tools (https://github.com/stschiff/msmc-tools).

If each diploid sample in a single population was generated through approximately the same demographic process, then each set of inferred pairwise coalescent times is in principle a different set of random draws from the same underlying coalescent distribution (realisations of the random variable T_2). Thus variation between individual curves is partly a function of variance in the underlying distribution. A standard approach to assessing this variance with small samples is through bootstrapping, which was implemented in the orignal suite of PSMC modules. Using many samples is a more direct way of assessing the range of plausible values PSMC might infer for individuals evolving under similar demographic conditions.

Scaling parameters PSMC infers the historical distribution of effective population sizes in coalescent units. These are scaled into years using an average autosomal mutation rate and generation time. We use the approximation that mutation rates and recombination rates are constant across the autosome. Generation time used is the average zygote-to-zygote time (in years). Both mutation rates and generation times are assumed to be constant across the time scales involved in these analyses. This is not believed to be true over long time scales, but there is considerable uncertainty involved in calibrating changes in these values across time and in undestanding the biological processes determining their history in the great apes [e.g. 142, 98]. The changes in these rates over long time scales are not implemented in the underlying inference scheme used by PSMC, nor in the rescaling of output. Parameters were also kept constant across each genus as there is no evidence to suggest large differences in values between the species in each of them. A discussion on the calculation of sex chromosome mutation rates is presented in Section 3.3.3.

As an indication of the effect of changes in these parameters, Figure 3.23 shows the consequence of using values which are approximately 20% greater and 20% smaller than those currently chosen in the *Pan* genus. If the true generation time is larger, for example, it will have the effect of pushing the divergence event between chimpanzees and bonobos further back in time. An increase in mutation rate would have the opposite effect, bringing divergence closer in time to the present, while also decreasing the estimated "effective population size". In either plot, a 20% change in the relevant parameter changes the estimate of the divergence time by as much as 1 million years. Errors in these inferred times would of course be compounded if both parameters are substantially wrong in opposite directions (if either is too small, while the other is too big). Nonetheless, the broad features of the plots will remain the same. Inferrences of the existence of bottlenecks, for instance, will not change, even if their absolute size and timing might. Importantly, scaling changes will not alter the sequences of events inferred from the plots, since they affect all curves in the same way. Uncertainty in these parameters does, however, make it difficult to correlate changes with environmental events.

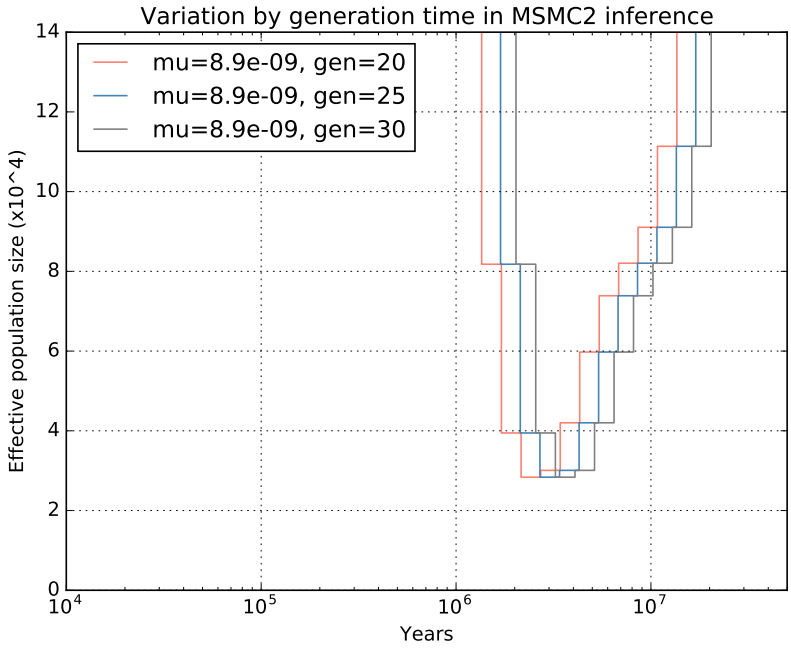
3.3.3 Cross-population MSMC2 analysis

Samples used Following Nater et al., the orangutans were grouped into seven populations based on genetic similarity and sample sizes. The three populations on Sumatra were the "Northeast Alas" which consisted of the North Aceh and Langkat orangutans, the West Alas, and the Tapanuli or Batang Toru orangutans. No cross-population comparisons with the West Alas population were performed since no male individual from that region is present in this sample. On Borneo, the North and South Kinabatangan orangutans have been combined into a single population ("Kinabatangan") as well as the Central and Western Kalimantan, ("Central/West Kalimantan"). The other two population on Borneo are the East Kalimantan and the Sarawak. I excluded the following low-coverage individuals from the analysis: PA_KB5883 and PA_KB4661 (North Aceh); PP_KB4204 and PP_KB5405 (Central/West Kalimantan); and PP_KB5543 (East Kalimantan).

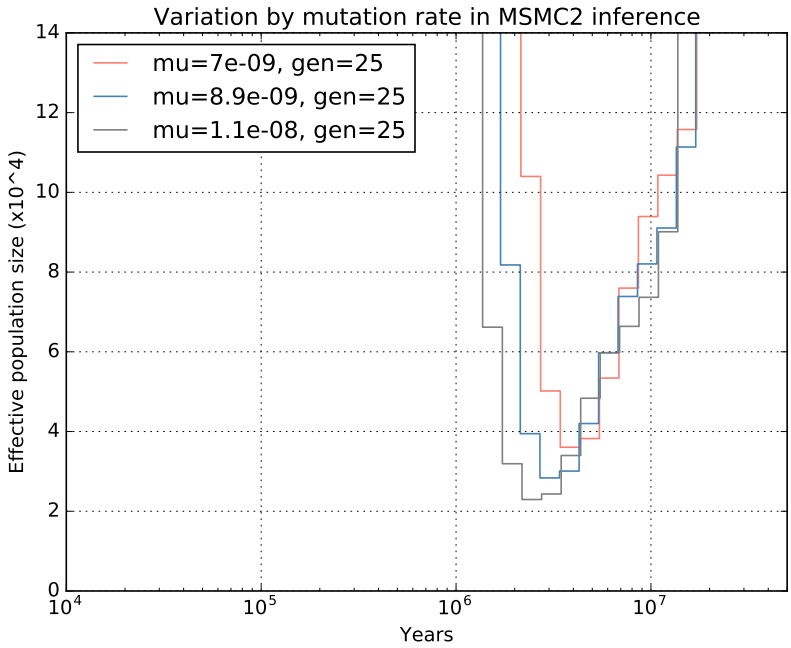
All the male chimpanzees and bonobos were used, and their population designations correspond to their subspecies and species classifications.

Processing I prepared input data files using BCFtools [71] and custom versions of the Python conversion tool generate_multihetsep.py (https://github.com/stschiff/msmc-tools). These input files incorporated the mappability mask described above, converted to the required format using BEDtools [124].

Default time discretization parameters were used throughout, but we note that reasonable modifications to them did not substantially affect our results. Cross-population comparisons were handled with -P flag. For example, since we had one male individual from Sarawak and



(a) Central chimpanzee and bonobo cross-population coalescent rates



(b) Central chimpanzee and bonobo cross-population coalescent rates

Fig. 3.23 Effect of 20% change in mutation rate and generation time on inferred cross-population coalescent rates. These plots were used to study the divergence between bonobos and chimpanzees using parameters corresponding to the blue curves.

two from Kinabatangan, we ran MSMC2 using -P 0,1,1 when analyzing gene flow between these two populations.

X chromosome mutation rates There are two very similar ways of deriving the X chromosome mutation rates, depending on the availability of estimate of other mutation parameters. We applied the different approaches here.

With the orangutans, MSMC2 results were scaled using an X chromosome mutation rate $\mu_X = 1.17 \times 10^{-8}$ mutations per base pair per generation. This was determined using the relationship $\mu_X = (4\mu_A - \mu_Y)/3$, where μ_Y is the Y-chromosomal mutation rate and μ_A is the autosomal mutation rate. The relationship is derived by considering the proportion of time X chromosomes spend in the male line and assuming the sex chromosomes have the same average mutation rate as the autosomal chromosomes [70, 17]. We assumed an autosomal mutation rate of $\mu_A = 1.5 \times 10^{-8}$ per base pair per generation, a Y-chromosomal mutation rate of 2.5×10^{-8} per base pair per generation [184], and used a generation time of 25 years [178].

The reasoning was expressed in a slightly different way in the chimpanzees. Since 2/3 of an X chromosome's history is spent in females, male-mutation bias causes X chromosomes to have a lower mutation rate than autosomal loci which, on average, divide their histories evenly between males and females. For a given ratio α of male-to-female mutation rates, we determine the X chromosome mutation rate μ_X using the expression $3(1+\alpha)\mu_X = 2(2+\alpha)\mu_A$, where μ_A is the autosomal mutation rate [12, 95]. We could determine α using an estimate of the Y chromosome mutation rate μ_Y, so we are able to derive an equivalent expression for μ_X in terms of μ_A and μ_Y. However, here we used the expression stated above with the values $\alpha = 7.8$ and $\mu_A = 1.2 \times 10^{-8}$ per base pair per generation as derived in Venn et al. (2014) [166]. These determine an X chromosome mutation rate $\mu_X = 0.89 \times 10^{-8}$ per base pair per generation. We used a generation time of 25 years [64].

Note that the reason we used different methods to estimate μ_X here is to directly work with the estimates of relevant additional quantities which are taken from previous research. In the case of the orangutans, this was the Y-chromosomal mutation rate, and in the case of the chimpanzee, the ratio of male-to-female mutation rates. Using the equations in the previous two paragraphs we can determine the following corresponding quantities: for orangutans $\alpha = 4.8$ and for chimpanzees $\mu_Y = 2.13 \times 10^{-8}$. Both these quantities are comparable to those in the other species (for chimpanzees $\alpha = 7.8$ and for orangutans $\mu_Y = 2.5 \times 10^{-8}$).

In Figure 3.24 I replot the cross-coalescence between North and South Kinabatangan with a doubled X chromosome mutation rate (while keeping the generation time constant). This curve very closely tracks those inferred from the automosomes of Bornean orangutans. Note that this similarity is the case both with the timing of the population bottlenecks, and with the estimated effective population sizes. It is also more consistent with a history

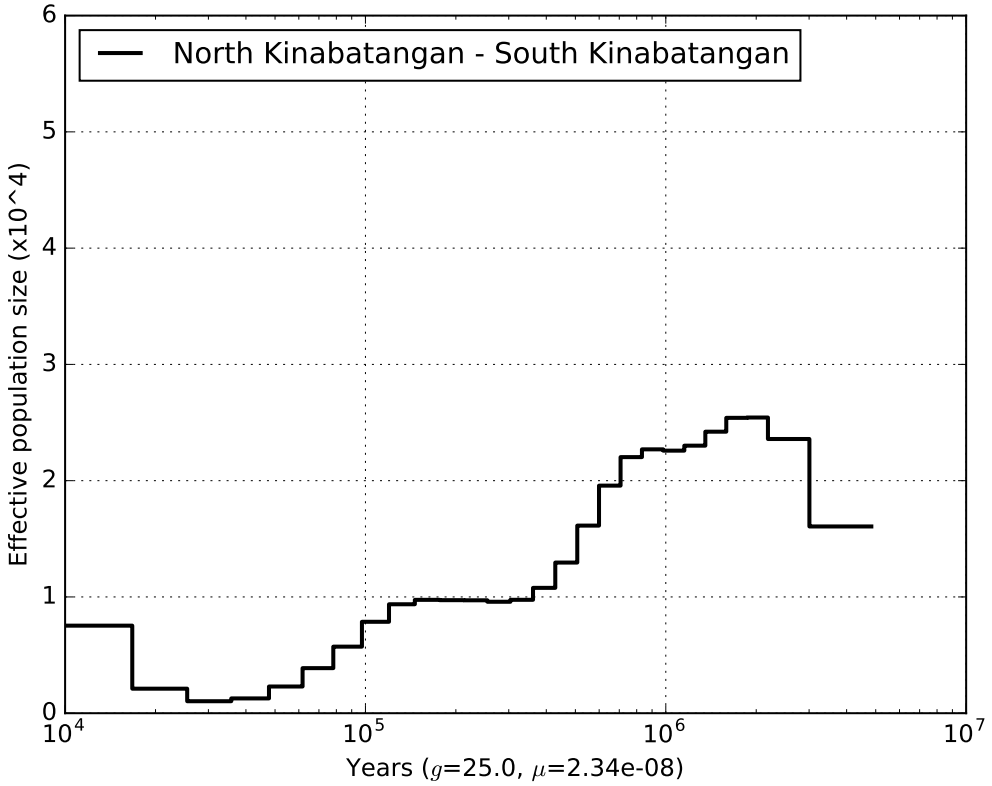


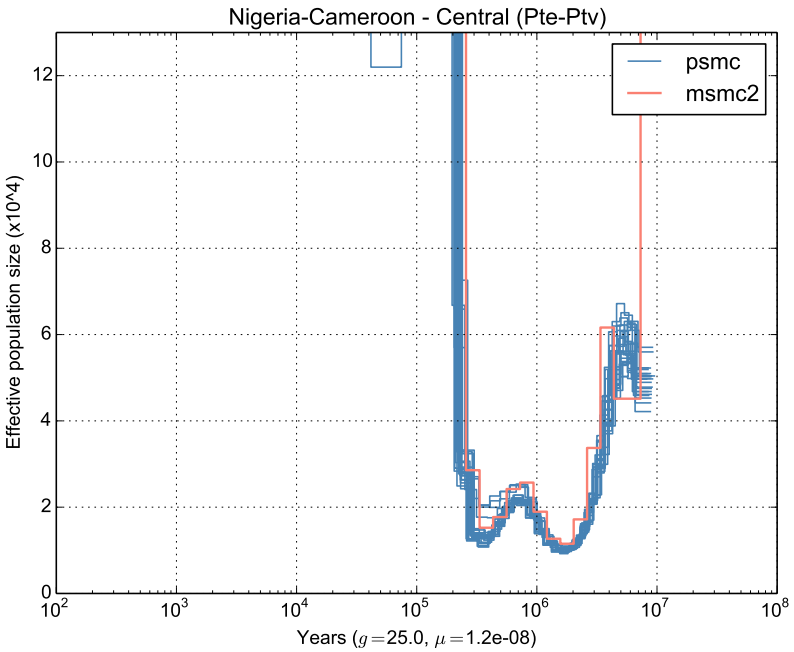
Fig. 3.24 Effect of doubling X chromosome mutation rate on Kinabatangan cross-coalescence analysis.

in which the Kinabatangan populations exchange gene flow continuously until the present. This inconsistency can be taken as evidence that at least one of the orangutan autosomal or sex-chromosome mutation rates are incorrect, though independently determining the correct mutation rates would be beyond the scope of this study.

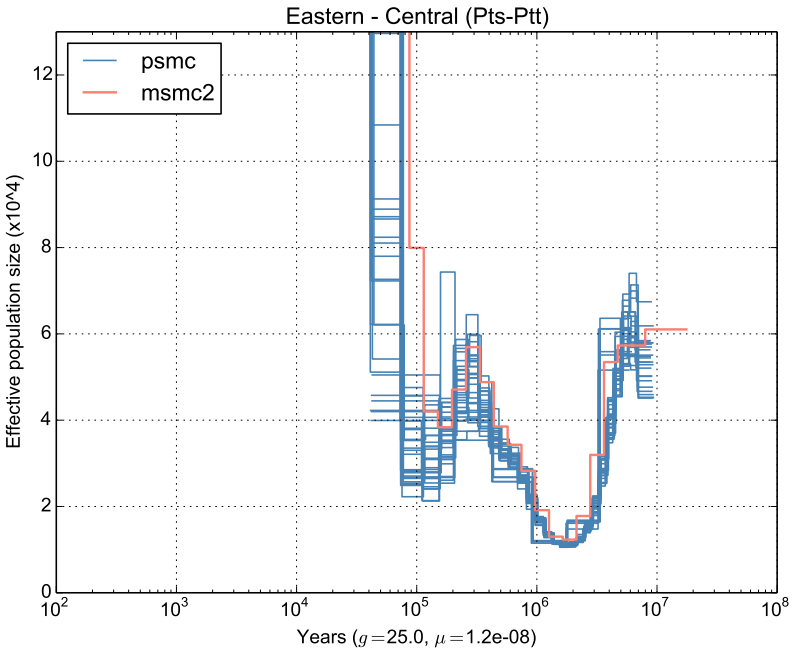
Comparison with PSMC For the sake of comparison, I show in Figure 3.25 the results of running PSMC and MSMC2 in order to infer the cross-population coalescent rates. In the first plot (3.25a), there is a close agreement between the analyses. Each PSMC run is conducted on a pseudo-diploid sequence constructed using a male X chromosome from a Nigeria-Cameroon individual, and one from a Central individual. In this case, the PSMC output for each pairing is very similar to the other PSMC curves. Notice that MSMC2, which combines the information used to construct the entire set of PSMC curves, does not infer a simple average curve. Over most time intervals its inferred N_e values are slightly larger than the PSMC-inferred ones, although it clearly conveys the same signal. In the second plot (3.25b), there is much more variance between the PSMC lines, which also seem to have more extreme average peaks and dips. This is surprising given that there are fewer

Nigeria-Cameroon males in the sample than each of the Central and Eastern subspecies (4 as opposed to 6), and the average sequence coverage of the Nigeria-Cameroon individuals (male and females) is lower than that of the other subspecies. The increased variation between individuals in the Eastern and Central populations was also present in the autosomal PSMC curves in these populations. MSMC2 is less sensitive to this variation, and perhaps it obscures genuine differences between individuals in these populations (which also have a higher genetic diversity).

Simulations In order to test the ability of MSMC2 to infer the history of gene flow detected by De Manuel et al. we also used the method on the output of simulated histories using the coalescent simulator SCRM [157]. For both histories, here designated "becn" and "becw" for the populations simulations, we used the command beginning `scrm 16 1 -t 9600 -r 7680 20000000 -I 4 4 4 4 4` (the rest of the command specifies the detailed demographic model fitted by de Manuel et al., mentioned in Results). This simulates 4 haplotypes of length of 20Mb for each population group, with a mutation rate and recombination rate scaled to be consistent with usage elsewhere. These haplotypes were put through the identical MSMC2 pipeline as the observed male X chromosomes, with the results described in Figure 3.22.



(a) Nigeria-Cameroon and Western Chimpanzee divergence



(b) Eastern and Central Chimpanzee divergence

Fig. 3.25 A comparison of the two methods used to infer cross-population rates of coalescence. Each red line is produced by MSMC2 and draws on the same data used to produce the set of blue curves in the respective plot. This data consists of the male X chromosomes taken from each population. A single blue line represents an the output of a PSMC analysis of a single pseudodiploid sequence. Note, that these plots were constructed early in the project as a comparison of methods and should not be used to draw demographic inferences as the final scaling parameters had not been settled at this stage.

Chapter 4

The Genomic Effects of Population Substructure

In this chapter we study the effects of ancient population structure on genomic variation and look at possible ways to detect such structure. We are specifically interested in the effects of transient historical structure on expected distributions of pairwise coalescent times. Picking up from the previous chapter, this can be understood as studying the extent to which population structure confounds the interpretation of inferred historical N_e . We study the effects of this kind of deviation from panmixia by generalising a demographic model briefly analysed by Li and Durbin (2011) [70], and by examining parameterisations of the model relevant for the inferrence of ancestral hominid structure. This is done through simulation and theoretical analysis. We present some evidence that major changes in human N_e , inferred using methods such as PSMC, are more consistent with models featuring genuine changes in census population size than with models featuring subdivision with varying levels of migration. We conclude that detecting transient historic substructure will always be difficult without ancient DNA, and discuss ways in which this additional source of information might be used. This leads us to the method developed in greater detail in the following chapter.

Population structure and rates of coalescence With the publication of PSMC, Li and Durbin presented an argument about the effect of a form of population structure on the distribution of pairwise coalescent times, concluding that structure will tend to inflate N_e above the value of the sum of deme sizes [70]. The argument is as follows. Consider a population which, looking backwards, split at time s into two demes which remerge at time t ($s < t$), and assume no migration between the demes while they are distinct. During this period of separation, the demes have the same effective population size as each other and the sum of these values is equal to that of the panmictic population before and after this period. Elementary coalescent arguments show that the probability that two distinct

(uncoalesced) lineages at s will coalesce before t, is given by the chance that they sort into the same subpopulation multiplied by the chance of coalescence there: $(1/2)\left(1 - e^{-2(t-s)}\right)$ [171]. On the other hand, in the scenario where no structure exists, the probability of those same loci coalescing between t and s is $\left(1 - e^{-(t-s)}\right)$. The probability of coalescing before t is greater in the panmictic population. (This can be seen by first noting that the terms are identical when $t = s$). Then, if viewed as a function of $\tau = t - s$, the first order derivative of the 2-island expression is $e^{-2\tau}$. When $\tau \geq 0$, this is less than or equal to $e^{-\tau}$, the derivative of the panmictic function. (Figure 4.19 presents a graphical illustration of this argument and develops it in a more general context.) Thus the rate of coalescence over many such loci will be lower in the structured population and we expect that a greater N_e will be inferred in that scenario. It is easy to see how this argument extends to a similar structure consisting of more than 2 subpopulations, and with each additional subpopulation it is straightforward to show, using a similar argument, that the coalescent probability decreases relative to the panmictic value.

This argument leaves open several questions related to the size of the effect. Namely, whether the inflation in coalescent rates is significant enough to confound the common interpretation of PSMC curves, under which the curves mostly reflect changes in census population size; if so, how long the structure would need to exist in order to produce specific N_e "humps" seen in, for example, curves drawn from human sequence data [e.g. 81]; the extent to which the effect changes according to the age of the structured period; and the extent to which inter-deme migration lessens the effect. As we shall also see, altering the probability of lineages sorting into the same subpopulations can change the direction of the effect.

One approach to answering these questions is offered by Mazet et al. (2015) who analyse coalescent distributions in n-island models with symmetric migration rates [87]. (Critical of the concept of effective population size and several of its common applications, they also offer an alternative theoretical framework for interpreting the results of methods like PSMC. In order to limit ambiguity, I do not adopt their use of terminology here and refer the reader to the discussion on effective population size in Chapter 2.) Mazet et al. agree with Li and Durbin that structure inflates metapopulation N_e above the sum of subpopulation N_e values. However, they point out that high migration rates between islands will produce higher overall coalescent rates *relative to periods of low-migration.* They use this fact to demonstrate that it is possible to produce human-like PSMC curves using an island model with no change in subpopulation sizes and with only three changes in a symmetric migration rate. Notably, they use varying migration rates to mimic the effect of population *bottlenecks* on inferred coalescent rates. It is not obvious which biological or environmental phenomena could cause a global increase or decrease in symmetric migration rates between large numbers

of islands, but this effect could plausibly be caused by more complex mixing scenarios or perhaps through forms of population structure less simple than island models.

Non-equilibrium models The model studied by Mazet et al. assumes that a single form of demographic structure (an n-island model with varying symmetric migration rate and fixed n) persists throughout a population's history. This restriction can be reasonable for some species and samples if the average TMRCA across the genome is recent enough that major changes in structure are too old to affect current patterns of genetic diversity. In contrast, we have reason to believe humans did not evolve under sufficiently stable structured population conditions. Ancient DNA and other studies show that human populations have undergone periodic mass migrations, population splits, and range alterations until the present [e.g. 104]. Such demographic changes have altered the relationships between local populations in ways that shape present day variation [e.g. 46]. A noteworthy example of this is in Western Europe, where current populations are thought to have been formed through the admixture of several ancestral populations which had been relatively separate, possibly from soon after the out-of-Africa migrations until at least as recently as 10 kya [68]. It took the recovery of ancient DNA from these early populations to reveal the demographic change. This observation is a key motivation behind the techniques developed in the next chapter.

Evidence for a varying population structure in humans is also provided by historical cross-population coalescent rate estimates (similar to those produced in Chapter 3). For example, between samples of Yoruban and East Asian, or European, ancestry these analyses show low cross-coalescence from the present until the putative out-of-Africa migrations (around 50-150 kya) [81]. Earlier than this, the cross-coalescent functions exhibit a time-dependence similar to functions estimated from within-population samples of the same groups. Estimates of the ancestral N_e of individuals also coincide over this period [144]. If populations like the Khoisan are included, these sorts of analyses support the idea that present-day global population structure arose slowly over the last 200-250 ky [81]. We cannot say whether stable and long-standing structure existed before this period, though at a minimum the coalescent analyses show that humans today draw similar proportions of their ancestry from any putative ancestral populations, and suggests that there was a period of panmixia, or high inter-population migration, before the out-of-Africa migrations. Since all ancestral humans would presumably have been found in Africa (ignoring archaic hominin ancestors), evidence supporting early structure might yet be found through the recovery of more ancient DNA from the continent, if not from samples predating 100kya, perhaps from extinct, long-isolated populations. The arguments of Mazet et al., and Li and Durbin before them, illustrate the difficulty with straightforward interpretations of historical N_e . As a result of the preceding discussion, and in order to apply this insight to natural populations, we study the effect of

historically varying, or non-equilibrium, population structure on the distribution of pairwise coalescent times.

While the population genetic effects of simple models of population structure are well studied, the effects of complex varying structure are analytically less tractable and less easy to predict. The central problem is that a particular signal of some form of structure can require significant time to reach a detectable level, and may not persist when demographic conditions change. This difficulty has been noted in the context of classical measures of population divergence. It has been shown for example that the various summary statistics, such as F_{ST} or π, reach equilibrium values at different rates and as a result exhibit different sensitivity to changing demographic conditions [17]. The form of structure may also not affect many ancestral lineages, even if it persists for a long time. Equilibrium models cannot account for varying levels of shared ancestry in present-day panmictic populations. An example of this is found again in comparison of N_e curves of African and out-of-Africa populations. I noted above that the curves are indistinguishable before 150 kya, at least at the resolution of current methods. This is noteworthy because non-African populations are known to trace varying amounts of ancestry to the known archaic hominins, neanderthals and denisovans (up to $\sim 2\%$ for neanderthal and $\sim 5\%$ for denisovans) [140]. Since the division between hominins can be viewed as a form of long-standing population structure, it is perhaps surprising that there is no inflation in estimated N_e in non-African relative to African populations, before 150 kya. Without explicit modelling or simulation, we cannot say what level of admixture would be large enough to detectably perturb coalescent rates.

Finally, models with varying structure might also allow multiple demographic changes to occur simultaneously in a single population, and these can have complicated interacting effects on distributions of coalescent times. The most common approach to studying the effects of complex demography is to use methods based on the Approximate Bayesian Computation (ABC) inferential scheme [e.g. 31]. These methods, discussed in more detail in Chapter 2, are able to use sample summary statistics to fit rich demographic models with varying structure. With this flexibility, however, the possible parameterisations of the history greatly increase and it can in practise be difficult to choose among models in any principled manner. Ultimately, accounting for changing population structure in demographic inference is a considerably harder challenge, often better suited to study via simulation. Even then, alternative histories may not be possible to distinguish. It would therefore be useful to use analysis to reduce the parameter search space.

4.1 A model of historical substructure

In order to quantify the effect of historical population structure, I analyse a generalisation of the model by Li and Durbin described above. I use this to predict the effects of structure

on the distribution of pairwise coalescent times (also referred to as the "T_2" distribution) and to propose a way of evaluating interpretations of historical N_e curves. I apply this to an assessment of the kinds of substructure which could cause us to misinterpret previously inferred N_e curves in human populations. The model bears an analogy to that proposed by Mazet et al.: where they study varying migration rates, I look at varying numbers of islands.

The distributions of coalescent times in stable n-island models are well studied [87]. Here we determine a general form for T_2 distributions under a simple non-equilibrium model of structure. We assume that population size is constant throughout the population history, and assume panmixia over all times other than a single period during which the population is split into some fixed number of islands. These islands are of equal size, and their sum equals the size of the population in the panmictic periods. We assume no migration between islands during the period of splitting. This extends the Li and Durbin argument by considering any number of islands, and by looking at the full distribution of coalescent times over the entire history of a population. We are also interested in quantifying the effect of structure, in addition to looking at the direction of the effect. The assumption that population is constant during the panmictic periods is incorporated in order to simplify the analysis of the coalescent time distribution and illustrate most clearly the effects of an island period.

The model consists of three time periods:

1. $[0, T_S]$: a period of panmixia in which the effective population size of the population is scaled to 1; it lasts from the present until the "split time" T_S (time is scaled in units of $2N_e$ generations)

2. $(T_S, T_M]$: a period in which the population is divided into N islands and surviving pairs of uncoalesced lineages from the first period are randomly sorted into islands; unless otherwise stated we assume the random sorting is uniform across the islands; each island has population size $1/N$; there is no migration; and this structure lasts until the "merge time" T_M

3. (T_M, ∞): the last period, in which we return to the panmictic conditions of the first period

The assumption that no migration occurs between the islands in the second time period greatly simplifies the analysis of the coalescent distributions. To see why this is the case, observe that if two lineages have not yet coalesced by time T_S, they can only coalesce in the second time period if they have been (randomly) sorted into the same subpopulation. In the simplest scenario of uniform random sorting, the probability of ending up in the same island is $1/N$ (as pointed out in the 2-island case by Li and Durbin). Thus we reduce the problem to a 2-island non-equilibrium model, in which one island has size $1/N$, the other island has size $(N-1)/N$, and sorting probability corresponds to island size. This is illustrated in Figure 4.1.

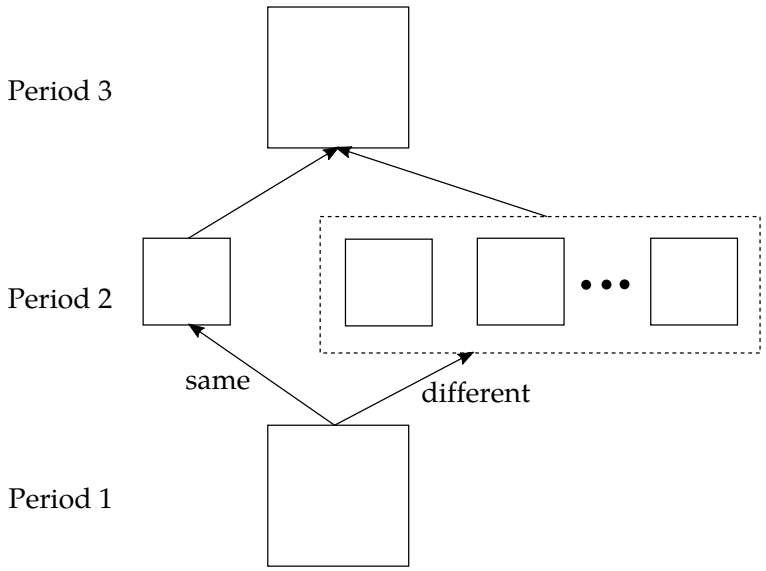


Fig. 4.1 An illustration of the population structure model studied in this section. It is a generalisation of the model looked at by Li and Durbin described above [70]. In this schematic illustration, the solid-line boxes represent subpopulations. In the first and third period the population is panmictic. In the second it is subdividied into N islands. The arrows indicate possible paths of coalescent lineages. In the transition to the second period the left arrow indicates the event that the lineages end up in the same island. In the general form of the model this occurs with probability p_s, though in the model where islands are chosen uniformly at random it is $1/N$. The right arrow is the complementary event that lineages choose different islands. The arrows may be interpreted as referring to the location of the second lineage given that the first lineage sorted into the leftmost subpopulation.

With this model, we can straightforwardly compute the distributions of pairwise coalescent times. Over the first time period, the probability that two lineages have not coalesced by time t is $1 - e^{-t}$ since we are assuming the first time period has population size 1. Following the standard coalescent argument, we scale time in units of $2N_e$. In the second time period, two lineages which have not coalesced sort into the same island with probability p_s. In this case $p_s = 1/N$. In the conditional case where the sorting is into the same island, the probability they would coalesce by time t (where $t > T_S$) is simply $1 - e^{-N(t-T_S)}$ since the population size of an island is $1/N$. Finally, in the last time period, once we have accounted for the probability that lineages survive uncoalesced until time T_M, the probability of coalescing by a given time t (where $t > T_M$) is simply $1 - e^{-(t-T_M)}$ since the population again has effective size 1.

The probability that lineages do not coalesce before T_M can be calculated by accounting for the probability that they do not coalesce in the first period, and the probability that they do not coalesce in either of the two cases in the second period (in the case of sorting into different islands there is no chance of coalescing). In the second time period, the conditional probability that they do not coalesce given that they end up on the same island is just $1 - e^{-N(T_M - T_S)}$. We can summarise these in the recursively defined distribution function,

$$F_I(t) = \begin{cases} 0 & \text{if} \quad t \in (-\infty, 0] \\ 1 - e^{-t} & \text{if} \quad t \in (0, T_S] \\ F_I(T_S) + p_s e^{-T_S}\left(1 - e^{-N(t-T_S)}\right) & \text{if} \quad t \in (T_S, T_M] \\ F_I(T_M) + \left(p_s(e^{-N(T_M-T_S)} - 1) + 1\right) e^{-T_S}(1 - e^{T_M - t}) & \text{if} \quad t \in (T_M, \infty). \end{cases}$$

With this we obtain the density function,

$$f_I(t) = \begin{cases} 0 & \text{if} \quad t \in (-\infty, 0] \\ e^{-t} & \text{if} \quad t \in (0, T_S] \\ p_s e^{-T_S} N e^{-N(t-T_S)} & \text{if} \quad t \in (T_S, T_M] \\ \left(p_s(e^{-N(T_M-T_S)} - 1) + 1\right) e^{T_M - T_S - t} & \text{if} \quad t \in (T_M, \infty). \end{cases} \tag{4.1}$$

We will be interested in comparing this history with one in which population size changes. To make the comparison as direct as possible, I analyse a three-period model in which the population has the same size in the first and last period, scaled to 1, and in the middle period allow N_e to vary by some scaling-factor λ. We are largely interested in increases of N_e during the middle period. The analysis of this model proceeds similarly to the one above, and it is

straightforward to show that the T_2 density of the "hump" model is

$$f_H(t) = \begin{cases} 0 & \text{if} \quad t \in (-\infty, 0] \\ e^{-t} & \text{if} \quad t \in (0, T_G] \\ \lambda^{-1} e^{(\lambda^{-1}-1)T_G - \lambda^{-1} t} & \text{if} \quad t \in (T_G, T_C] \\ e^{(\lambda^{-1}-1)T_G + (1-\lambda^{-1})T_C - t} & \text{if} \quad t \in (T_C, \infty). \end{cases} \tag{4.2}$$

Here we signify the beginning of the hump period as T_G and the end as T_C (for "growth" and "contraction").

The three population models used in this chapter can formally be described as follows.

The constant model, $K = (N_e)$: This is a baseline comparison model in which population size is fixed as N_e and the entire history is panmictic. Due to the choice of time scaling, its T_2 density is given by $f_K(t) = e^{-t}$.

The historical n-island model, $I = (N_e, N, T_S, T_M)$: This is the structured model as described above. Its density is given by $f_I(t)$, with corresponding time and population parameters (N is the number of islands and N_e the population size). Where time and baseline N_e can be understood from context, or are not pertinent, I will sometimes refer to this simply as I_N.

The hump model, $H = (N_e, \lambda, T_G, T_C)$: This is the main comparison model with density function provided above by $f_H(t)$, again with corresponding time and population parameters. Similarly to the island model, where time and baseline N_e can be understood from context, this model will be referred to as H_λ.

The effects on coalescent rates of historical island structure In Figure 4.2, I use simulations and Equation 4.1 to show the difference between the T_2 distributions of models I_2 and K. The panmictic population has the same size (and sum of island size) as the structured. The split in I_2, $T_M - T_S$, lasts between 1 and 2 in coalescent units, equivalent to 0.5-1 Mya if generations last 25 years and $N_e = 10^4$. Observe that in the first time period the coalescent rate decays at the same rate in both cases since the population histories over this period are identical. On the second time period, there is a noticeable reduction in coalescent rate under I_2 relative to K. This is followed at T_M with a spike in the number of coalescent events since lineages which were prevented from coalescing during the island period have moved into the same population. Consistently thereafter, the number of coalescent events is greater in the structured history, a result of the "missing" coalescent events during the island period. Shown along with these theoretical curves is a histogram of observed coalescent times in the simulation of 50000 independent coalescent trees, with sample size 2 and conditions identical to the structured history.

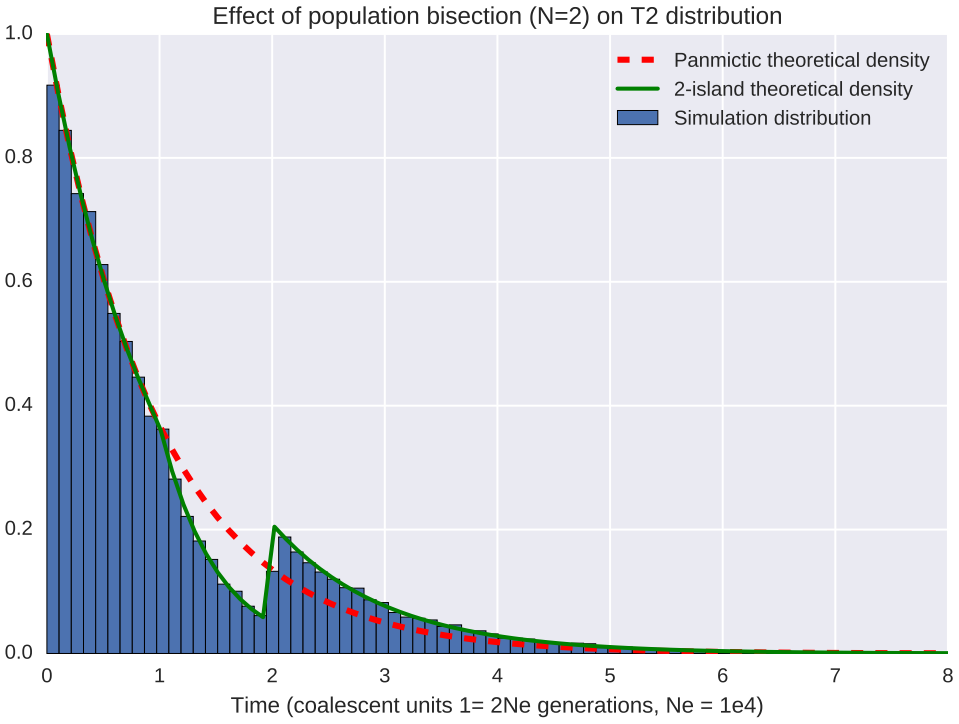


Fig. 4.2 Differences between coalescent distributions of I_2 and K. The effects on pairwise coalescent time distribution of population bisection lasting from 500 kya to 1 Mya (1-2 in coalescent units, assuming $N_e = 10^4$ and generation time is 25 years). The blue histogram summarises observed coalescent times of 50000 independent 2-coalescent processes simulated under subdivided history using `msprime` [58]. The green line is a theoretical prediction using density function f_I shown in Equation 4.1. The red dashed line is coalescent time prediction under panmictic model K with consistent $N_e = 10^4$.

As suggested in the discussion which started this chapter, we can straightforwardly extend the argument we used for I_2 to illustrate the effects of higher numbers of islands. Using f_{I_N} with $N > 2$, and assuming that lineage sorting into islands is still uniformly random, we can readily obtain the distributions of pairwise coalescent times. These are shown in Figure 4.3. As expected, the direction of the effect on coalescent times is the same when more islands are allowed and the magnitude of the effect increases: the probability of sorting into the same island reduces as N increases, while the reduction in size of the islands, and presumed increase in coalescent rates of lineages which do sort into the same island, do not compensate for the effects of barriers to coalescence. Also observe that a scenario in which there are more than 16 islands will not have a much greater effect on the coalescent distributions than that shown here. The reason is that the chance of coalescence on the middle interval is already very low as a result of the small chance that lineages end up in the same island. Additional islands will admit fewer coalescent opportunities though the difference will be relatively smaller, and with some number of additional islands the distribution of coalescent times will appear indistinguishable from a model in which there is a complete stop in coalescence over the second time period.

As a useful comparison, I use simulation and Equation 4.2 to show in Figure 4.5 a similar plot generated under a history in which population size is doubled over the same period during which the previous history was subdivided. This model, H_2, is illustrated in Figure 4.4. Observe that the broad features of the plots are similar, in that over both middle periods there is a sharp reduction in coalescent rate and thus an increase in effective population size. The shape of the depression in coalescent rates demonstrates some visually distinguishable features. In comparison with I_2, for example, the reduction in coalescence in the hump model is initially more drastic at the start of the middle period, but remains at a higher level through the period. At the end of this period, the change in coalescent rate is less drastic than the change in rate at the time of the merging of separated populations.

These figures illustrate close agreement between theoretical prediction and simulation. They also show that while the direction and broad magnitude of the effects of population size change can be quite similar to the effects of structure, with sufficiently high resolution it might be possible to visually distinguish between the scenarios in idealised and extreme cases such as the situation above. However, it seems unlikely that this sort of resolution would be obtainable in practise, and it will be difficult to interpret slight differences in a context in which several demographic changes are occurring simultaneously.

4.2 An approach to identify plausible structured models

I propose a strategy for identifying a range of structured island histories consistent with estimated distributions of coalescent rates. Since it will always be possible to explain increases

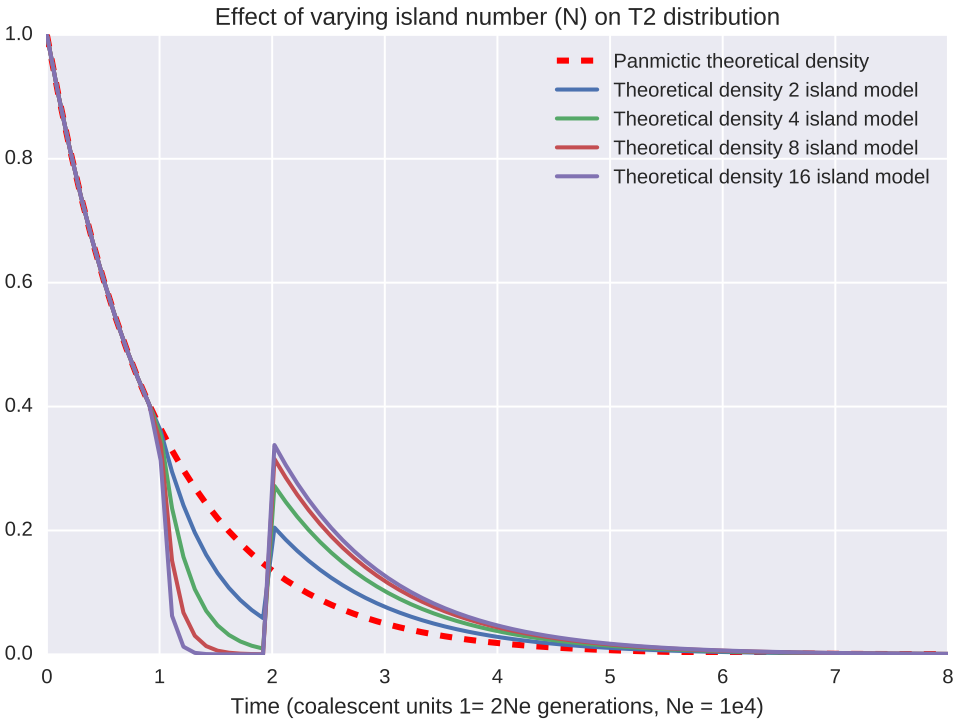


Fig. 4.3 The effects of multiple islands (I_N with $N \geq 2$) on distribution of pairwise coalescent times. Each line represents the theoretical prediction of a distribution using density function f_{I_N}, shown in Equation 4.1. The lines differ by the number of islands simulated. Each history assumes uniform random distribution of lineages into islands at split time. Here split time is 500 kya, and merge time is 1 Mya (1-2 in coalescent units, assuming $N_e = 10^4$ and generation time is 25 years).

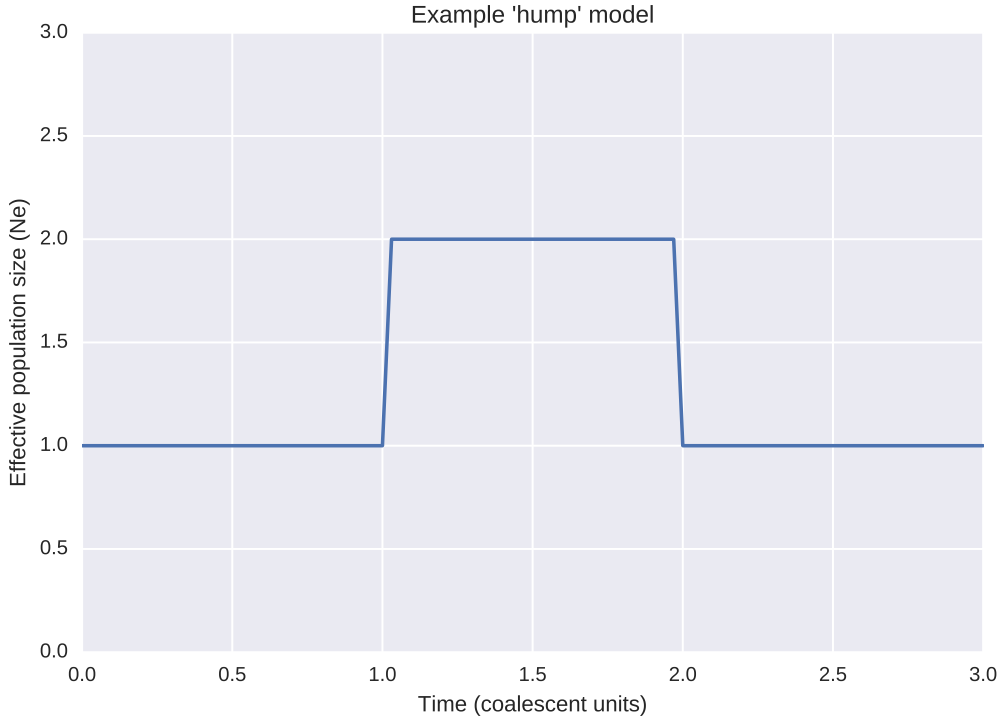


Fig. 4.4 Illustration of the comparison hump model, H, in which population size scales by factor λ between times T_G and T_C. In this example $\lambda = 2$ on the time interval (1,2) in coalescent units.

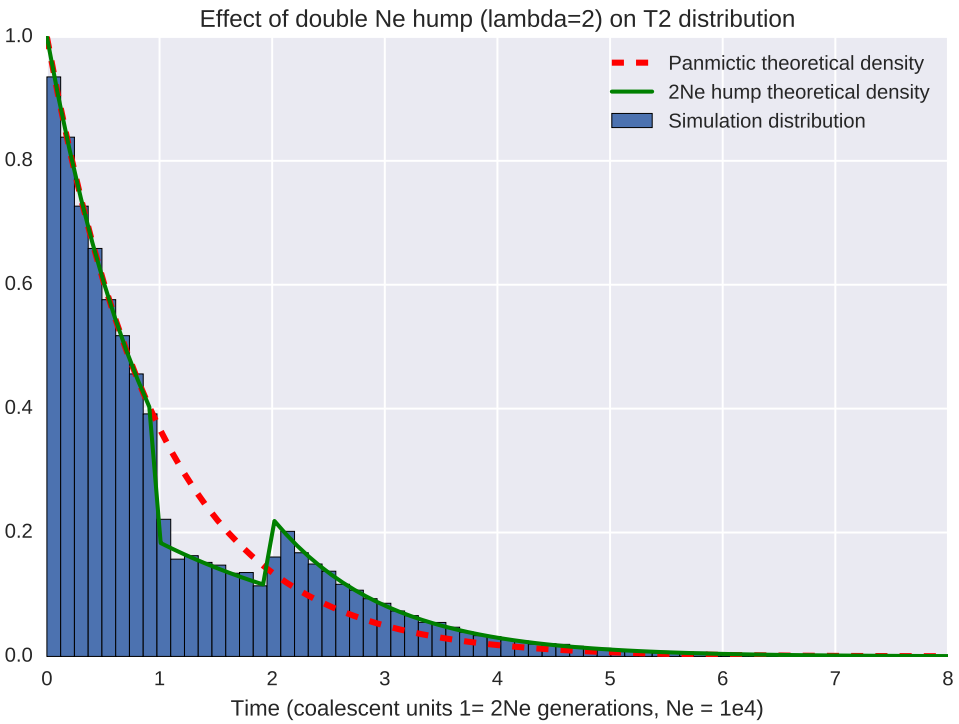


Fig. 4.5 The effects of H_2, showing the effects on pairwise coalescent time distribution of population doubling in size from 500 kya to 1 Mya (1-2 in coalescent units, assuming $N_e = 10^4$ and generation time is 25 years). The blue histogram summarises observed coalescent times of 50000 independent 2-coalescent processes simulated under H_2 using msprime [58]. The green line is theoretical prediction using density function f_{H_2} shown in Equation 4.2. The red dashed line is coalescent time prediction under panmictic history with consistent $N_e = 10^4$. Simulation conducted using MSPRIME [58].

in effective population size with a period of island structure, we would like to determine a minimum number of islands necessary to produce certain observed increases. Using only the modelling above we will be able to determine no more than this minimum number of islands, since migration will increase rates of coalescence and reduce the effect of larger numbers of islands, and thus the ability to detect them. For the purposes of deciding whether some change in N_e is a result of a change in census population size or population structure, this need not be a problem. This minimum number of population divisions can be compared with other sources of information to determine the plausibility of different explanations for changes in coalescent rate. I note however, that some histories will still not be distinguishable using this approach on available data.

The approach is as follows. Assume that we are trying to decide whether some increase in observed N_e is the result of population structure. Observed N_e will likely be estimated using a method like PSMC. It is also possible to use the so-called decoded output of a method like PSMC, which produces the raw estimates of coalescent rates before they are converted to effective population sizes. However, since the approach described here is likely to be used without access to the decoded data, such as when the results published by others are explored, we present it in a form which can be used in more settings. Using these N_e estimates, we obtain a measure of the divergence between two coalescent distributions: an H model featuring an increase in population size similar to that observed; and a constant-sized population model K which we choose as the baseline comparison. We apply the Kullback-Liebler (KL) divergence, commonly used in information theory as a measure of divergence between probability distributions, discussed below. There is some freedom involved in choosing the size parameter of K, and we discuss implications of this choice in applications below. After determining the divergence, we try to identify the minimum number of islands, and time of separation, which would be required to induce a similar level of divergence between the coalescent time distributions of the kind of structured models analysed above, and the baseline panmictic model chosen. In addition to identifying the size of the hump (λ) which we would like to interpret, we also locate an approximate time at which the hump peaks, and use this as the central time around which we vary the length of the island structure when identifying suitable island models.

More formally, the approach can be expressed as follows:

1. Identify hump of interest in estimates of historical rates of coalescence using, for example, PSMC

2. Choose a baseline population size N_e^b and estimate the scale of the hump λ, as well as the times T_G and T_C over which the hump will be assumed to exist. From these, also determine the point in time at which the hump has its central point, or peak, usually $t_p = (T_C - T_G)/2$.

3. Calculate the KL divergence of distribution times describe by f_H above, compared to a constant-sized coalescent distribution f_K. Use the parameters for the hump model chosen above, and assume the population is constant-sized and panmictic outside of the hump, with effective size N_e^b.

4. With various numbers of islands N, determine the length of time interval $(T_M - T_S)$, centred at t_p, which would be required to produce a KL divergence similar to that observed above. Here the divergence is between the density f_{I_N} of an island model I_N and f_K.

5. Compare matching island numbers and times with external evidence for plausibility.

As discussed above, Mazet et al. illustrate a way in which population structure can cause spurious signals of bottlenecks, not just expansions. The approach proposed here does provide a means of studying this scenario, even though it generally assumes a population expansion ($\lambda > 1$), not contraction. The reason is that the apparent signal of a population bottleneck will, under the alternative explanation, be caused by an increase in the symmetric migration rate between established islands, relative to the migration rate preceding the bottleneck. As a result, we can explore the kinds of plausible island structure which might produce the hump preceding the bottleneck. We illustrate an example of this below, in the context of human population history, looking at the signal of population contraction usually associated with the out-of-Africa migration.

One simplifying assumption we make here is that the history before or after the period of structure, in I, or the hump, in H, does not vary in size or structure. This condition does not hold in any population we are interested in, but since we are comparing I and H models against the same K, it will matter less what the absolute change in coalescent rate induced by either situation is, and more what they are relative to each other.

Kullback-Liebler divergence In the approach used to measure the divergence between coalescent distributions under structured and hump models we apply the widely-used Kullback-Liebler (KL) divergence, also known as the relative entropy. To illustrate this, if two random variables F and P have continuous density functions f and p, then the relative entropy *from* P to F is determined by the expression

$$D_{KL}(f||p) = \int_{-\infty}^{\infty} f(t) \log\left(\frac{f(t)}{p(t)}\right) dt \tag{4.3}$$

As reflected in the language used, the measure is not a true metric since it is not symmetric. We will in general use the T_2 density of some constant model K as the divergence *from which* other densities will be compared. In other words, when we refer to the divergence between the coalescent distribution of a structured model and that of a panmictic, constant-sized one,

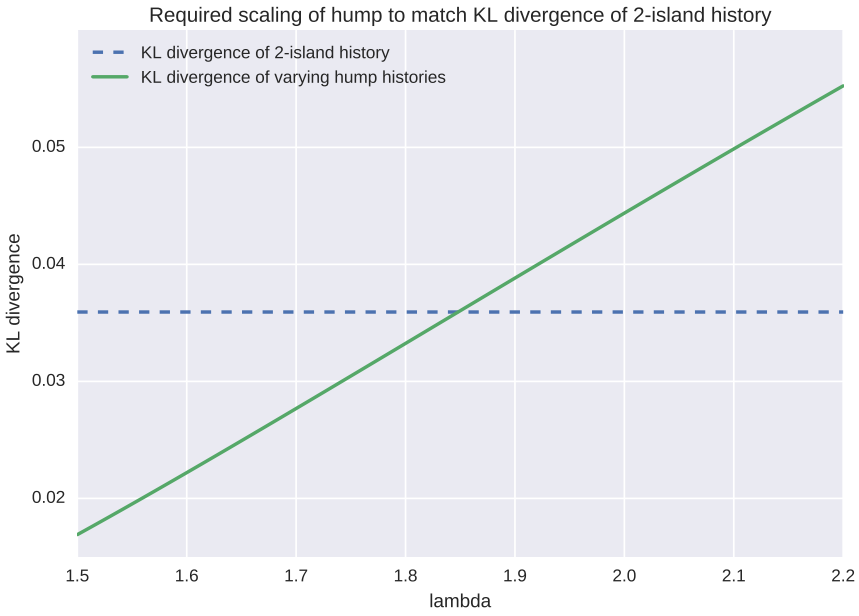
the latter model will always correspond with P in the expression above. Note that instead of writing $D_{KL}(f_{I_N}||f_K)$ for the Kullback-Liebler divergence from f_K to f_{I_N}, I will indicate this measure with $D_{KL}(I_N||K)$, and use an analogous shorthand for divergence measures from K to H models.

To illustrate the comparison of models using KL divergence values, I show two distributions of coalescent times in Figure 4.6b. These models have matching KL divergence values. One is an I_2 model, and the other is an H in which the N_e increase of the middle period of the history was determined in such a way that the KL value of the distribution is identical to the structured model, in other words $D_{KL}(H_\lambda||K) = D_{KL}(I_2||K)$ (Figure 4.6a). The middle periods of both models occur between coalescent time units 1-2, which can be translated into years as described in Figure 4.2. The size of the hump population on the interval (T_G, T_C) is $\lambda^* N_e$ where $\lambda^* = 1.85$ and $N_e = 1$, this value is obtained in the previous figure. Observe that the shape of the curves is different over the periods where their demographic structure differs. However, the curves are identical in the last period, when both populations are panmictic and constant-sized. This suggests that the proportion of "missing" coalescent events are the same in both histories at the time when their middle periods end (where "missing" here means relative to the proportion of coalescent events in that period in the baseline K model).

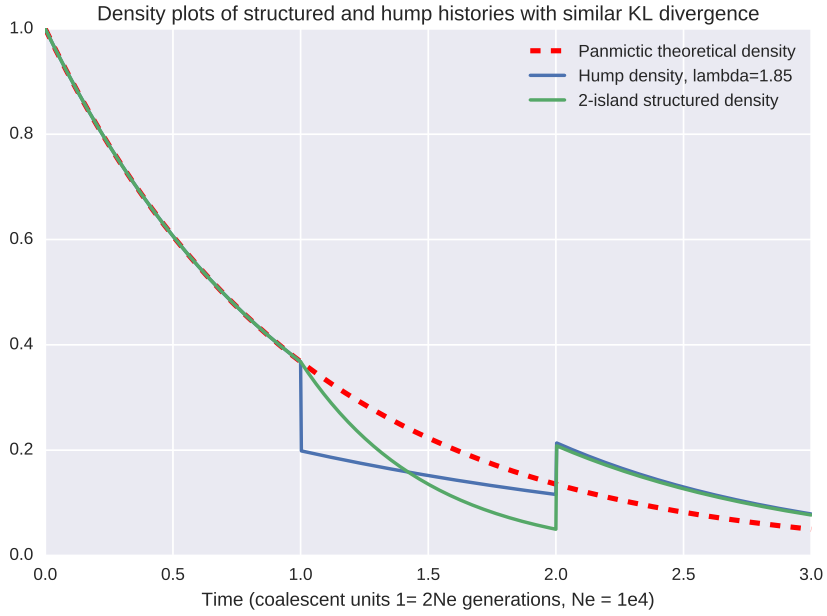
Effects of time and start of separation on structure-induced divergence Before looking at empirical applications of the approach, we briefly illustrate the interacting effects of the time of onset and length of separation of islands in the island model described above. This is equivalent to studying the effects of varying the endpoints of the interval (T_S, T_M) in the distribution f_I. We compare this with panmictic models using the KL divergence discussed above.

The broad trend is the same in Figure 4.7 and Figure 4.8. Each of the four plots in this pair of Figures shows KL divergence values under varying time regimes, with structured periods starting at different time points and lasting for different lengths of time. The different figures correspond to different numbers of islands (N) existing during the structured periods (of models I_N). The diagonal trend in KL values shows an increase in divergence from f_K when the structured periods are earlier and when they last for longer lengths of time. The greater effect on earlier events is due to there being more pairs of lineages available which can be affected by barriers to coalescence. Note also that an increase in the number of islands will cause a greater depression in coalescent rates, and thus induce a greater divergence value. This is to be expected from Figure 4.3.

For comparison sake, a similar effect is also seen in PSMC analyses of sequences produced under simulation. In Figure 4.9 we see PSMC-estimated N_e curves of various simulations of I_2 models. The base simulation in this case was chosen as an idealised model of the population size history of an out-of-Africa population. It approximates a population bottleneck like

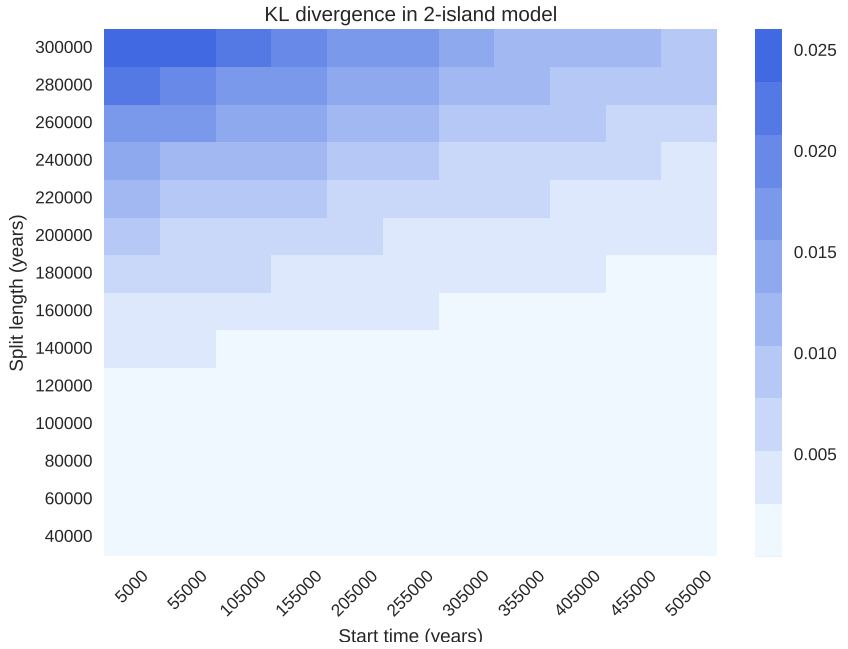


(a) Determining λ so that $D_{KL}(H_\lambda||K) = D_{KL}(I_2||K)$

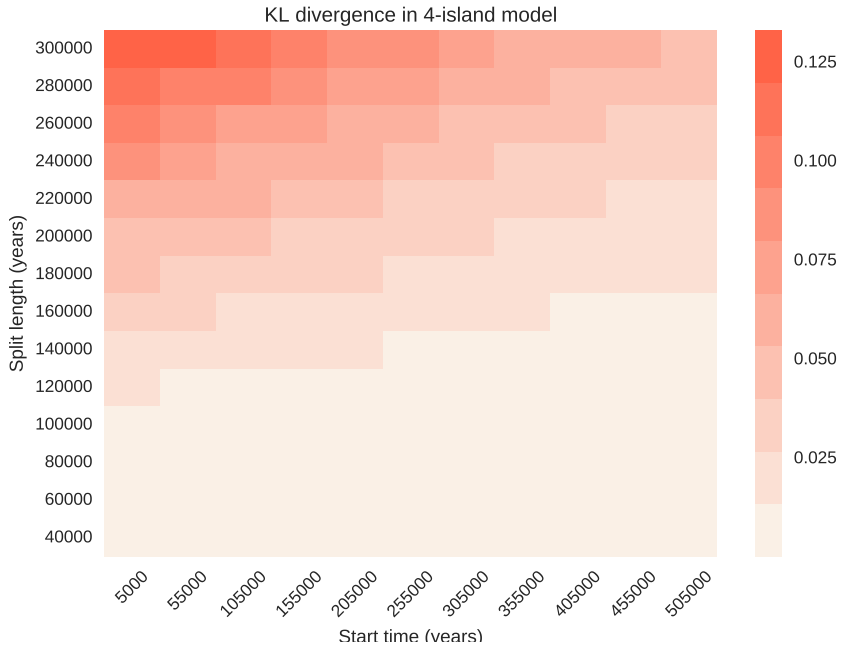


(b) Plot of theoretical densities of I_2 and H model T_2 times with $D_{KL}(H_\lambda||K) = D_{KL}(I_2||K)$

Fig. 4.6 An illustration of I_2 and H_λ models with $D_{KL}(H_\lambda||K) = D_{KL}(I_2||K)$. In (a), the blue dashed line represents the KL divergence of an I_2 model in which the split occurs between 1-2 in coalescent time units (to translate into years, see the caption to Figure 4.2). The green line represents the value of the KL divergence of H_λ. Intersection is at $\sim \lambda^* = 1.85$. In (b) we see the distribution of coalescent times under both the I_2 (green) and H_λ with $\lambda = \lambda^*$ (blue). For comparison, the distribution of the corresponding K model is also plotted.

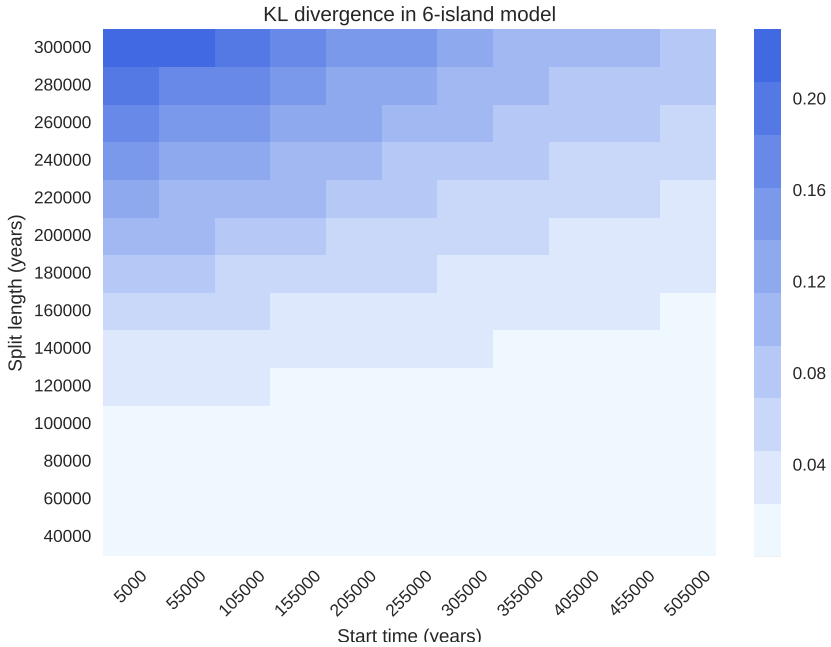


(a) $D_{KL}(I_2||K)$ under varying time regimes

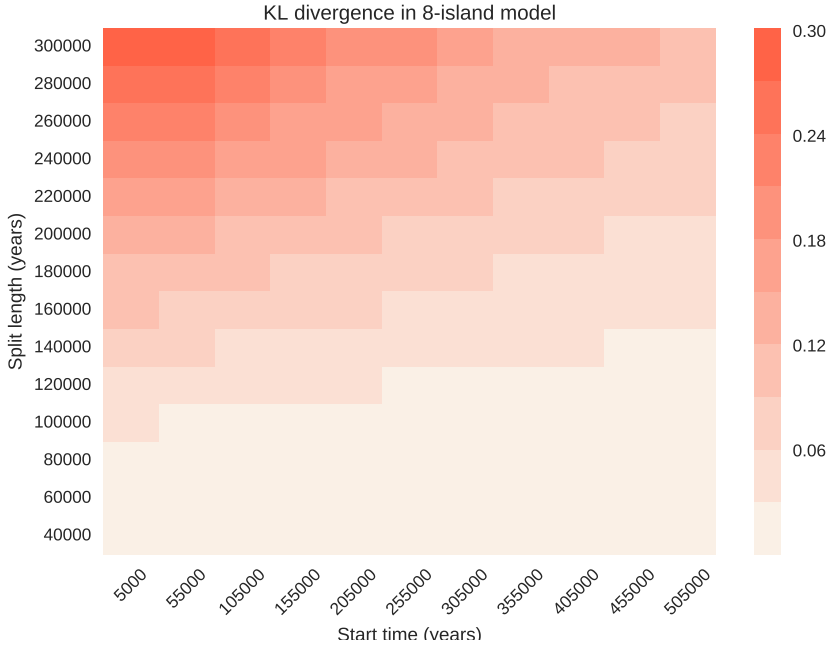


(b) $D_{KL}(I_4||K)$ under varying time regimes

Fig. 4.7 Heatmap showing values of Kullback-Liebler divergence in island models, $D_{KL}(I_N||K)$, using coalescent distribution times given by f. Figure (a) features an I_2 model, and (b) an I_4 model. Both show the effect of varying the split time interval (T_S, T_M). Assumed baseline $N_e = 10^4$, and conversion from coalescent times assumes generations are 25 years. The colour difference between the two plots indicates only that the scaling is different.



(a) $D_{KL}(I_6||K)$ under varying time regimes



(b) $D_{KL}(I_8||K)$ under varying time regimes

Fig. 4.8 Heatmap showing values of Kullback-Liebler divergence in island model, $D_{KL}(I_N||K)$, using coalescent distribution times given by f. Figure (a) features an I_6, and (b) an I_8 model. Both show the effect of varying the split time interval (T_S, T_M). Assumed baseline $N_e = 10^4$, and conversion from coalescent times assume generations are 25 years. The colour difference between the two plots indicates only that the scaling is different.

the one experienced by humans of European or Asian ancestry. The bisection in population occurs at various times preceding the bottleneck and lasts for various lengths of time, as indicated in the figures. The effect shown is as expected from the KL divergence plots: a smaller distortion in coalescent distribution as the length of the split period is shorter, and as the start of the split is pushed further back in time. All simulations were conducted here using SCRM [157].

This effect is also seen in Figure 4.10, a heatmap which summarises the trend, and exhibits a similar diagonal pattern to that seen in Figures 4.7 and 4.8. The comparison model is the same idealised out-of-Africa model used in the previous figures. Note that the measure of divergence in this case, between the estimated historical N_e and the true history, is not the KL divergence. It would be inappropriate to use a measure designed for the comparison of probability distributions on unnormalised estimates of historical N_e . We use instead an integral measure, which is identical to one used in the original publication of PSMC. Here it is called the "scaled fractional difference" and, using the original notation on the time interval $[t_0, t_1)$, it is given by

$$d(t_0, t_1) = \frac{1}{\log t_1 - \log t_0} \int_{t_0}^{t_1} \frac{|N_0(t) - N_1(t)|}{N_0(t) + N_1(t)} \frac{dt}{t} \tag{4.4}$$

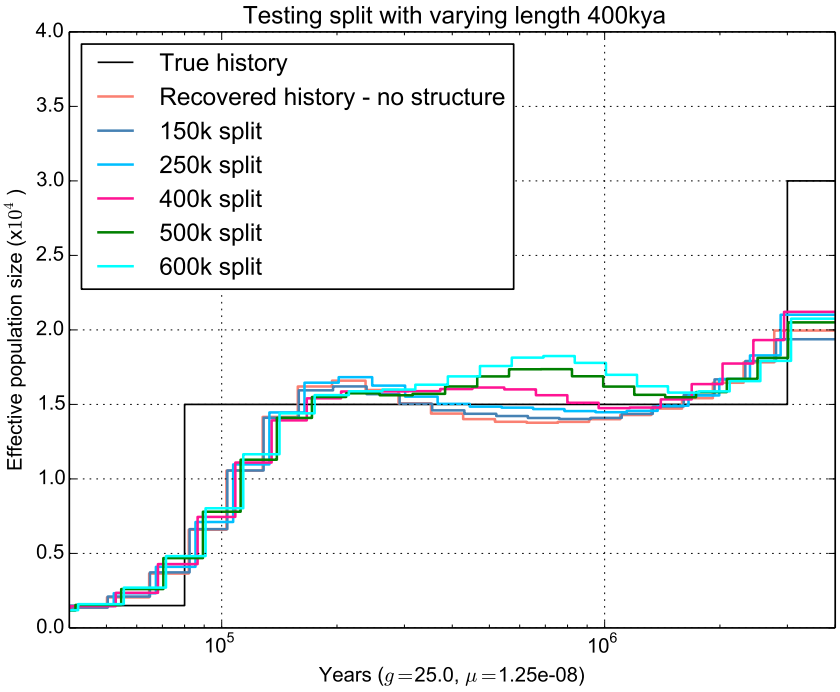
where N_1 and N_0 are the population size curves being compared.

It is evident from both these sets of analyses that island structure lasting for a short period of time will not appreciably alter coalescent rates. The relevant length of time depends on the sensitivity of the inference scheme, the number of islands, and the structured period's proximity to the present.

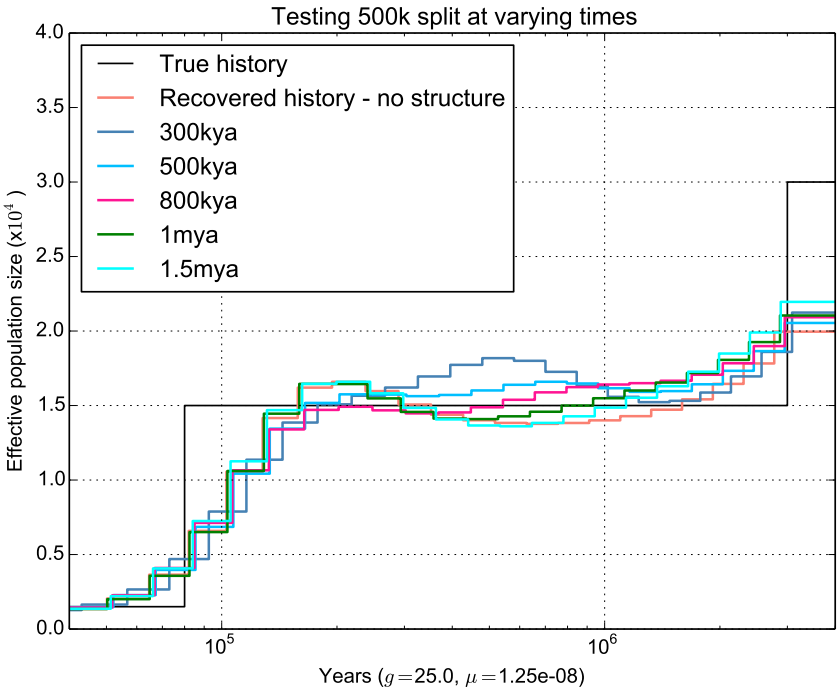
4.3 Applications of the approach

Population structure in central chimpanzees As a first application of this approach we examine a question which arose in the previous chapter in the context of the population history of chimpanzees. We observed in Figure 3.15 that the central chimpanzees experienced a large increase in population size within the last 100 kya and this was followed by a decrease closer to the present. Because of the increased variance between individuals over this time period, we were concerned that this increase was due to the presence of some kind of population structure, rather than to an increase in census population size. Recall that although this effect is diminished in subsamples consisting only of high-coverage individuals, it is not entirely absent from this data (Figure 3.16a).

One way to detect cryptic population structure on this recent time period would be to use the cross-population coalescent rates between various subgroups of individuals in a single present-day population. The time periods over which these inferred curves differ significantly



(a) Analysis of the effects of varying split time



(b) Analysis of the effects of varying start of split period

Fig. 4.9 PSMC-estimated N_e history of simulated structured histories. In (a) we see the inflationary effect of lengthening the time of the structured period. In (b) we see a similar change in the estimated historical N_e if we bring structured periods closer in time to the present. True history in both cases is an idealised version of the apparent population bottleneck history observed in out-of-Africa human populations. Simulations conducted using SCRM [157]

.

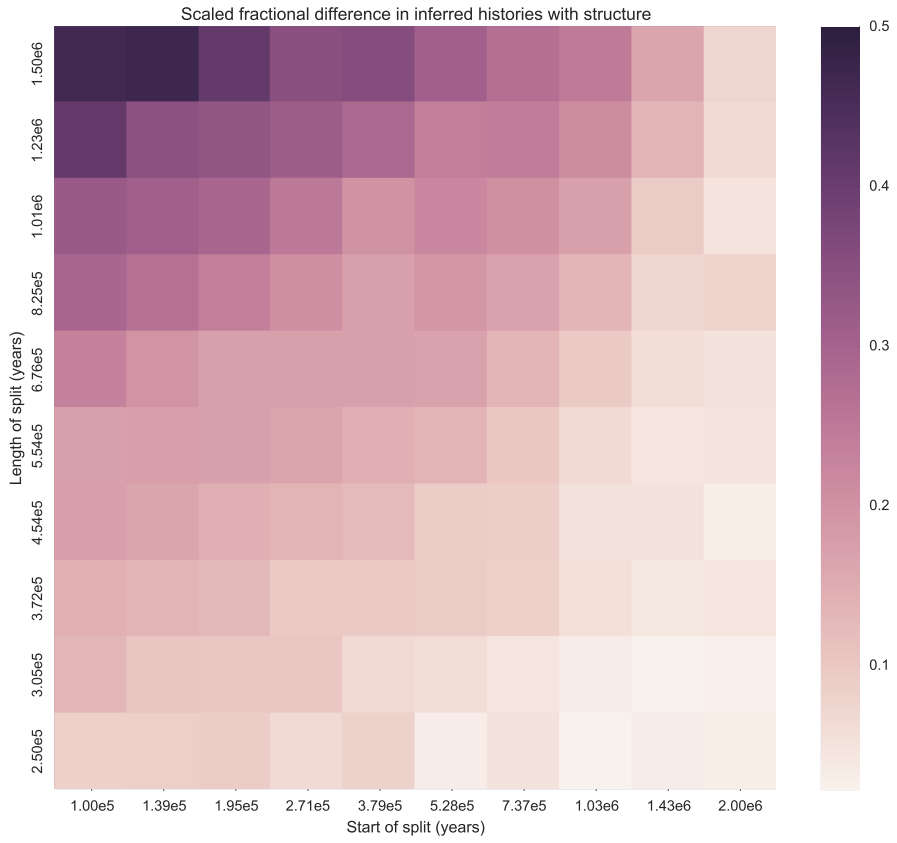


Fig. 4.10 Heatmap showing time-dependence of divergence between various PSMC-estimated 2-island N_e histories and the model of a panmictic population. True, panmictic history is idealised out-of-Africa human population model, shown in Figure 4.9 above. Generation time is 25 years and mutation rate, $\mu = 1.25 \times 10^{-8}$. For details of simulations and divergence measure between N_e curves ("scaled fractional difference") see main text.

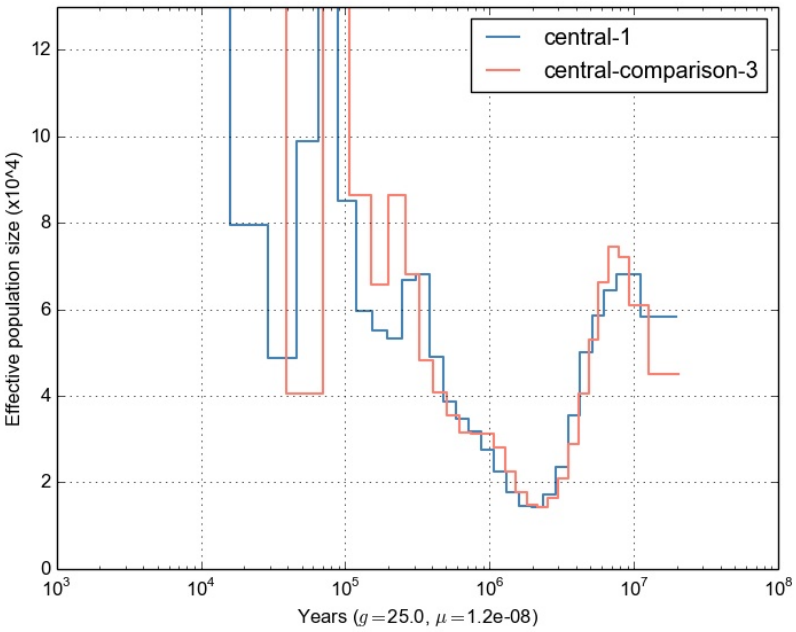
from each other, or from the autosomal effective population size of an individual in the population, might be those periods with uneven rates of coalescence across the population. This could be caused by low migration rates between subregions and indicative of the sort of structure relevant to the island models in this chapter.

However, this approach is not always conclusive. An example of this is shown in Figure 4.11. In Figure 4.11a I selected two subgroups of Central chimpanzees based on clustering patterns taken from a FINEstructure analysis [67] undertaken by de Manuel et al [24]. The first "central-1" compares gene flow between those in Gabon East and the rest, while the second "central-comparison-3" compares two groups which showed no indication of being distinct from each other, outside of Gabon-East (see Figure 3.18 and [24]). A similar analysis is shown in Figure 4.11b. Here the first comparison is between Eastern chimpanzee males from DRC-South and Tanzanian versus Eastern males from outside those regions, while the second is on randomly chosen individuals drawn strictly from the second group, where individuals cluster together.

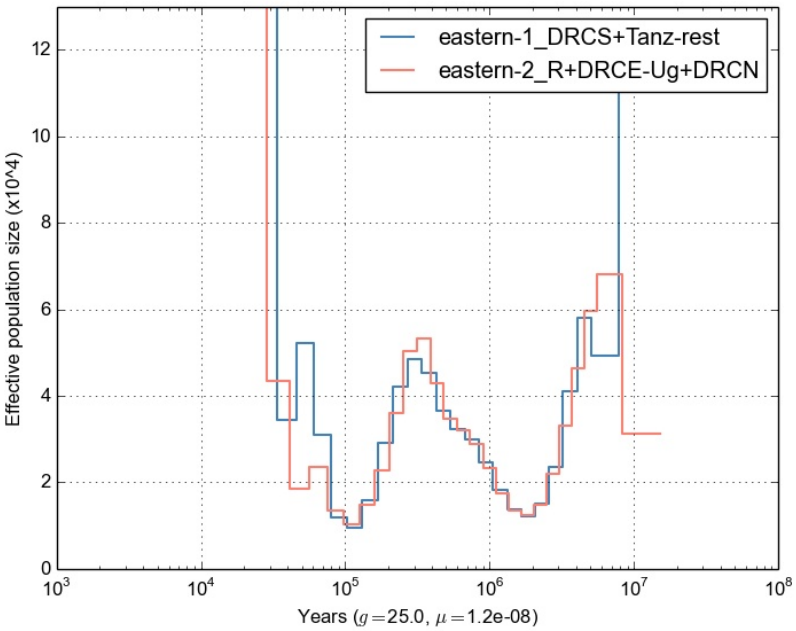
Taken alone, the first (blue) curves in either plot might seem to suggest a decrease in gene flow, stopping entirely around 20-40 kya. However, the fact that this pattern is also seen in the analysis of individuals which are indistinguishable in other clustering methods suggests this effect may be an artefact. Methodological limitations prevent easily interpretable conclusions being drawn from the amount of data available. We cannot say that the substantial increases in the most recent time periods are not merely due to a paucity of coalescent events on that time scale. On the other hand, it is possible that the application of FINEstructure misidentified the likeliest division of individuals into structured populations and that in fact both plots show evidence of genuine decrease in within-subspecies gene flow. Further analysis, based on more and better quality male X chromosomes, or, ideally, high-quality phased autosomal data, may make this approach feasible.

A last exploration of possible substructure does not involve MSMC2, but is based on running PSMC on simulated sequences drawn from panmictic and substructured populations. In an attempt to replicate the most recent expansion seen in each chimpanzee subspecies, I formulated a simplified version of a chimpanzee population history, and simulated sequences under this demographic history. In comparison, simulations were also conducted under a history similar to this, but with a population bisection and no migration between subpopulations over recent time periods. I tried to replicate the hump seen in the chimpanzee curves using substructure alone. As can be seen in Figure 4.12, even a population bisection as long as 80 ky does not sufficiently alter the coalescent rates and produce the observed increase.

These plots are suggestive, but it would be ideal to have a more principled way of determining the plausible range of structured models which might produce such increases. Applying the approach proposed in this chapter to examine this scenario, we characterise a hump model H over the period of apparent population expansion. This requires specifying



(a) MSMC cross-coalescence analysis of gene flow between male Central chimpanzees from Gabon-East and all other male central chimpanzees (central-1), and between randomly chosen males from outside Gabon-East who cluster closely in tests of extant population structure (central-comparison-3).



(b) MSMC cross-coalescence analysis of gene flow between male Eastern chimpanzees from DRC-South and Tanzania versus the rest (eastern-1...), and then between Eastern males randomly chosen from outside DRC-South and Tanzania (eastern-2...).

Fig. 4.11 Investigation of substructure based on gene flow within chimpanzee subspecies. Groupings are derived from putative clustering inferred by FINEstructure analysis conducted by de Manuel et al. [24].

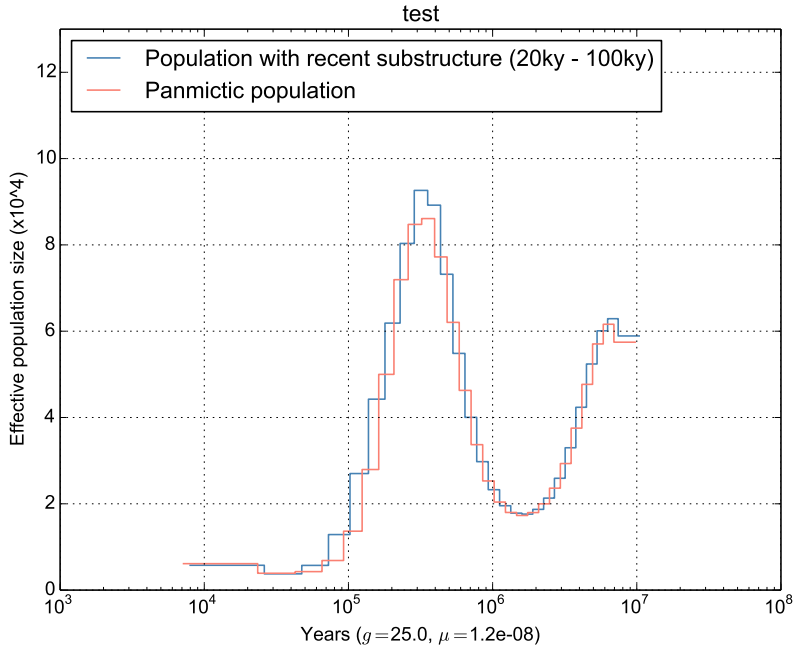


Fig. 4.12 Running PSMC on two simulations based on a schematic version of the chimpanzee population history. Output from substructured history in blue is indistinguishable from panmictic history in red. Substructure in this case refers to bisection of the population into subpopulations with summed effective population size equal to panmictic size and no migration. Further discussion of this question is presented in Chapter 4.

the end points of the increase time T_G, a baseline population size for the comparison model K, and the scale of an increase, λ. We take the peak of the hump to be 6×10^4, and produce analyses for a baseline N_e of 5000 and 10000, corresponding to a scale increase of $\lambda = 12$ and $\lambda = 6$, respectively. We assume conservatively that the hump period lasts from 20000 to 60000 years ago, with the peak t_p at 40000 years ago (and assume a generation time of 25 years).

In Figures 4.13 and 4.14 we show the KL divergence curves of various I_N models under these two baseline population size scenarios. These plots suggest that the time available is too short for island structure to sufficiently depress coalescent rates to produce the observed pulses in PSMC curves. Barring more complicated demographic histories, it seems most likely that the increase in N_e is due to genuine changes in census population size.

Human demographic history In Figures 4.16 and 4.17 we show an attempt to determine whether the apparent bottlenecks seen in populations which underwent a putative out-of-Africa migration, can be explained by island structure. As described above, this is a matter of exploring whether the hump preceding the contraction is caused by islands between which there was little migration. In this scenario, the signal of the bottleneck is largely a result of a subsequent increase in migration between islands or a complete transition to panmixia.

For this section we use the effective population size histories estimated by the Simons Genome Diversity Project [81]. In their Figure 2f, reproduced here in Figure 4.15, they collect representative PSMC curves for several out-of-Africa populations, including French, Han, and Mixe. The curves exhibit the characteristic decrease in N_e to 50kya, from a high point around 300kya. As mentioned above, this is widely interpreted as the population contraction caused by migration from Africa [e.g. 79], partly because it is not evident in the same degree in populations with exclusively African ancestry [144, 81].

Applying the approach proposed here to explore structured histories, we need to characterise a hump model for the period before the contraction. We choose the conservatively small time interval (200 kya, 400 kya), with a peak $t_p = 300$kya. This may not be unrealistic, since PSMC tends to smooth out large sudden changes in population size (see for example the inferred histories in Figure 4.9). In Figures 4.16, 4.17, and 4.18 I show plausible island histories I_N with three choices of baseline N_e (2000, 5000 and 10000 respectively) in each of which I have fixed the height of the hump in H_λ at 20000 and adjusted λ accordingly. This corresponds to expansion size scalings of 10, 4, and 2, respectively. In order to allow some uncertainty I have shown in each Figure the values of island time lengths corresponding to three values of λ centred on these values. Note that the structured histories assume the same baseline effective population size as the corresponding hump model in each plot, in other words the comparison between corresponding I and H models is to the same model K.

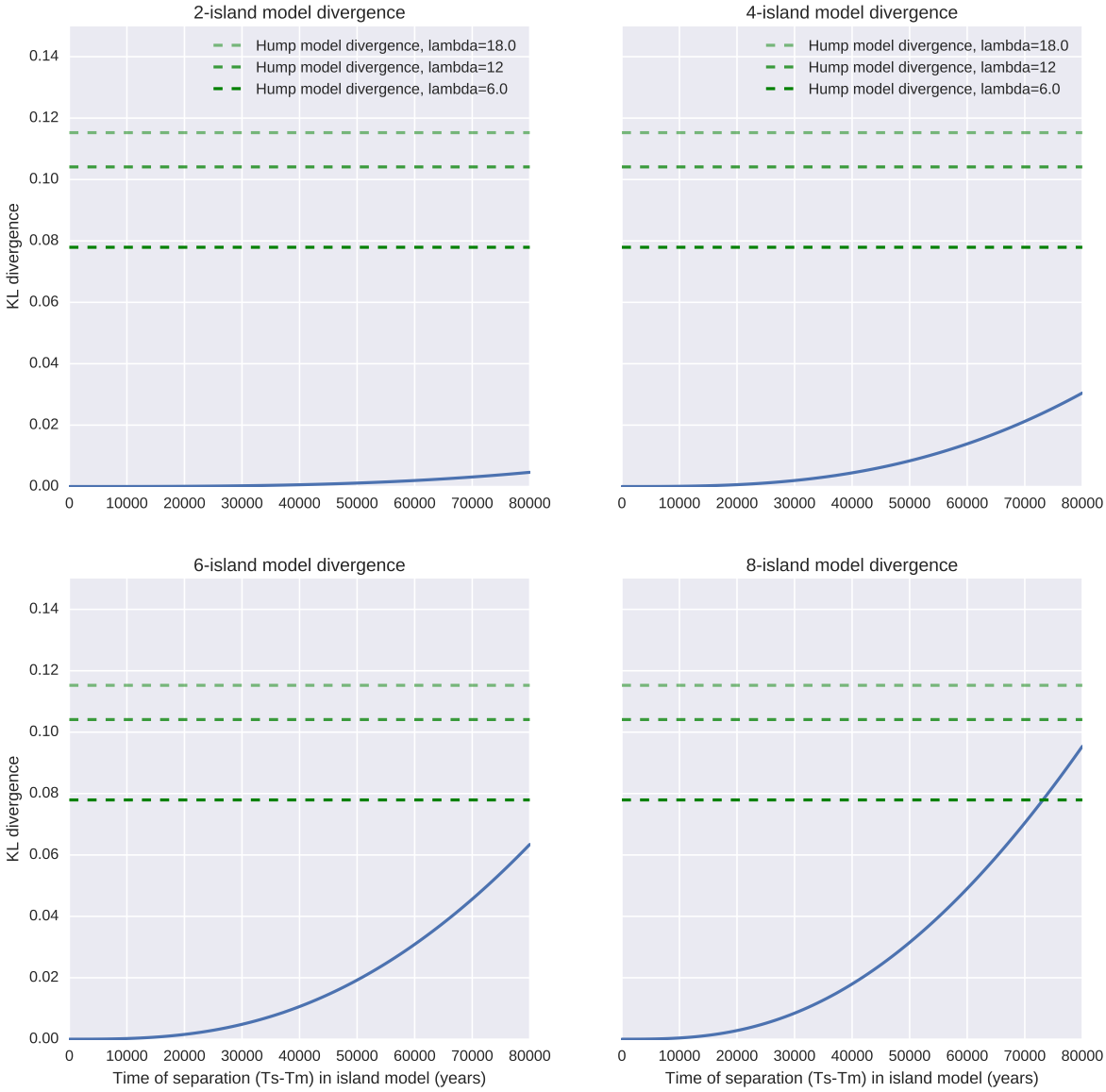


Fig. 4.13 Island models I_N consistent with central chimpanzee population expansion when baseline $N_e = 5000$. Figures represent dependence of KL divergence values on length of island intervals centered at $t_p = 40$ kya. Intersections with dashed green lines represent matching KL values with hump models in which increase lasts from 20-60 kya. Scale of the population increases in the hump models H_λ vary according to λ values associates with dashed lines (see legend).

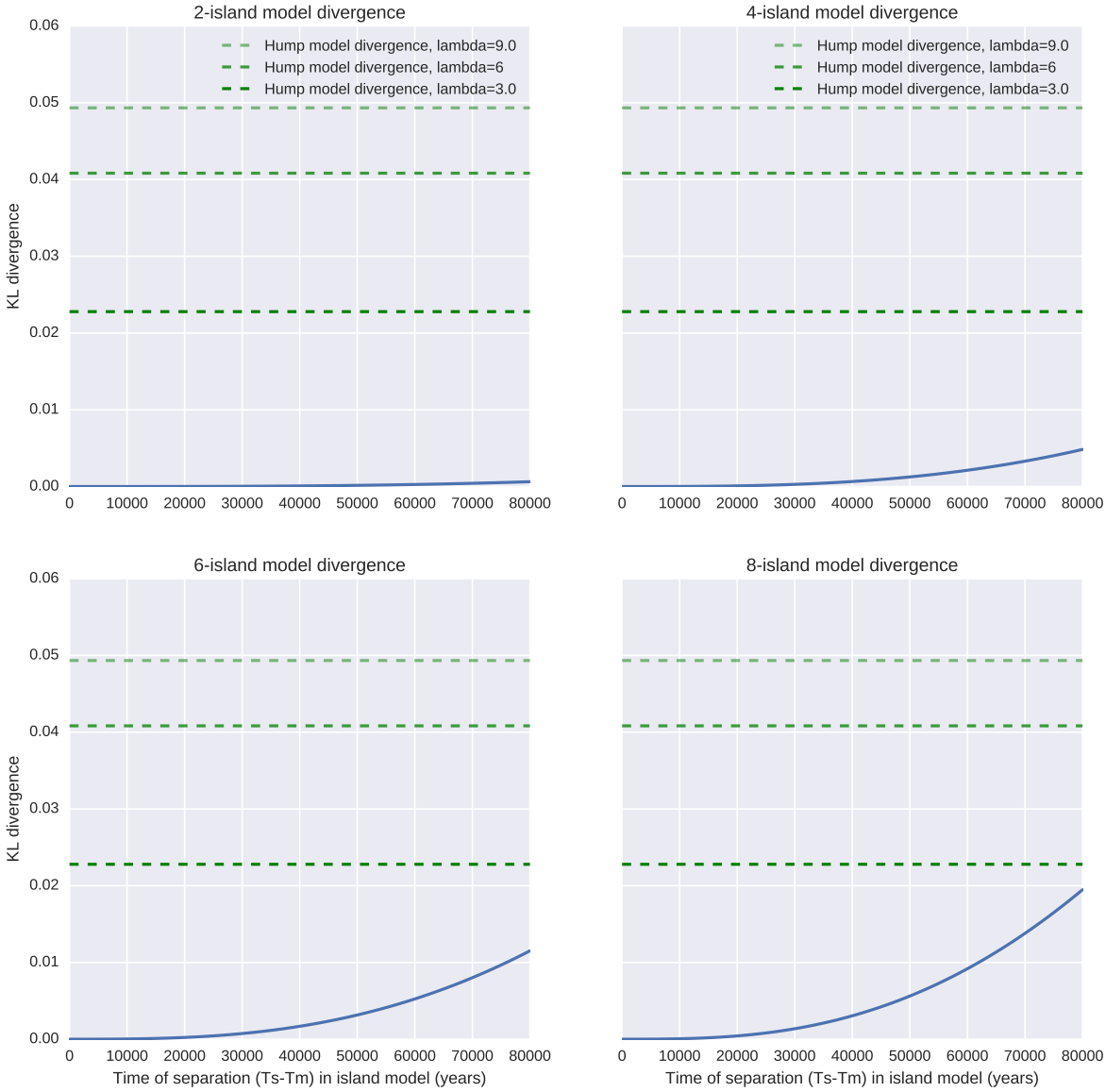


Fig. 4.14 Island models I_N consistent with central chimpanzee population expansion when baseline $N_e = 10000$. Figures represent dependence of KL divergence values on length of island intervals centered at $t_p = 40$ kya. Intersections with dashed green lines represent matching KL values with hump models in which increase lasts from 20-60 kya. Scale of the population increases in the hump models H_λ vary according to λ values associates with dashed lines (see legend).

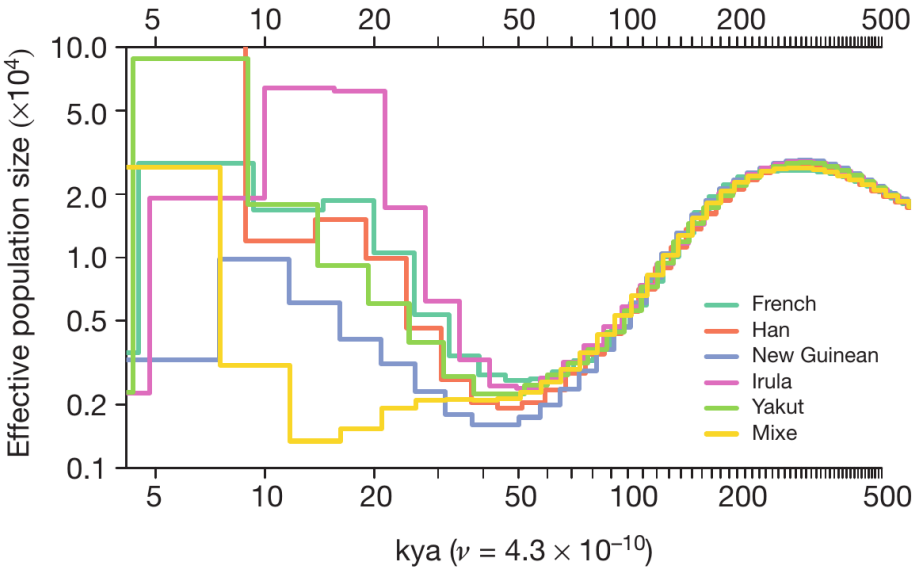


Fig. 4.15 PSMC-estimated historical effective population sizes of out-of-Africa populations. Produced by the Simons Genome Diversity Project [81]. We focus on the increase in population size occuring between approximately 50 - 300 kya.

As an example of one reading of these figures, if we consider the baseline effective population size to be 5000 (Figure 4.17) and we believe that the most accurate scaling of the population hump (contraction in historical terms) is $\lambda = 4$, then we can see that the a 4-island period centred around 300kya would need to last for approximately 230-240 ky in order to depress coalescent rates sufficiently, and an 8-island period, on the other hand, should last for at least 150 ky. With only 2-islands you would need a division lasting longer than 300 ky. As these models assume zero migration, they set the minimum number of islands required for their respective lengths of time. In other words, if some external evidence suggested it was only plausible that structure could last for 150 ky, then we would need more than 8 islands if a significant amount of migration is expected. Observe the small differences between the 6-island and 8-island models. These suggest that at these time scales we will not be able to distinguish models with more than these number of islands, due to the effect observed in Figure 4.3.

If structure existed on this time scale, we expect it would indicate divisions between ancestral populations on the African continent. However, this period is likely to pre-date the appearance of the earliest extant population structure in humans (between ancestors of Khoisan populations and other Africans) [150]. Its plausibility is thus difficult to assess using modern sequence data alone. External evidence, in the form of the fossil record, suggests the out-of-Africa migration coincided with the contraction [94], and as mentioned in the introduction to the chapter it also coincides with the decrease in cross-coalescence between populations in and out of Africa. This is strong prior evidence that the migration is causally

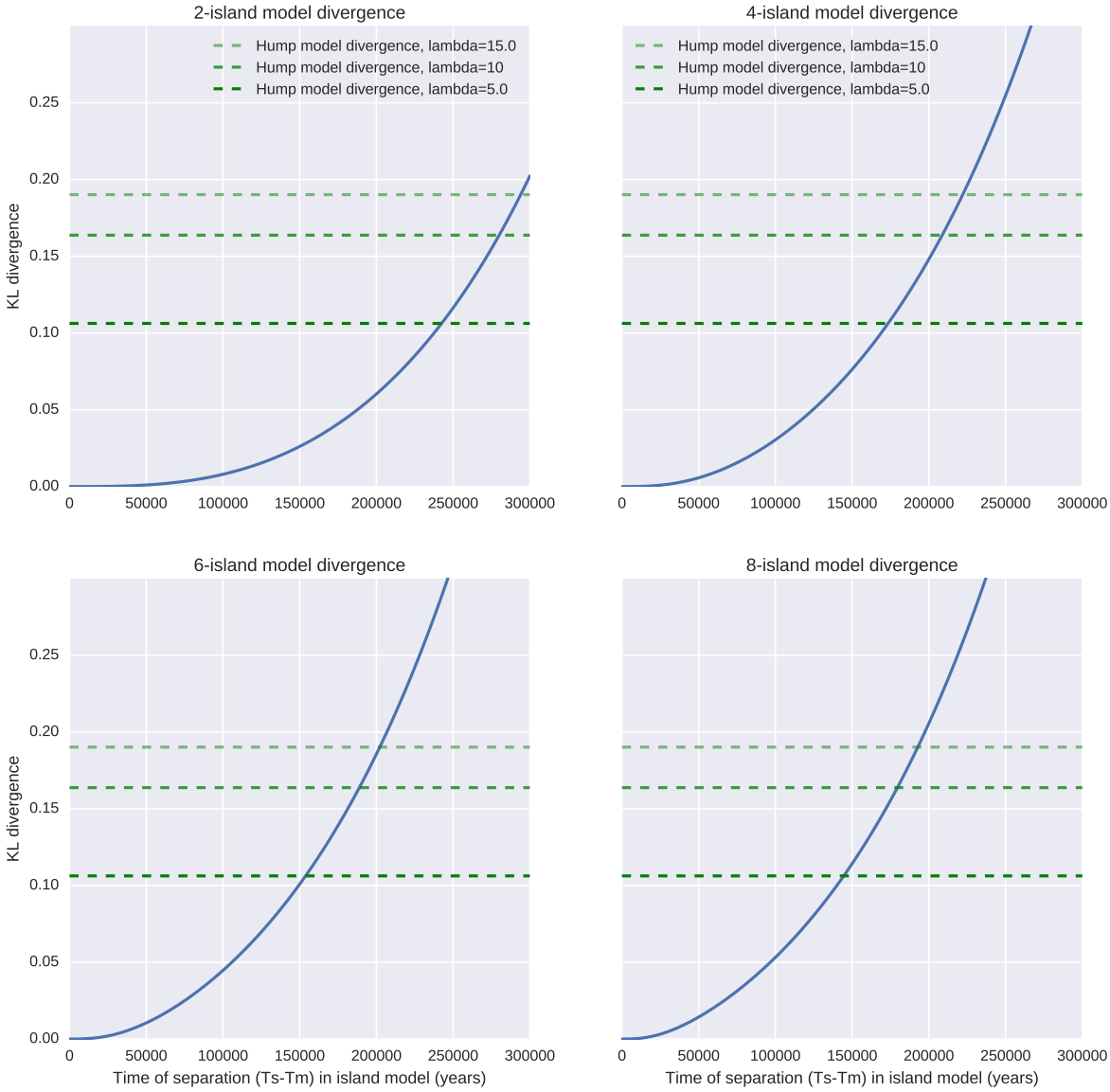


Fig. 4.16 Island models I_N consistent with OOA population contraction when baseline $N_e = 2000$. Figures represent dependence of KL divergence values on length of island intervals centered at $t_p = 300$ kya. Intersections with dashed green lines represent matching KL values with hump models in which increase lasts from 200-400kya. Scale of the population increases in the hump models vary by dashed line.

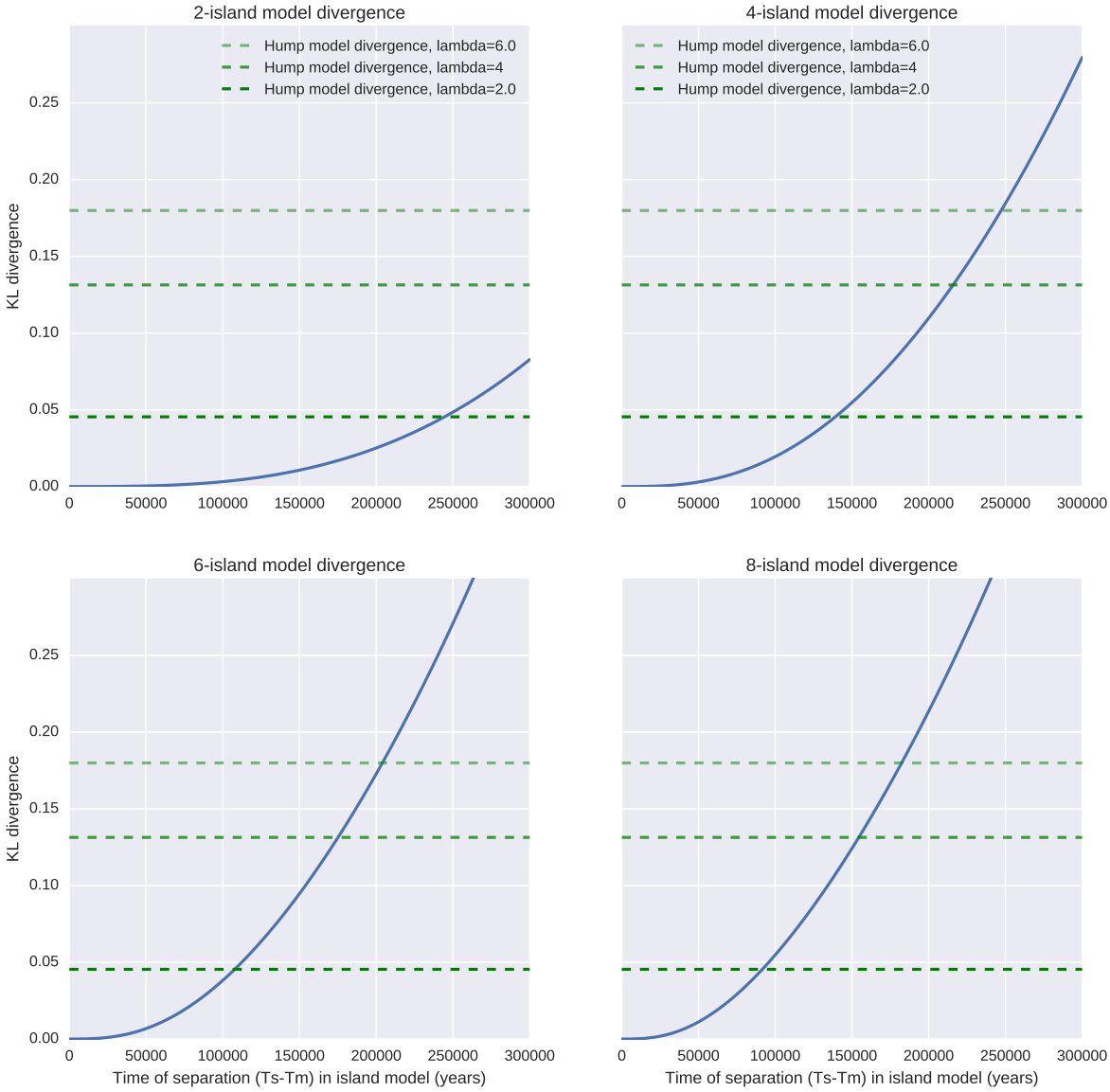


Fig. 4.17 Island models I_N consistent with OOA population contraction when baseline $N_e = 5000$. Figures represent dependence of KL divergence values, on length of island intervals centered at $t_p = 300$ kya. Intersections with dashed green lines represent matching KL values with hump models in which increase lasts from 200-400kya. Scale of the population increases in the hump models vary by dashed line.

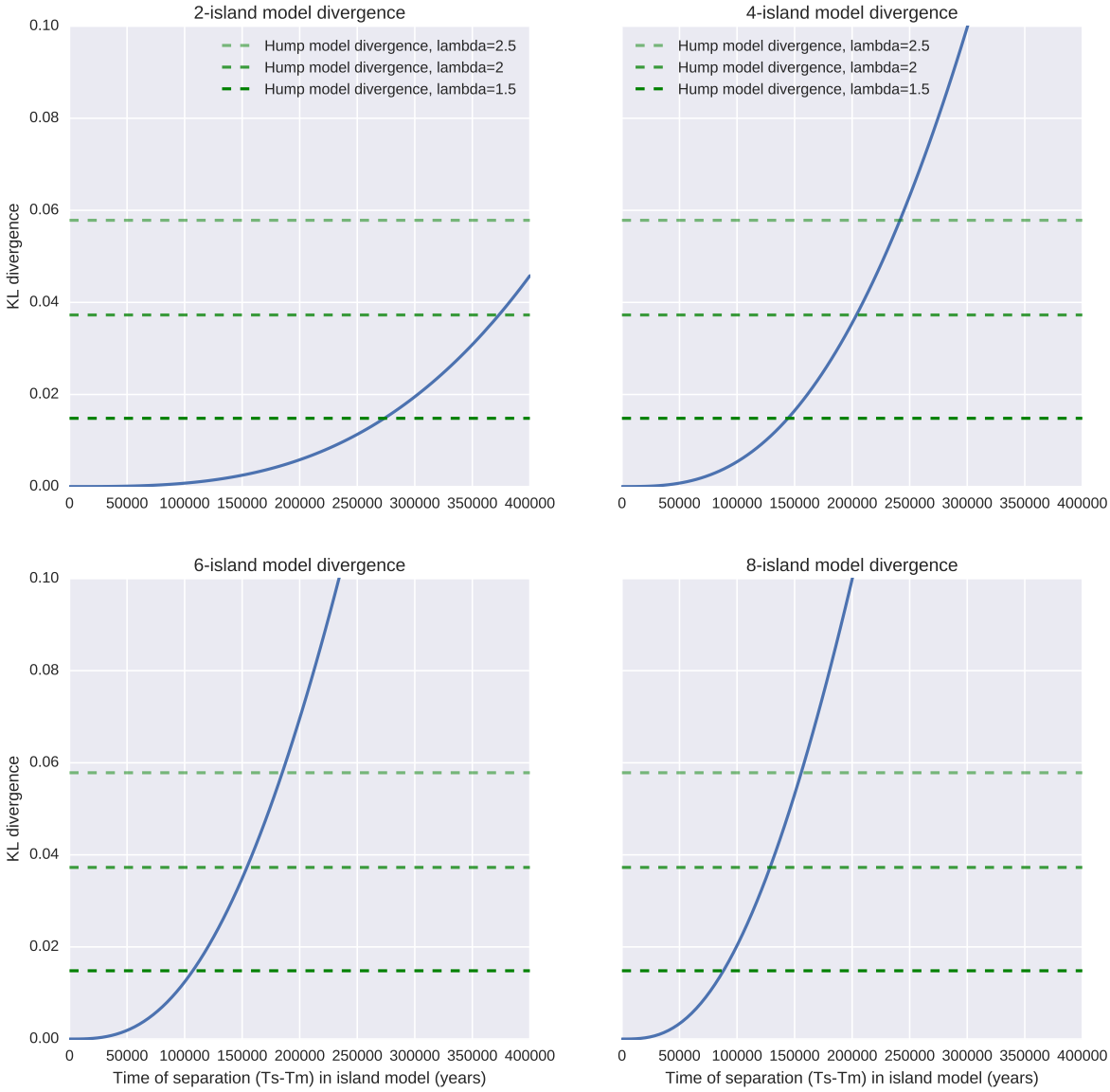


Fig. 4.18 Island models I_N consistent with OOA population contraction when baseline $N_e = 10000$. Figures represent dependence of KL divergence values on length of island intervals centered at $t_p = 300$ kya. Intersections with dashed green lines represent matching KL values with hump models in which increase lasts from 200-400kya. Scale of the population increases in the hump models vary by dashed line.

related to the decline in effective population size. The number of strictly isolated islands required within Africa to produce a similar increase in population size, under the simple island model proposed here, seems large enough to support the view that the contraction is due to a genuine change in census population.

4.4 Additional effects of historical substructure

Subpopulation lineage sorting Under the demographic scenario briefly analysed in the beginning of this chapter, Li and Durbin observe an inflation of the effective population size due to the 2-island structure. The island models I_N which we developed further also exhibit this behaviour. However the effect depends on an assumption about the sorting of lineages into islands at the time of population splitting (T_S in our model): it assumes uniform random island choice by uncoalesced lineages.

This is significant, because where lineages are allowed to choose islands non-uniformly, and depending on the length of time of the separation, coalescent rates can increase. Under the I_N models used in this chapter, it is straightforward to determine the conditions under which this will occur. Focusing only on the conditional case in which lineages are uncoalesced at time T_S, we know that the probability of coalescence by the end of the island period is given by $g_{\text{isl}}(\tau) = p_s(1 - e^{-N\tau})$, where p_s is the probability of sorting into the same island and N is the number of islands, here assumed to be the same size. For the sake of brevity we set $T_M - T_S = \tau$, and note that in the case of uniform random sorting $p_s = 1/N$. The probability of coalescence by T_M in the comparable panmictic model is $g_{\text{pan}}(\tau) = 1 - e^{-\tau}$.

We know from before that under uniform random sorting, $g_{\text{isl}}(\tau) < g_{\text{pan}}(\tau)$ for any $\tau > 0$. This is true since both expressions are 0 when $\tau = 0$, while the first order derivatives $g'_{\text{isl}}(\tau) \leq g'_{\text{pan}}(\tau)$ when $\tau \geq 0$. However, this is not the case under all sorting regimes. More generally,

$$g'_{\text{isl}}(\tau) = p_s N e^{-N\tau} \quad \text{and} \quad g'_{\text{pan}}(\tau) = e^{-\tau}.$$

Thus $g'_{\text{isl}}(0) > g'_{\text{pan}}(0)$, and coalescent probability is initially greater in the structured model, when

$$p_s > 1/N.$$

Probability for coalescence in this case is identical at the separation time τ^*, when $g_{\text{isl}}(\tau^*) = g_{\text{pan}}(\tau^*)$. This phenomenon, in which structure decreases effective population size, is illustrated in Figure 4.19.

This situation is biologically plausible. It can occur if some population C is formed through the admixture of equally-sized populations A and B but admixture proportions are not equal. This can happen, for example, if C is formed out of founding populations of different sizes which merge soon after breaking away from A and B, or if one of A and

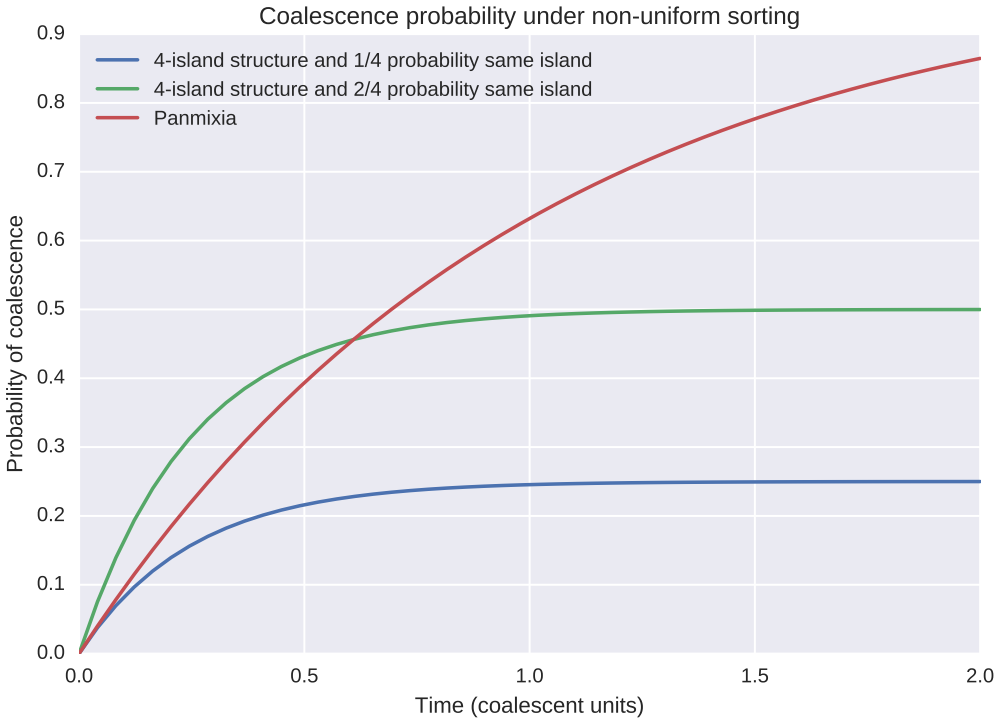


Fig. 4.19 Probability of coalescence under non-uniform lineage sorting in an I_4 model by time t. This illustrates the probability that two uncoalesced lineages at time 0 coalesce before time τ. We assume that island structure exists from the start. Functions plotted in non-panmictic cases are the probabilities of sorting into the same island multiplied by probabilities of coalescing under condition that sorting is into the same island. In the non-uniform sorting case, when probability of lineages sorting into same island is 1/2, as opposed to 1/4, there is a time interval in which coalescence probability is higher than in the panmictic case, and N_e lower, unlike the case of uniform sorting. In other words, if structured period ends before intersection of curves, the direction of the effect on coalescence rates will be different to the uniform random sorting case.

B change size near in time to the merging, due to disease, or conflict over resources in the process of merging. If external evidence (as might be provided by a rich ancient DNA record) suggests this is likely in the case of the history of some specific population, the island model developed earlier in this chapter can readily be generalised.

Chapter 5

Haplotype-Based Analyses of Ancient Population Structure

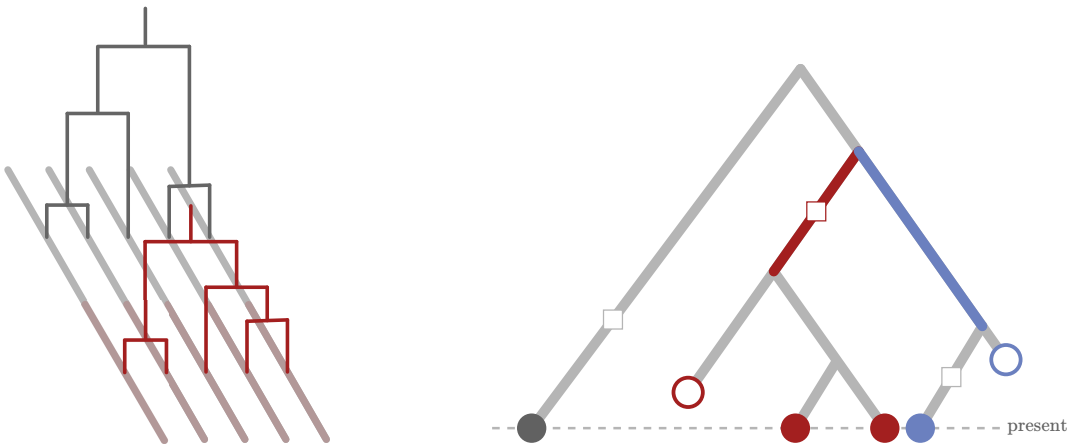
We saw in Chapter 4 that it can be difficult to detect and characterise past population structure using modern sequence data alone. With the development and proliferation of ancient DNA sequencing we are able to observe genetic variation at specific times in the past. This variation can be compared with other ancient samples to learn about past gene flow, or with modern samples to reveal the histories of extant populations. Some of the methods commonly used to draw these inferences have not expressly been designed to analyse mixed sets of modern and ancient sequences. While they can still be informative, we do not always understand the extent to which this assumption affects inferences. However, new approaches are increasingly being developed with such datasets in mind. In this chapter I present a tentative addition to this family of techniques.

5.1 Method

5.1.1 Modelling ancient structure with the ARG

The approach in this study requires a sample consisting of two ancient sequences and a set of reliably-phased modern sequences. We speak hypothetically of two different populations from which the ancient sequences are drawn, but the relationship between these populations is left open. Indeed, the populations may be identical. We begin by identifying haplotype segments in the modern sample which share a long common ancestry. We then compare the derived mutations on these haplotype segments with those shared privately with either of the ancient sequences and use this to "match" the loci with either of the ancient individuals. This quantifies shared ancestry in a way particularly amenable to demographic analysis.

We use this pattern of shared inheritance to inform us of ancient population structure. The underlying motivation is that a modern panmictic population cannot trace significantly



(a) Marginal ancestral trees **(b)** Detectable branch-sharing

Fig. 5.1 The ancestral recombination graph (ARG) Figure (a) shows two marginal genealogical trees of adjacent loci, obtained from the ancestral recombination graph of a sample of 5 sequences. Figure (b) shows the unobserved branches which the matching procedure is attempting to infer. The tree represents the history of a single locus. It describes the times at which the sample ancestral lineages find their common ancestors and the order in which these ancestors are found. Each solid circle represents a segment of a modern sequence, while the hollow circles represent corresponding regions from ancient sequences. The red line is the shared "private" history of all the red-coded segments; similarly with blue. The only history the black locus shares with any other is the history it shares with all of them. These private branches form the basis of the matching process. Since we cannot observe the trees directly, they are inferred from the sharing of private derived alleles, ie. the shared history of the red-coded segments is detectable only if a mutation occurres on the red line.

more ancestry to one or other ancient population unless those populations have significantly distinct histories. Nonetheless the extent, and possibly direction, of the difference will depend on demography and the sampling time of the ancient sequences. Thus in the second phase we use explicit demographic models to interpret the results of the matching process. The first phase is agnostic with respect to models chosen in the second and we foresee several possible extensions of our broad approach.

Segmentation

As with the ChromoPainter approach outlined in Chapter 2, the haplotype segments we require are those which correspond to single ancestral trees. We use these trees because they are well-modelled by coalescent processes [39]. However, the trees at adjacent such intervals are not independent since they are identical up to the point at which some past recombination event separates them. Thus closely positioned trees will be correlated. The set of such trees across an entire sample of sequences, along with their complex linkages, can be obtained from the ancestral recombination graph (ARG) [40], which is what we infer in the first phase of the method (see Figure 5.1 (a)). We will be working with the minimal set

of haplotype segments for a given ARG and reserve the word "locus" or "non-recombining locus" for these intervals.

Due to its large parameter space, the ARG is a notoriously difficult structure to infer. ARGweaver, the method we use, implements a Markov Chain Monte Carlo (MCMC) approach to draw samples from a probability distribution over the space of all ARGs, estimated under the assumptions of the sequentially Markovian coalescent (SMC) [130]. Unlike methods which might produce a maximum likelihood or max posteriori ARG, this stochastic sampling introduces uncertainty in our inferences. However, we do not use the entire data structure. The ARG samples drawn by ARGweaver are only used to identify the endpoints of non-recombining loci. While the differences in positions that do exist can affect downstream summary statistics, our results show consistency between samples drawn (see analysis below). We also average over randomly choosen samples (see Modelling and Inference section).

Matching

For a given ARG sample, each haplotype produces a potentially distinct set of loci drawn from the modern sequences. For each locus, we apply a matching procedure which potentially pairs each of its segments with either of the ancient individuals. Note that this matching requires phased ancient sequences.

1. Identify the derived alleles on the locus which are shared with one of the ancient individuals, but not the other. We call these the *private shared derived alleles*.

2. If the locus shares such alleles with only one of the ancient sequences, declare it a match with that one.

3. If it shares them with *both* or *none* of the ancient sequences, reject the segment as ambiguous.

For the sake of clarity, let $m = m_1 m_2 \dots m_b$ be a modern haplotype segment, with $m_i \in \{0, 1\}$, $i \in \{1, 2, ..., b\}$ and b the number of base pairs in the locus. Let the allelic type be 1 if it is derived and 0 if ancestral. Similarly, let $x = x_1 x_2 \dots x_b$ and $y = y_1 y_2 \dots y_b$ be the corresponding ancient loci, with $x_i = 1$ or $y_i = 1$ whenever x_i or y_i are heterozygous *or* homozygous for the derived allele. Otherwise, $x_i = 0$ and $y_i = 0$. With this, we can succinctly summarise the matching process:

Haplotype matching process Locus m matches ancient sequence x if $m_i = x_i = 1$ for at least one $i \in \{1, 2, ..., b\}$ and if for all i, $m_i = y_i$ only when $m_i = 0$. Similarly with y. If m matches neither x nor y we declare it ambiguous.

The justification for this process is as follows. If for some locus we had access to the true genealogy of our entire sample (including the ancient individuals) we could directly observe

the shared branch lengths between haplotypes. The distribution of these values over all the loci in a sequence can be used to draw inferences about population history (see Appendix). Without this information, we have to infer the existence of these branches by comparing the private derived alleles that individuals share. These correspond to mutations on the shared branches. The requirement that only one of the ancient sequences shares the mutation ensures that it occured more recently than the common ancestor of both ancient sequences (see Figure 5.1 (b)). This follows from the assumptions of the infinite-sites mutation model, in which recurrent mutations and back-mutations are assumed not to occur.

If both modern and ancient sequences are reliably phased and methods existed to infer ARGs with sufficiently good time resolution, and which allowed sequences to have different sampling times, it would be possible to draw inferences from branch-lengths taken directly from the whole-sample ARG. Since this is not yet possible, we infer the existence of these branches by the presence of derived alleles shared exclusively by branch descendents. Note that not all private branches will be detected. We model detectability in the next section. Also observe that segment ambiguity caused by the sharing of private alleles from both ancient sequences may have several causes, namely, a failure of the infinite-sites assumptions to model the mutational process correctly, an incorrect inferrence of the locus endpoints, poor variant-calling, or switch-errors resulting from poor phasing.

Comparison with the D statistic There is evident similarity between this procedure and the popular "ABBA-BABA", or D-statistic, test, first used by Green et al (2010) [38]. Indeed, our matching can be seen as a repurposed and haplotype-adjusted version of it. The D is a four-population test for admixture. Given genotypes from base populations X and Y, a third population Z, and an outgroup population O, the test counts the differences between the number of a certain class of variants shared by Y and Z and compares it with that shared by X and Z. The relevant variants are those which occur in the base population and Z, but neither the other population nor O. The outgroup variant is a proxy for the ancestral type which we use in this study. The difference is usually summarised with the statistic

$$D = \frac{C_{ABBA} - C_{BABA}}{C_{ABBA} + C_{BABA}}, \tag{5.1}$$

where C_{ABBA} and C_{BABA} represent the counts of the respective shared variants across the entire genotype (B represents the derived and A the ancestral alleles in the order (X, Y, Z, O)). A significant difference from 0 is taken as evidence of historical admixture from population Z into Y if positive, and into X if negative. Significance is usually tested by block jackknife procedure since SNPs in LD will have complicated correlations. This test was used by Green et al to infer an excess of Neanderthal alleles in the descendents of ancient Eurasians relative to those found in Africans, which they took as evidence of ancient admixture between Neanderthals and the humans who migrated out of Africa.

Using this notation, our focus is on the population Z, which we assume is modern. We are attempting to understand the shared history of Z with X and Y, which we take to be our hypothetical ancient populations. We have no explicit outgroup, but infer the derived alleles using previous determinations of ancestral variants by the 1000 Genomes Project [19]. The more important difference to our approach is that we attempt to prevent the overcounting of certain trees by regarding segments as single matches even if they contain multiple shared private derived alleles. Overcounting can bias inferences if gene flow between populations occurred at different times and recombination had different time-windows during which to break up the clustering of alleles.

The D is known to have certain robustness guarantees which our measure is not designed to have, notably to variation in ancestral effective population size [29]. Durand et al. recommend not using the D for demographic inference because of its complex sensitivity to various historical parameters. This difficulty is shared by our approach, although we attempt under simplified models to gain some understanding of these interactions. Under these conditions, and setting several parameters using external information, it is possible to do a limited inference of some aspects of the history.

Modelling and inference

The marginal genealogical trees obtained from the ARG, those corresponding to single loci, can be modelled by coalescent processes. Under simple historical demographic models, we can use this fact to analytically compute the probabilities of observing matches with either ancient sequence. This involves first determining the distribution of shared branch lengths between modern and ancient sequences. Conditioned on this distribution we can determine the probability that any given tree has a *detectable* private branch shared with either of the ancient sequences. This is equivalent to asking whether a mutation occurs on the branch.

Our structured demographic model is as follows. We assume that the haplotype segment A_0 at a given locus is sampled at present. It can follow one of three ancestral paths. The first path, which it follows with probability p_1, allows it to migrate to the population from which the ancient sequence A_1 is drawn, and it has the potential to coalesce with the corresponding locus of A_1 from "join time" j_1 until some "merge time" t_M, when the populations of A_1 and A_2 merge. Similarly, with probability p_2 it is allowed to coalesce with the corresponding locus of A_2 after j_2 and until t_M. With probability p_0, on the other hand, the ancestry of the locus is traced through some "ghost" population and can coalesce with neither A_1 nor A_2 until time t_M. We assume that each subpopulation is panmictic and that selection is not strong or common enough to systematically bias the probabilities of coalescence or mutation. We also allow the subpopulations of A_1 and A_2 to have effective population size scaled relative to the (constant) effective population size N of the global ancestral population, by parameters $1/\lambda_1$ and $1/\lambda_2$, respectively. No migration is permitted between these ancient subpopulations, so

that A_1 and A_2 cannot themselves coalesce until the global merging of populations at t_M. Note that we also only compare sites at which we successfully call alleles from both ancient sequences, and thus assume that differences in variant calling and missing ancient sites will not introduce bias.

We seek the conditional matching probability

$$P_1 = \frac{\mathbb{P}(S_1)}{\mathbb{P}(S_1) + \mathbb{P}(S_2)}, \tag{5.2}$$

where S_1 and S_2 are the probabilities that a segment matches with ancient sequence A_1 or A_2 respectively. We outline the argument in the case of $\mathbb{P}(S_1)$, though relegate the detailed arguments to the Appendix. Observe that there are three possible topologies for the gene tree of this 3-sample locus. We are interested in the topologies under which the modern locus A_0 and the first ancient sequence A_1 share a private branch. This occurs in precisely one case: when the ancestral lineages of A_0 and A_1 are the first to coalesce. We shall label the length of their private branch L_{01}, and the shared branch lengths under the other possible topologies L_{02} and L_{12}, following the obvious notation. The detectability of this branch, under the infinite sites assumption, is the probability that a mutation occurs on it. We shall refer to the random number of mutations on branch L_{01} as M_{01}. Observe that the probability of a segment match occuring is thus simply $1 - \mathbb{P}(M_{01} = 0)$. We determine this by conditioning on the distribution of L_{01}. Thus

$$\mathbb{P}(M_{01} = 0) = \mathbb{P}(L_{01} = 0) + \int_0^\infty \mathbb{P}(M_{01} = 0 | L_{01} = l)\mathbb{P}(L_{01} = l)\,\mathrm{d}l. \tag{5.3}$$

The first term is the probability that A_0 and A_1 only share branches which are also shared by A_2 (and thus share no private recent mutations). The second term is the probability that no mutations are observed conditioned on L_{01} being nonzero. (With a slight abuse of notation, we assume in the second term that $l > 0$, in order to avoid including the point mass at $l = 0$. We also suppress the dependence on demographic parameters.) The distribution of L_{01} can be determined by conditioning on the ancestral path that A_0 follows, and on the event that it coalesces with A_1 before t_M. This argument is shown in the Appendix. For the sake of testing and subsequent expressions, the cumulative distribution function of shared branch lengths is given by

$$F_{01}^3(l) = \begin{cases} 0 & \text{if} \quad l < 0 \\ p_2(1 - e^{-t_2/\lambda_2}) + (2/3)B & \text{if} \quad l = 0 \\ F_{01}^3(0) + A_1\left(\lambda_1 e^{l/\lambda_1} + e^{-l} - \lambda_1 - 1\right) + (1/3)\left(1 - e^{-l}\right)B & \text{if} \quad 0 < l < t_1 \\ F_{01}^3(t_1) + \left(A_1\left(e^{(1+1/\lambda_1)t_1} - 1\right) + (1/3)B\right)\left(e^{-t_1} - e^{-l}\right) & \text{if} \quad t_1 < l. \end{cases} \tag{5.4}$$

The superscript of F_{01}^3 signifies the third of the models analysed in the Appendix and the time t_1 is window before t_M during which A_1 and the modern locus can coalesce, so that $t_1 = t_M - j_1$. Also,

$$\begin{aligned} A_1 &= \frac{p_1 e^{-t_1/\lambda_1}}{\lambda_1 + 1} \quad \text{and} \\ B &= p_0 + p_1 e^{-t_1/\lambda_1} + p_2 e^{-t_2/\lambda_2}. \end{aligned} \tag{5.5}$$

Similar expressions exist for F_{02}^3.

Given a Poisson model for the distribution of mutations on branches as a function of branch lengths, as well as μ, the per-site per-generation mutation rate, and b, the length of the locus in base pairs, we solve Equation 5.3 to obtain

$$1 - \mathbb{P}(M_{01} = 0) = F_{01}^3(0) + I_1 + I_2 + I_3$$

$$\begin{aligned} \text{where} \quad I_1 &= A_1 \left[\frac{e^{(1/\lambda_1 - \theta/2)t_1} - 1}{1/\lambda_1 - \theta/2} + \frac{e^{-(1+\theta/2)t_1} - 1}{1 + \theta/2} \right], \\ I_2 &= A_1 \left(e^{(1+1/\lambda_1)t_1} - 1 \right) \left(\frac{e^{-(1+\theta/2)t_1}}{1 + \theta/2} \right), \\ \text{and} \quad I_3 &= \frac{B}{3(1 + \theta/2)}. \end{aligned} \tag{5.6}$$

Here, $\theta = 4N\mu b$ the scaled classical parameter governing the rate of mutations in coalescent processes.

Ultimately, this allows us to obtain the value of P_1, with which we might fit a binomial model to the observed proportion of *non-ambiguous* segments which match the first ancient sequence. However $P_1 = P_1(\Theta)$ where Θ is the set of demographic parameters under which the loci have evolved. We should not expect that by fitting P_1 to the observations the values of the individual parameters will be identifiable. There are two further difficulties with doing this directly. The first is that in the inference, we need to account for the uncertainty induced by the MCMC sampling procedure. For each haplotype we randomly choose an ARG from the sampled graphs, and compute the relevant quantities using these random choices. We then redo this multiple times and average over the values obtained through each set of random samples. As shown in Results, the average estimate of several required statistics converges after 30-40 iterations.

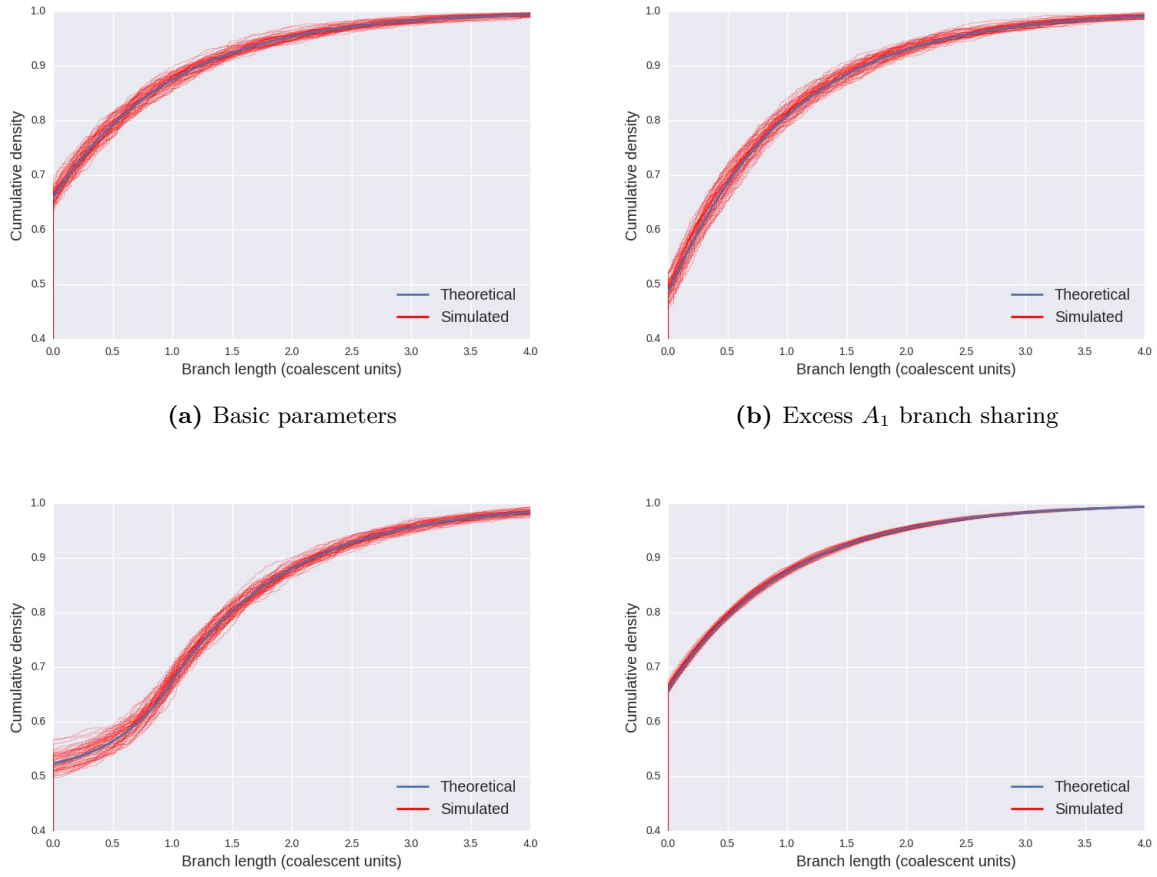
5.2 Results

Validity of analytical results

Several features of the analytical arguments above are validated here using simulations, all of which are run on the coalescent simulator `msprime`. We first check that the theoretical prediction for the distribution of shared branch lengths is correct. While this random quantity is not observable, `msprime` allows us to generate coalescent trees under a large variety of ancient demographic models, and it supplies the topologies of these trees along with all relevant branch lengths. We also run these simulations in order to obtain some insight into the distributions of simulated trees about their theoretically predicted values. In the first set of simulations we generate independent coalescent trees under the structured demographic model. Later, we illustrate the effects of linkage through simulations of sequences.

The first set of simulations validate the theoretically-derived cumulative distribution functions (cdf) of the random shared branch length L_{01}. This is the shared coalescent history of the modern haplotype segment and the first ancient segment. In Figure 5.2, each subplot shows the theoretical cumulative distribution 5.14 and the observed cumulative distributions of 1000 trees per simulation (there are 50 simulations in each subplot). Parameters specifying the demographic histories are stated in the caption. The plots confirm several salient features of the theoretical cdf. First is the proportion of trees in which no private branch-sharing occurs between the relevant haplotype. This is given by the intersection of the curves with the y-axis. Second is the shape of the cdf, which the simulated curves closely approximate even when the demographic conditions are chosen to exaggerate greatly the branch-sharing of the two haplotypes and when the merge time is increased in order to alter the convexity of the theoretical cdf. If for each simulation, we increase the number of trees by a factor of 10, we see that the simulations adhere more closely to their theoretical prediction, suggesting the asymptotic accuracy of the theoretical cdf. Observe as well that the simulations in each plot appear to be symmetrically distributed around the theoretical curve.

In the next set of simulations, shown in Figure 5.3, the demographic history is fixed and asymmetric (the parameter values stated in the figure captions). We validate the use of the theoretically derived quantities P_1 and P_2 as estimators for the relative proportion of haplotype matches under this demographic model. Each simulation consists of 10000 coalescent trees. For each of these we have counted up and plotted the number of modern segments matching (unambiguously) with either ancient haplotype. In addition, we plot the point corresponding to $(10000 \times P_1, 10000 \times P_2)$. Observe that the parameters have been chosen to illustrate the ambiguity of the relative number of segment matches. In this case, a separation time of 200 generations between the two ancient sequences is not sufficient to guarantee that more matches with the more recent ancient sequence will be made in each simulation (illustrated by the number of blue points above the dashed green line), even



(a) Basic parameters

(b) Excess A_1 branch sharing

(c) Merge time extended

(d) Increased number of trees per run

Fig. 5.2 Comparison of theoretical and simulated L_{01} branch length distributions. (a) Cumulative distribution of L_{01}, the random branch length shared between modern locus and ancient sequence A_1, under the following demographic parameters: $t_M = 1000$ generations, $j_1 = j_2 = 500$ generations, $N = 1e4$, $\lambda_1 = \lambda_2 = 2$, and $p_1 = p_2 = 0.5$. The blue line represents theoretical prediction of branch lengths. Red lines correspond to cumulative histograms of 50 simulations, each consisting of 1000 trees generated with `msprime`. Figure supports demographic analysis under this set of parameters, showing close agreement of simulations with the theoretical prediction as well as an approximately symmetric distribution about the theoretically-derived curve. (b) Comparison under demographic conditions chosen to exaggerate the branch-sharing with ancient sequence A_1: $p_1 = 0.7$, $p_2 = 0.3$, $j_1 = 0$ generations, $\lambda_1 = 10$ (remaining parameters as in previous plot). Support for the theoretical curve is maintained as branches shared become longer on average. (c) Comparison in which all parameters are the same as in (a), except the merge time $t_M = 20000$ generations. Agreement with theoretical curve maintained as overall shape of distribution curve changes. (d) Number of trees in each run increased to 10000, showing the tightening of the simulated distributions about the theoretical curve.

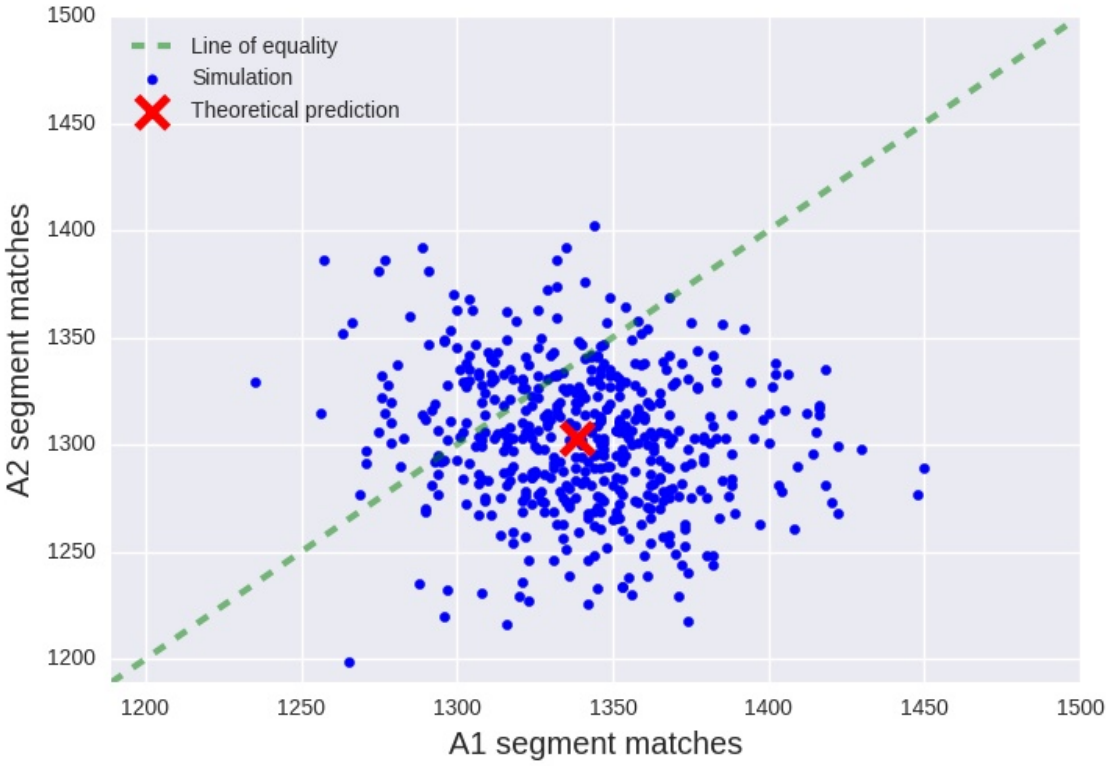


Fig. 5.3 Example of relative segment matching under asymmetric demographic history Each of the 500 blue dots corresponds to a simulation under demographic parameters: $N = 1e4$, $j_1 = 800$ generations, $j_2 = 1000$ generations, $t_M = 2000$ generations, $p_1 = p_2 = 0.5$, $\lambda_1 = \lambda_2 = 2$, $\mu = 1.2 \times 10^{-8}$ per site per generation, $b = 2500$bp. The dots are plotted according to the number of segments which match with the respective ancient sequences in a simulation of 10000 trees (each) using `msprime`. The red cross corresponds to the independent theoretical prediction for each count of segments shared. For the sake of visual comparison, the green dashed line represents equality of segment sharing. The plot supports the theoretical predition of segment-sharing under this illustrative demographic model, and demonstrates the spread of matches that can be expected.

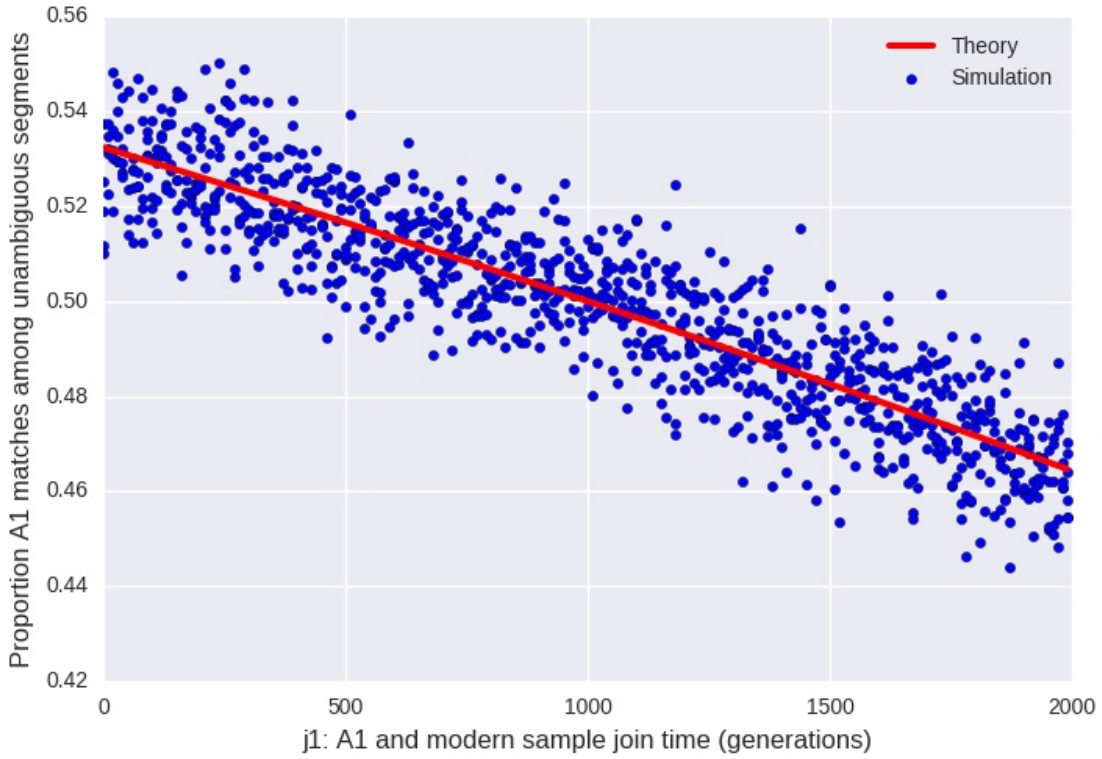


Fig. 5.4 Proportion of A_1 segment matching under varying join time j_1 Each blue dot corresponds to the proportion of segment matches (among unambiguous matches) with ancient sequence A_1 under simulations of 10000 trees. Join time j_1 varies from the present (0 generations) to the merge time $t_M = 2000$ generations. Other demographic parameters are $N = 1e4$, $j_2 = 1000$, $p_1 = p_2 = 0.5$, $\lambda_1 = \lambda_2 = 2$, $\mu = 1.2 \times 10^{-8}$, $b = 2500$ bps. At every 10 generations 5 simulations were run. The red line corresponds to the theoretical expectation of the proportion of matches. Observe that the theoretically predicted proportion of segment matching with ancient populations is equal when the join times are identical, at 1000 generations.

though this is the case for most simulations. The theoretical prediction for the relative number of segments matching is in this case below the dashed line, showing the degree of the expected bias in favour of the first ancient sequence. The roughly symmetric distribution of simulated points about the theoretical prediction suggests the multinomial distribution is an appropriate model in the case of unlinked trees.

In Figure 5.4 this observation is illustrated on a larger variety of conditions. Here we vary the bias in preferential matching of haplotype segments by varying one set of join times. We fix the time (at 1000 generations) at which the modern population merges with the second ancient population, and we vary the join time with the first ancient population from 0 to 2000 generations. At every tenth generation we run 5 simulations (note that the population size parameters are the same for each ancient population). We plot the ratio of

the segment matches with the first ancient sequence to the total number of segments which match unambiguously with either of the ancient sequences. In other words, the proportion of segment matches with the first haplotype, excluding the segments which match with neither (or, in practice, both). Along with these simulations, we also plot the theoretical prediction, which in this is case is merely P_1 as a function of the join time j_1. The plot illustrates that the trend in the simulation results follows the linear-like decrease in the theoretical curve. Observe that the theoretical prediction is 0.5 when the join times are identical. The scatter of the simulations about the theoretical curve also appear to be symmetric, supporting the appropriateness of using the P_1 quantity as an estimator of the empirical ratio of segment matches with one ancient haplotype to the total number of matching segments. It also suggests that on this narrow time scale it might be sufficient to use a linear approximation of our model.

In Figures 5.5 and 5.6, we show the interaction, under our model, of the effective population size, ancestry proportion and join time parameters. In both figures the scaled difference of probabilities, which is the theoretical counterpart of the chunk count statistic, can vary in magnitude and direction when demographic parameters are altered. The previous simulations suggest that the inferences we draw from this theorectical quantity allow us to assess the interpretation of the empirical chunk counts. As the effective population size of the one population is increased relative to the other, the probability of matching with that one decreases, although the effect is greater when the merge time of those populations is relatively later than the other. Differences in the parameters we call ancestry proportions have a similar effect, albeit in the opposite direction.

5.2.1 Application to the peopling of the Americas

In this section we look at substructure in an early American population. Evidence from modern and ancient DNA suggests that indigenous peoples in South and Central America trace almost all of their ancestry to a single founding population [133, 148, 127]. Modern North Americans are thought to draw a varying amount of their ancestry from this same population, with some Canadian groups receiving additional smaller contributions from ancestral Inuits while others, like the Athabascan-speaking Chipewyan, possibly receiving contributions from populations which left no modern descendents outside northern North America [149]. It has been claimed several times that the relative paucity of shared genetic drift between North and South Americans reflects an *early* separation in their shared ancestral population, possibly starting prior to migration into the Americas and reflecting different migratory waves, and perhaps different routes, into the continent [129, 148, 127]. Following Skoglund and Reich (2016), we refer to the more southern American ancestral lineage as SA and the northern Native American lineage as NNA [149]. Recent work, in which the analysis presented in this section was first published, supported this deep division, while offering

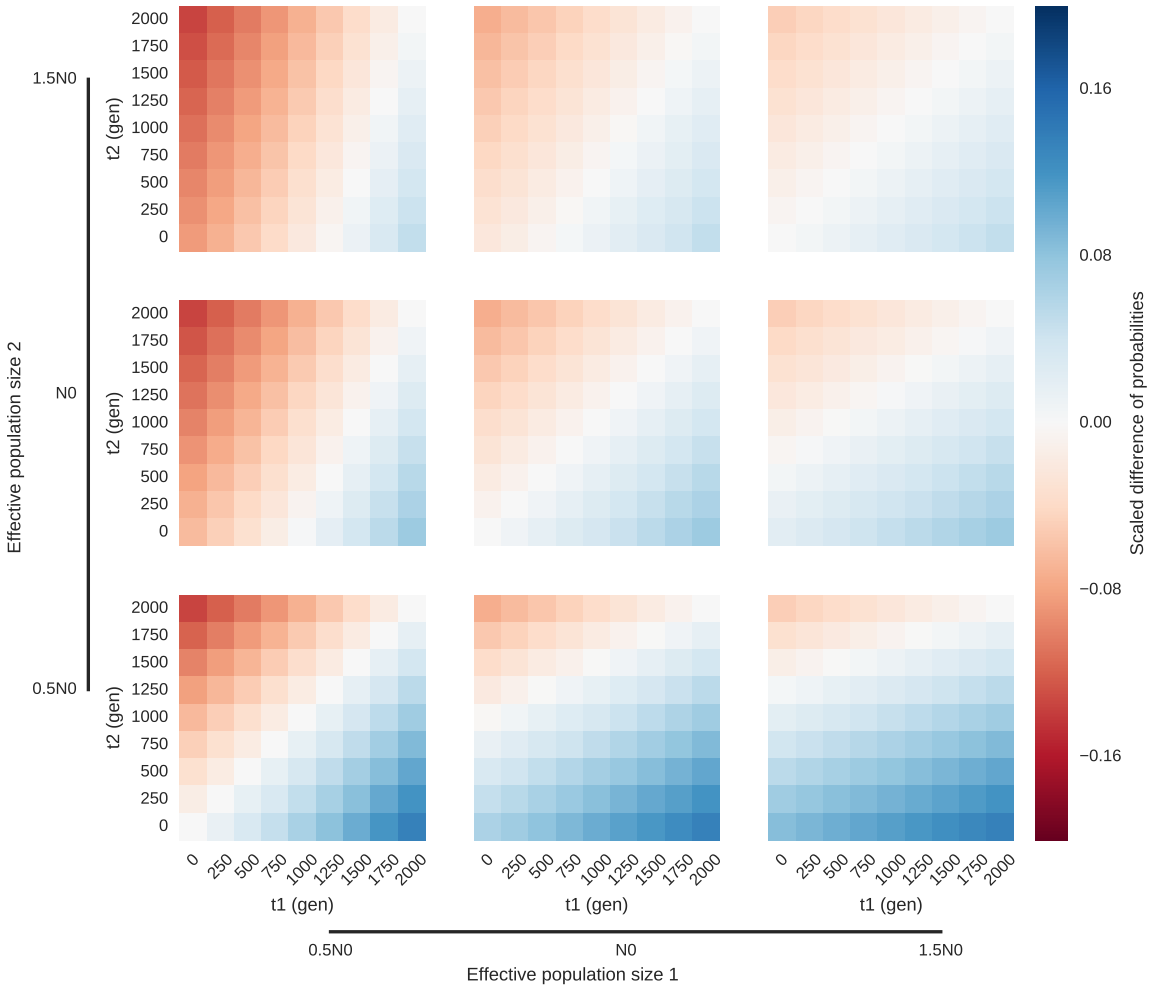


Fig. 5.5 Theoretical matching probabilities Heatmap illustration of matching probabilities under demographic regimes in which effective population size and join time vary.

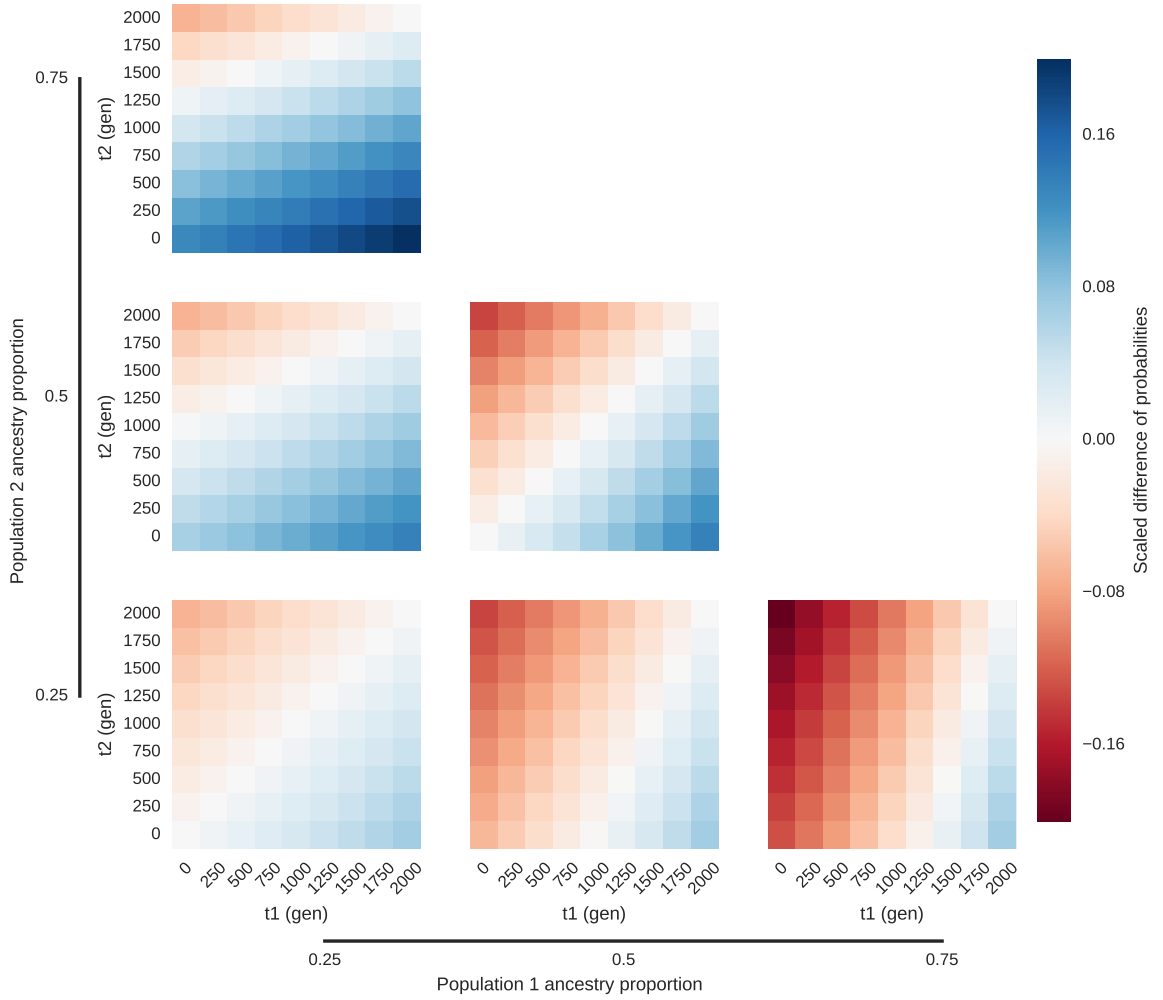


Fig. 5.6 Theoretical matching probabilities Heatmap illustration of matching probabilities under demographic regimes in which ancestry proportion and join time vary.

evidence for later gene flow between these subpopulations [143]. Using the approach set out in this chapter, I look into the question of the existence of substructure in the founding population, and the timing of any population subdivision.

Under the simplest model, the founding population broke away after ∼23kya from a population in Siberia which had exclusively East Asian ancestry [127]. However, a more complex origin for the population was suggested by the recovered genome of a ∼24ky-old individual from Mal'ta in Siberia, known as MA-1. This individual shared a large ancestry with most Native Americans (estimated at 14-38% using admixture graph fitting) and modern West Europeans (among whom there appears to be clinal variation in shared ancestry), but significantly less with East Asians [126]. This more complex origin might decrease the previously estimated upper bound on the separation time between the Siberian and ancestral North American populations [149]. It is not known when the population migrated into North America, though a growing body of archaeological evidence supports a possibly related human presence on the continent from at least ∼14.7kya, and perhaps from as early as ∼18kya [e.g. 27]. Environmental evidence suggests that the earliest path into America was a coastal route via the now submerged Beringian land bridge, since a viable ice-free corridor through modern-day Canada and Alaska appears only to have opened up after ∼12.6kya [116].

An additional ancestry component, from what has been called "Population Y", has been proposed to explain the fact that some Amazonian populations appear to share significantly more ancestry with Australo-Melanesians and Andaman Islanders than other Native Americans [148, 127]. This is thought to be the result of either an additional stream of migration from an Australasian-derived population, or of early substructure in the ancestors of the founding population, unrelated to the putative North-South divide. However, ancient DNA is yet to be recovered from this hypothesised population, and since as little as 2% additional Australasian ancestry is thought to be sufficient to explain this signal [148], we ignore this component during our analysis.

The oldest evidence of a separation comes from the genome sequence of an infant known as Anzick-1, associated with the Clovis cultural complex and dating from ∼12.6kya [129]. Anzick-1 was shown to have shared more drift with Native Americans than any other population in a worldwide survey, and to cluster with Native Americans in a global ADMIXTURE analysis and a PCA. This suggests that the infant belonged to the founding population and is evidence of long-standing population continuity on the continent. Similar analyses showed it to be significantly more closely related to SA-derived populations than NNA ones. On the basis of this, it has been argued that the Anzick individual did not come from a population basal to both SA and NNA, suggesting that population division might predate the establishment of the Clovis cultural complex [127].

An alternative hypothesis is that division between SA and NNA is more recent, or perhaps that there is a long-standing North-South cline, and NNA populations received additional gene

flow from outside the founding population. This might account for a significant proportion of the difference in the amount of drift that Anzick-1 shares with SA and NNA, and is suggested by the geographic proximity of Anzick-1 to northern groups, as well as the evidence of more recent contact between SA and NNA-derived populations [143]. Previous studies have discounted this hypothesis on the grounds that NNA-derived populations are no more closely related to any modern Siberian or East Asian population than are SA-derived populations [129]. However, it is uncertain whether an additional population, not extant and distantly related to modern Siberians, could have contributed to the NNA populations. Analysis produced to support the claim of early structure have also largely consisted of admixture graph fitting approaches based on SNP array data, and little of whole-genome methods, which introduces the possibility of systematic bias related to the choice of SNPs. In addition, Native American populations are known to have experienced significant founder efffects in the last 10-20kya [81], and f_3 statistics are harder to compare across such populations since bottlenecks can obscure drift-based signals of admixture [114].

Results Scheib et al (2018) present several additional ancient genomes from across Northern America, including one from an ancient Southwestern Ontario sample [143]. This 4000 year old sample, named CK-13, is regarded as coming from the NNA lineage since it clusters with northern samples in America-wide PCA and ADMIXTURE analyses. Using Anzick-1 and CK-13 I attempted to distinguish between these two scenarios: whether structure emerged early, predating Clovis complex, or whether it emerged later and the admixture graph results are more likely the result of additional gene flow into NNA. In order to apply the likelihood approach developed in Section 5.1, I use sequence data from the Pima and Surui, extant populations in Central and Southern America [81].

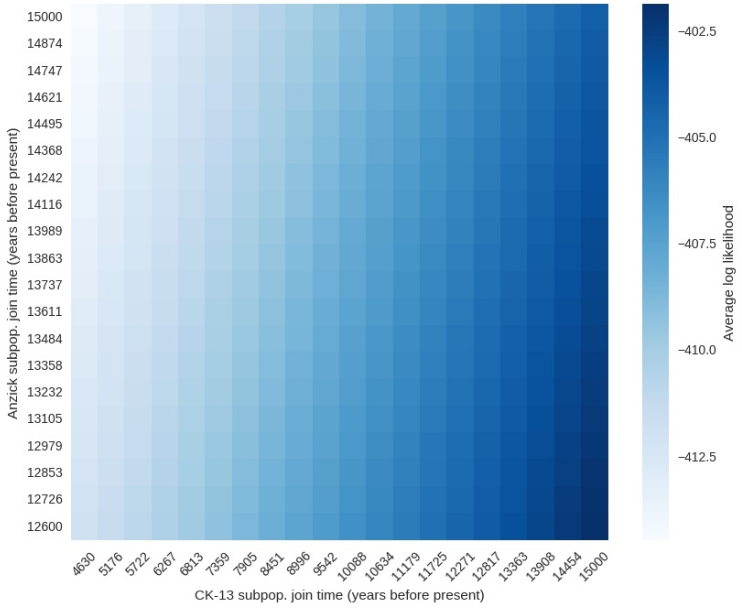
The log-likelihood heatmap shown in Figure 5.7 shows that under the assumptions of the population model and the basic parameters stipulated in the Analysis section above, it is most likely that the ancestral lineages of the Pima and the Surui begun to coalesce with Anzick-1 lineages around the age of sample, 13kya, but that they only started coalescing with the CK-13 lineages much earlier, possibly as early as a 15kya. The fact that coalescence with CK-13 does not start close to the age of the sample, is an indication of population structure predating the sample. One alternative explanation for a greater amount of Anzick-1 segment matching might be that the Clovis-associated SA population had a much lower effective population size than the NNA population. This would increase the rate at which coalescence occurs within that population and the related population matching probability. However, given the similar levels of heterozygosity in Anzick-1 and CK-13 [143], a large difference between them in recent effective population size seems implausible. Note that the additional NNA-SA gene flow which Scheib et al claim existed would likely increase the amount of segments shared with NNA by Surui and Pima, and would make us less likely to

find evidence of an early split. Our signal here seems to have been produced in spite of this gene flow rather than because of it.

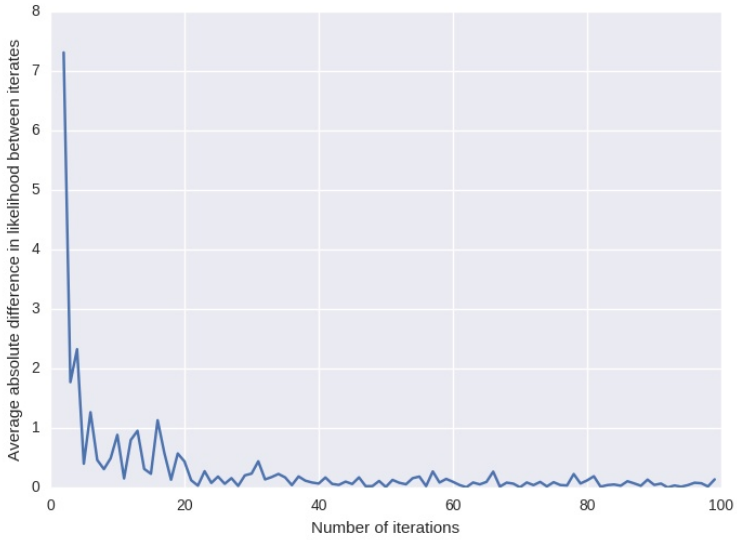
Methods Phased Pima and Surui chr1 genotypes were taken from the Simons Genome Project [81], along with the universal mappability and low-complexity mask x75.fa. Derived alleles were identified using the set of ancestral alleles taken from the 1000 Genomes Project primate EPO panel [19]. I generated 200 MCMC samples of the chr1 ancestral recombination graph (ARG) by applying ARGweaver to the 8 Pima and Surui haplotypes. This was done by allowing 2000 burn-in iterations and sampling every 10th ARG. This process was parallelised by splitting the chromosome into 5Mb regions with a 1Mb overlap and running ARGweaver on each region independently. The subsequent results were combined across the chromosome separately for each MCMC iterate. Haplotype segments for each iterate were identified by the intervals corresponding to single trees in the resulting ARG.

Using the demographic model shown in the main section we derive probabilities of observing the proportion of modern segments matching with either of the ancient sequences. As described above, the ancient sequences are taken from Scheib et al (2018) [143] along with estimates of the ages of the samples. The arguments deriving the relevant probabilities are found in the Appendix to this chapter. Haplotype segments were binned by length using increments of 10bp, since the probabilities of segment matching are dependent on segment length and thus likelihood calculations of the proportions of segment matches per sequence needed to vary by length.

Note that under this model, the ancestral lineages of segments are not independent. This will lead to an overdispersion in observed segment-sharing relative to the binomial variance derived from the demographic model described here. We expect, nonetheless, that the observed mean under the assumption of independence is an unbiased estimate of the expectation derived from the true demographic model. In order to control for this increase in variance it would be appropriate to use a block jackknife approach. However, here we infer a likelihood surface using a beta-binomial distribution (conservatively, we set dispersion parameter $\alpha = 10$, while β is defined so that the mean of the distribution corresponds to the relevant binomial) to account for the increased variance. The log-likelihood is calculated 100 times, each time drawing 8 "observations" corresponding to one random choice of a segment matching proportion for each haplotype from among the MCMC samples of the ARG. Figure 5.7(b) shows the differences in sum of log-likelihoods between each run of the averaging process, illustrating that 100 is sufficient in this case to reach convergence in average likelihood. Scripts for haplotype segment matching, likelihood inference, and plot generation can be accessed at https://github.com/td329/NA-hapmatch-2018.

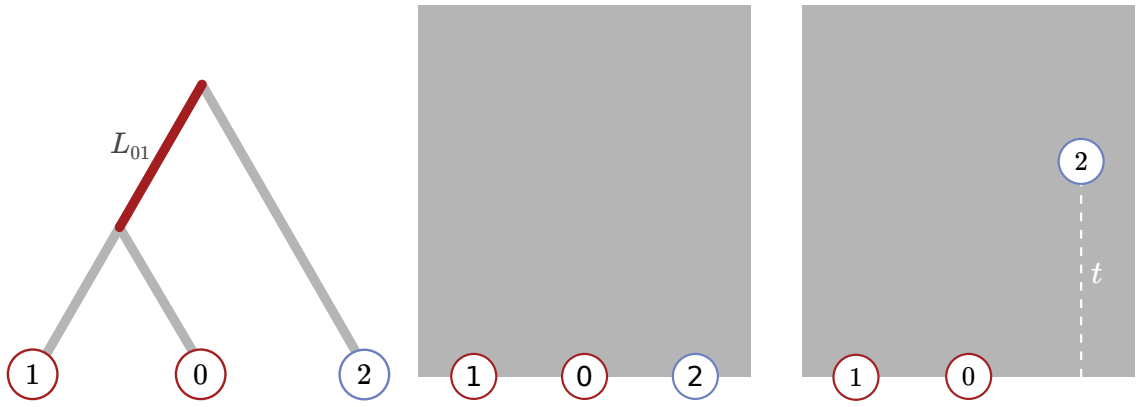


(a) Log-likelihood heatmap under structured population model



(b) Convergence of average likelihood values between iterations

Fig. 5.7 Likelihood analysis of segment matching under the transient structured coalescent model Figure (a) shows a heatmap at which each point represents the likelihood of the observed proportion of segment matches under a transient structured coalescent model with CK-13 and Anzick-1 merge times given by the axes. Description of the model parameters and the sources of those not inferred here are described in the main text. Figure (b) shows convergence of the average values of the heatmap between iterations of random sampling of an MCMC iterate for each haplotype. After each round of sampling, the new average is compared with the previous one, and the absolute difference summed over all points is plotted. The difference between consecutive averages changes little after 40 rounds of sampling. The surface in (a) is taken after 100 iterations.



(a) Branch to be detected **(b)** Case 1 history **(c)** Case 2 history

Fig. 5.8 Case 1 and 2 sampling schemes

5.3 Appendix

A Coalescence Analysis of Mutation-Sharing

We present a coalescent analysis of the distributions of branch- and mutation-sharing under three possible demographic histories. The scenarios increase in complexity. The first is the classical set-up for the 3-coalescent. Here, the sample is exchangeable, representing 3 present-day loci with a selectively neutral history of evolution. In the second, the analysis is extended to allow two loci to have been sampled at (possibly different) earlier times. The third allows differences in sampling time and also relaxes the assumption of panmixia. It builds on the analysis of the first two cases.

In all cases we refer to the modern locus as 0 and the potentially ancient loci as 1 and 2 (Figure 5.8). We are interested in determining the distribution of shared branch lengths between loci 0 and 1, and the probability of a mutation occurring on that branch. We refer to the length variable as L_{01} and the variable number of mutations on the branch as M_{01}. Note that we only consider mutations in the last case, the first two being largely illustrative. By analogy, this analysis also gives us L_{02}. We follow a standard notation in coalescent theory, scaling time in units of $2N_e$ and referring to the random time to first coalescence in the 2- and 3-coalescent processes as T_2 and T_3 respectively [62].

Case 1: Simultaneous sampling and panmixia

It is equally likely that any of the three possible pairs of lineages coalesces after some time T_3. Until this event, no ancestral branches are shared and only if the first coalescence is between

the lineages of loci 0 and 1 will L_{01} be non-zero. If we condition on this event, we obtain

$$\mathbb{P}(L_{01} = l) := f_{01}^1(l) = \begin{cases} \frac{2}{3} & \text{if} \quad l = 0 \\ \frac{1}{3}e^{-l} & \text{if} \quad l > 0. \end{cases} \tag{5.7}$$

The last expression is the chance that the tree topology is correct multiplied by the distribution of T_2, the waiting time between the last coalescent event (Figure 5.8).

Case 2: Staggered sampling and panmixia

We look only at the case in which locus 2 is sampled in the past (Figure 5.8). If we allow both loci 1 and 2 to be ancient sequences, with potentially different sampling times, the analysis is in fact the same: if 1, for example, is more recently sampled, then the problem reduces to the case below since no coalescence can occur before another sample enters the population. Observe that this can also be interpreted as a history in which 0 evolves in a panmictic population and the other loci evolve in parallel populations, migrating into the population of 0 at different historical times (Figure 5.8)

Assume locus 2 is sampled at time t in the past. No ancestral lineage can coalesce with the lineage of 2 before t. We proceed by conditioning on the event that the lineages of 1 and 0 coalesce before then. Let this event be $C_t = \{S < t\}$, where S is a T_2 waiting time. Note that if C_t does not occur (an event we designate with the symbol $\bar{C}_t$), then the problem reduces to Case 1. Thus

$$\begin{aligned} \mathbb{P}(L_{01} = l) := f_{01}^2(l) &= \mathbb{P}(C_t)\mathbb{P}(L_{01} = l|C_t) + \mathbb{P}(\bar{C}_t)\mathbb{P}(L_{01} = l|\bar{C}_t) \\ &= \left(1 - e^{-t}\right)\mathbb{P}(L_{01} = l|C_t) + e^{-t} f_{01}^1(l). \end{aligned}$$

We evaluate the first term by observing that when $S < t$,

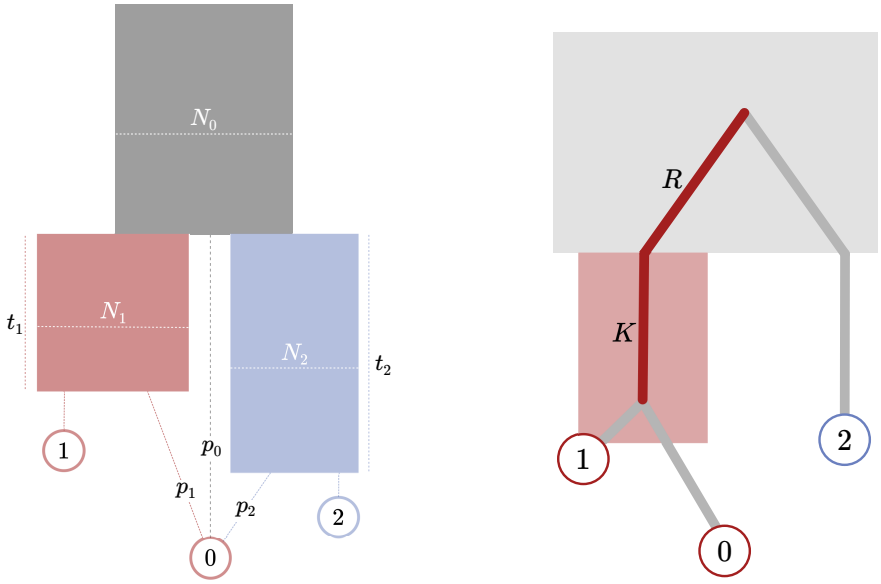
$$L_{01} = K + R \tag{5.8}$$

where $K = t - S^*$ and R is a T_2 waiting time (see Figure 5.8). In this conditional case S^* is distributed according to the truncated exponential $f_{S^*}(s) = e^{-s}/(1 - e^{-t})$. Since K and R are independent, their joint density is

$$f_{K,R}(k, r) = \frac{e^{k-t-r}}{1 - e^{-t}} \qquad 0 < k < t, \quad 0 < r. \tag{5.9}$$

Note that (5.8) implies $K \leq L_{01}$. Coupled with (5.9), this determines the convolution

$$\mathbb{P}(L_{01} = l|C_t) = \int_0^{\min\{l,t\}} f_{K,R}(k, l - k)\, \mathrm{d}k.$$



(a) Case 3 history **(b)** Branch to be detected

Fig. 5.9 Schematic illustration of staggered structured models. (a) is the full parameterised model studied in the Native American case study and for which segment matching probability is analytically computed in the Appendix. (b) indicates the branches used in the Case 3 argument

Evaluating this and combining it with the conditioned expressions, we obtain

$$f_{01}^{2}(l) = \begin{cases} \frac{2}{3}e^{-t} & \text{if} \quad l = 0 \\ e^{-t-l}\left(\frac{1}{2}\left(e^{2\min\{l,t\}} - 1\right) + \frac{1}{3}\right) & \text{if} \quad l > 0. \end{cases} \tag{5.10}$$

Case 3: Staggered sampling and population substructure

We now set up the model in the following way (Figure 5.9). Allow the ancient sequences 1 and 2 to have derived from populations, 1 and 2 respectively, which exchange no migrants. At some time in the past these populations merge into population 0. Locus 0 enters population 1, 2, or neither, with different probabilities. For some time it might coalesce with the ancient locus in the relevant subpopulation, otherwise it can only do so after the merge time, in which the case the problem reduces to Case 1. Despite the additional parameters, the analysis is very similar to that of Case 2.

We will use the following notation:

$N_i = \lambda_i N_0$ the effective population size of population i
P_i the event that path i is followed
p_i the probability of P_i
t_i the time during which 0 can coalesce with i in population i if P_i occurs
C_i the event that coalescence occurs between locus i and 0 in population i before populations mer

The argument again proceeds by conditioning on coalescence occuring within some time window. This time, however, we also need to accomodate population structure, thus additional terms are needed to model the paths lineage 0 might take. First consider the case in which $L_{01} = l$, where $l > 0$:

$$\mathbb{P}(L_{01} = l) = p_1\mathbb{P}(C_1|P_1)\mathbb{P}(L_{01} = l|C_1) + f_{01}^1(l)\left[p_0 + p_1\mathbb{P}(\bar{C}_1|P_1) + p_2\mathbb{P}(\bar{C}_2|P_2)\right] \tag{5.11}$$

The first term is the probability that locus 0 takes path 1 and coalesces with locus 1 before the merge time. Here $\mathbb{P}(C_1|P_1) = 1 - \exp(-t_1/\lambda_1)$. On the other hand, the second term is the probability that no coalescence occurs before the merge time. We see that $\mathbb{P}(\bar{C}_1|P_1) = \exp(-t_1/\lambda_1)$ and similarly, $\mathbb{P}(\bar{C}_2|P_2) = \exp(-t_2/\lambda_2)$.

All that remains to calculate is $\mathbb{P}(L_{01} = l|C_1)$. The argument is as before. Let $L_{01} = K+R$ where R is a T_2 waiting time and $K = t_1 - S^*$ where S^* is distributed according to a truncated exponential. Adjusting for the difference in population size, the joint distribution of independent variables K and R is

$$f_{K,R}(k, r) = \frac{\lambda_1^{-1} e^{(k-t_1)/\lambda_1 - r}}{1 - e^{t_1/\lambda_1}} \qquad 0 < k < t_1, \quad 0 < r.$$

The convolution, as before, is readily determined. Combining it with expression (5.11), we obtain

$$\begin{aligned} \mathbb{P}(L_{01} = l) &= A_1 e^{-l}\left(e^{(1+1/\lambda_1)\min\{t_1, l\}} - 1\right) + (1/3)Be^{-l} \\ \text{where} \quad A_1 &= \frac{p_1 e^{-t_1/\lambda_1}}{\lambda_1 + 1}, \\ \text{and} \quad B &= p_0 + p_1 e^{-t_1/\lambda_1} + p_2 e^{-t_2/\lambda_2}. \end{aligned} \tag{5.12}$$

Another exercise in branch-accounting provides the probability

$$\begin{aligned} \mathbb{P}(L_{01} = 0) &= p_2\mathbb{P}(C_2|P_2) + f_{01}^1(0)\left[p_0 + p_1\mathbb{P}(\bar{C}_1|P_1) + p_2\mathbb{P}(\bar{C}_2|P_2)\right] \\ &= p_2(1 - e^{-t_2/\lambda_2}) + (2/3)B. \end{aligned} \tag{5.13}$$

For the purposes of testing, we can express these probabilities as the cumulative distribution function

$$F_{01}^3(l) = \begin{cases} 0 & \text{if} \quad l < 0 \\ p_2(1 - e^{-t_2/\lambda_2}) + (2/3)B & \text{if} \quad l = 0 \\ F_{01}^3(0) + A_1 \left(\lambda_1 e^{l/\lambda_1} + e^{-l} - \lambda_1 - 1\right) + (1/3)\left(1 - e^{-l}\right) B & \text{if} \quad 0 < l < t_1 \\ F_{01}^3(t_1) + \left(A_1 \left(e^{(1+1/\lambda_1)t_1} - 1\right) + (1/3)B\right)\left(e^{-t_1} - e^{-l}\right) & \text{if} \quad t_1 < l. \end{cases} \tag{5.14}$$

Detecting segment matches

If it were possible to extract marginal genealogical trees from an accurately inferred ARG, we might be able to infer demography based purely on the joint distribution of shared branches. Since these branches are not directly observed, we are interested in the probability that loci share "private" derived alleles. In the case we have been working with so far, this is the same as asking whether loci 0 and 1 share any mutations which are not shared with 2. Under the infinite-sites assumption this is possible if and only if the lineages of 0 and 1 coalesce first. If this kind of mutation has occured, we say that a segment match has been made, or that some branch-sharing is *detectable*.

Under the models we have considered thus far, the probability of detecting branch-sharing can be analytically determined. We demonstrate this in the third case. If M_{01} is the random number of shared private mutations of loci 0 and 1, then a segment matches if $M_{01} > 0$, and thus the relevant quantity to determine is $1 - \mathbb{P}(M_{01} = 0)$. Expressions (5.13) and (5.12) allow us to get this by conditioning on shared branch-lengths. In other words,

$$\mathbb{P}(M_{01} = 0) = 1.\mathbb{P}(L_{01} = 0) + \int_0^\infty \mathbb{P}(M_{01} = 0 | L_{01} = l)\mathbb{P}(L_{01} = l)\, \mathrm{d}l. \tag{5.15}$$

(With a slight abuse of notation we assume, in the second term, that $l > 0$ to avoid including the point mass at $l = 0$. We also suppress the dependence on demographic parameters.) We use the standard coalescent model of mutation and assume that the random number of mutations on some branch is governed by a Poisson process with intensity parameter $\theta/2$. In our case $\theta = 4N_0\mu b$, with μ being the per site per generation mutation rate and b being the number of sites in our locus. The resulting integral is straightforward to evaluate, though is

perhaps easiest to express as a sum of three terms:

$$\begin{aligned} I &= \int_0^\infty e^{-l\theta/2} \left[A_1 e^{-l} \left(e^{(1+1/\lambda_1)\min\{t_1,l\}} - 1 \right) + B(1/3)e^{-l} \right] \mathrm{d}l \\ &= I_1 + I_2 + I_3 \end{aligned}$$

$$\begin{aligned} \text{where} \quad I_1 &= A_1 \left[\frac{e^{(1/\lambda_1 - \theta/2)t_1} - 1}{1/\lambda_1 - \theta/2} + \frac{e^{-(1+\theta/2)t_1} - 1}{1+\theta/2} \right], \\ I_2 &= A_1 \left(e^{(1+1/\lambda_1)t_1} - 1 \right) \left(\frac{e^{-(1+\theta/2)t_1}}{1+\theta/2} \right), \\ \text{and} \quad I_3 &= \frac{B}{3(1+\theta/2)}. \end{aligned} \tag{5.16}$$

Using this notation, the probability that we detect a match between loci 0 and 1 under the third demographic scenario is given by the expression:

$$\mathbb{P}(M_{01} > 0) = 1 - \left(F_{01}^3(0) + I_1 + I_2 + I_3 \right). \tag{5.17}$$

Several extensions of these analyses are obvious next steps. For instance, you might estimate the distributions of any specific number of shared mutations. More complicated demographic models might also be considered.

Chapter 6

Conclusion

Through the course of this thesis I have demonstrated various ways in which demography affects the genetic variation of populations. I have also shown several limitations in our current understanding of the way demographic effects can interact to shape variation. Although I have indicated some ways in which these can be overcome, it is evident that we have a long way to go. Improvements will come from additional data, especially relevant ancient DNA, better theoretical models, as well as more sophisticated statistical tools.

Hominid population size history and the benefit of large samples Methods like PSMC demonstrate that relatively little sequence data is required to produce meaningful estimates of ancestral N_e. In this study we have emphasised the advantages of using larger samples drawn from several dozen individuals. In using more sequences, we improved our understanding of ancestral N_e in *Pan* and *Pongo* populations, and obtained a better understanding of uncertainties inherent in the methods used here. We were also able to combine data from multiple male X chromosomes and obtain new information about the divergence of populations. Wider geographic sampling helped clarify population structure and the historical relationships between subpopulations. In addition to obtaining more sequences, we showed that higher-coverage data improved the consistency of results between individuals from the same population.

In common with humans and gorillas, the populations discussed here have undergone several major bottlenecks and expansions which can be associated with changing environmental conditions and with migrations. Hominid effective population sizes are strikingly similar in magnitude over the last 5 million years. This reflects persistent similarities in the life-histories and ecologies of these species.

Among the great apes, orangutans on Sumatra have the most consistently large N_e over the last 5 million years. This fact is reflected in them having the highest average heterozygosity of any of the great apes [120]. Lowland gorillas, Central chimpanzees and Bornean orangutans have similar levels of heterozygosity, also high, while Western chimpanzees, bonobos, and

humans of recent non-African descent have the lowest. None of the populations here show levels of inbreeding similar to the Eastern lowland gorillas [183]. Given the continuing range contraction of both *Pan* and *Pongo* populations, we might anticipate similar patterns of reduced diversity to arise in the near future.

While the broad features of these population histories are similar to previous results, with larger samples we were able to better understand uncertainty in the method. One way in which this mattered was in the Sumatran orangutan curves, where we noted that it was difficult to pick out the sudden collapse in N_e which other researchers have suggested may have been caused by the Toba supereruption 70 kya [77]. Another advantage of large sample sizes was seen in the way we had a better appreciation of variation within a population, such as in any of the chimpanzee subspecies or the orangutan island populations. In addition, it is striking that even though we observe that it is important to interpret these curves with an awareness that some fluctuations may be caused by cryptic structure, there are distinct differences between the N_e curves of recognised populations. This comparison can give us some confidence that we are already observing the deepest extant subdivisions in the population.

There is still some way to go before we can explain these results by changes in the environment. Such work will require better understanding of genetics, such as clarifying the mutation rate and generation length parameters, as well as environmental changes. One possibility to improve our understanding of the consequences of changing environments might be to study the dynamics of populations of other animals known to occupy similar habitats to those studied here. It would be interesting, in the case of the Toba supereruption, for example, if we could observe population contraction in several arboreal animals on the Sundaland shelf around the time of the eruption. Nonetheless, since selection in the hominids has been shown to vary in intensity according to N_e [120], these results are likely good enough to inform studies comparing selection between great ape populations.

Cross-coalescence estimates refine histories of gene flow We have refined the understanding of the gene flow history of orangutans, chimpanzees and bonobos. These cross-coalescence analyses were new and demonstrated some consistency and some differences which might be expected from the taxonomy of the genera. In the case of the chimpanzees, gene flow history inferred here largely coincides with histories inferred elsewhere, other than the fact that we did not detect the recently proposed secondary gene flow between bonobos and some chimpanzee subspecies. In the orangutans, we noted that there is ongoing gene flow until very recent times between the Tapanuli and Sumatran orangutans. These improved estimates of population divergence were also used as priors in ABC-based analyses [e.g. 24].

It would be useful, looking forward, to obtain high quality phased data from several populations in either genus, and thus be able to run MSMC2 on a larger source of information.

Obtaining more male X chromosomes, to mention one obvious advantage of larger samples, would allow us to improve the accuracy of the cross-population coalescent analyses, but phased autosomal data with low rate of switch-error would be ideal.

There were indications in Chapter 3 that population substructure may be inducing changes in effective population size which make the PSMC curves subtle to interpret. In Chapter 4, in one striking case, that of the central chimpanzees, we showed that island structure with minimal migration was unlikely to produce the effects we were concerned about. It is less commonly observed that cross-coalescence rates may also be influenced by population structure. Using a theoretical analysis of migration estimates between island populations, Slatkin (2005) showed that the presence of an unsampled "ghost" population can influence values of F_{ST} in such a way that two isolated populations appear as if they were exchanging migrants [152]. This would be caused by migration between the observed populations and the ghost population (but not between each other). Given the relationship between F_{ST} and average coalescent rates, it is likely that interpretations of cross-coalescence curves will be subject to similar difficulties. A signal of mergers between two populations may result simply from the presence of a third mediating population. For the purposes of studing speciation this effect may not be consequential, though it has some bearing on the accuracy of demographic models. Richer ancient DNA records could allow us to discount the likelihood such possibilities.

Studying the effects of transient population structure In Chapter 4, we looked at the effects on pairwise coalescent time distributions of transient island structure. In contrast to previous analyses showing the difficulty of interpreting PSMC curves as changes in census population size, we provide some grounds to believe that it is possible to restrict the range of plausible island models in such a way that external evidence might rule out structure as being the cause of some changes in coalescent rate.

The approach that we develop, based on the Kullback-Liebler divergence measure, indicates that in the case of the inflation in coalescent rates that we observed in the central chimpanzees, it seems mostly likely that structure was not the significant cause. However, in the examination of a human population bottleneck we found that our method could not strongly rule out varying structure as an explanation of the change in N_e, although other sources of information do support the bottleneck.

If the main part of the chapter was largely a cause for optimism, the chapter ended with a less encouraging observation: some forms of island histories reduce rates of coalescence and under certain conditions the effects of this structure end up "cancelling out". We point out as well that this non-identifiable situation is not unlikely to occur in biological populations.

Using the ARG to study ancestral structure In the final substantive chapter we proposed looking at segments of shared ancestry extracted from marginal genealogical trees

derived from the ARG of a sample. Although methods like this have been applied to study historical population structure, we show that it is important to model the effects of inheritance, and use a simple demographic model to study the sharing of ancestry with ancient samples. We apply this to study the early structure of the waves of migrating populations in the Americas, supporting a model of deep structure in the ancestors of extant Native American populations.

The explicit structured and time-staggered model can be generalised in several straightforward ways. The first is to allow the population size of the subpopulations to vary in time. The time-dependence might be inferred from a method like PSMC. Other generalisations might adapt the method for a larger number of sample loci, or allow multiple changing island periods. On the statistical implementation of the method, it might be possible to leverage an approach like ABC to jointly fit some of the model parameters, though questions of identifiability pointed out here will likely remain. In this vein, it would also be useful if methods for the inference of ARGs were more computationally efficient. This would allow us to gather much more data about haplotype segment inheritance.

Some further directions for future research One obvious set of developments of the methods in this work would result from relaxing the underlying assumptions I set out in Chapter 2 (in the section "Key assumptions"). To begin with, since we have only looked at neutral sites, we cannot predict the effects of various forms of selection in understanding the demographic models and techniques used throughout this thesis. Another key limitation described in Chapter 2 is the assumption that population structure consists of continuous blocks of individuals which are exchangeable from the perspective of the coalescent process. Although other models, like the stepping-stone or hiearchical island models, can be used here, they suffer from the same stringent limitation. We will not know how more complicated spatial relationships between individuals affect inferences of population size and gene flow histories until we have generative models which are sophisticated enough to model continuous spatial variation.

It would be useful to obtain a principled way of deciding the identifiability of certain models. In Chapter 2, I mentioned several previous results which clarify the kinds of population size changes that we are likely to be able to detect with a certain amount of information about the site frequency spectrum or the distribution of pairwise coalescent times. Analogous results do not exist for population structure. Although my model sets us on a path to being able to distinguish between histories, it would be useful to know how much information we require in the present to detect distinctions between, say, the number of demes in a transient island model. One emphasis in the analyses provided in Chapters 4 and 5, is that where we do not know the principled bounds on our knowledge of the past, we

might have to be content with being able to clarify families of plausible models and appeal to external evidence, where it exists, to reduce the possibilities.

My results have also underscored the importance of finding ancient DNA in order to understand histories of substructure and migration. The reason for this is that changes in demography can quickly obscure signals of previous structure. We aleady knew this from studies of ancient Europeans [104], for example, but it is emphasised in the difficulty we had in narrowing down plausible ranges of island structure in Chapter 4 to explain certain humps in N_e. It would also be helpful, as a general point, to expand the search for ancient DNA on the African continent, if we are looking to clarify human population structure before the out-of-Africa migrations.

On the point of expanding the search for ancient DNA, while there is an ongoing search for human ancient DNA across the world, there has been less of an emphasis on trying to obtain such data from non-hominins. The other great apes are perhaps not as migratory as humans, but by analogy with our own species, we can expect to learn a lot about their past if ancient DNA was discovered. In both the non-human genera studied in this thesis, it would be useful to obtain some ancestral DNA. With the chimpanzees it would be interesting to try to trace the migratory routes of both bonobos and chimpanzees to determine their relationships to the river systems in central Africa, and since the orangutans in Batang Toru might have a distinct history from other Sumatrans, it would be useful to try to capture some for the genetic variation of orangutans in the southern parts of Sumatra, in order to try and trace in more detail the relationships between the populations across islands.

In order to improve on or corroborate the coalescent rate inferences drawn here, we might repeat the analysis using several other approaches which have been developed to address this question. Several of these alternative approaches are described in Chapter 2. A recurrent theme in this kind of analysis is that inferred coalescent rates, whether within populations or between them, can be difficult to interpret, regardless of their accuracy, due to the effect of complex demographic histories. Some of the methods described in Chapter 2 have been developed to determine more complicated demographic models, notably the ABC-based approaches [e.g. 31], but also methods like dical2 [159]. These might be used to complement some of the inferences made by PSMC and MSMC2 in this chapter by helping, for example to exclude population substructure as a cause of some of the fluctuations in historical N_e , or by detecting small amounts of gene flow between populations. At the moment it can be difficult to reconcile the results of PSMC and a method like fastsimcoal. The results in De Manuel et al. [24] are instructive on this point. It would also be useful to use other methods to learn things about the more recent and more distant past, time periods over which the SMC-based methods cannot reliably estimate coalescent rates, at least not without more sequences (in the case of MSMC).

We assume that mutation rates and generation times are constant throughout the periods of interest, and note that we have fixed the mutation rate across the genome. The latter assumption can be relaxed when using the methods here, provided sufficiently accurate mutation rate maps have been produced for the relevant species, though the former set of assumptions are currently fixed into the approaches. There is evidence of mutation rate and generation time evolution across the hominids, most obviously seen in the fact that these values differ between extant species. Therefore the development of methods which can model these parameter shifts might be necessary. Failing that, it would be valuable to understand, by simulation or analysis, the magnitude of these effects on coalescent rates.

References

[1] Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–64.

[2] Amorim, C. E., Nunes, K., Meyer, D., Comas, D., Bortolini, M. C., Salzano, F. M., and Hünemeier, T. (2017). Genetic signature of natural selection in first Americans. *Proceedings of the National Academy of Sciences of the United States of America*, 114(9):2195–2199.

[3] Anka, Z., Séranne, M., and di Primio, R. (2010). Evidence of a large upper-Cretaceous depocentre across the Continent-Ocean boundary of the Congo-Angola basin. Implications for palaeo-drainage and potential ultra-deep source rocks. *Marine and Petroleum Geology*, 27(3):601–611.

[4] Arora, N., Nater, A., van Schaik, C. P., Willems, E. P., van Noordwijk, M. A., Goossens, B., Morf, N., Bastian, M., Knott, C., Morrogh-Bernard, H., Kuze, N., Kanamori, T., Pamungkas, J., Perwitasari-Farajallah, D., Verschoor, E., Warren, K., and Krützen, M. (2010). Effects of Pleistocene glaciations and rivers on the population structure of Bornean orangutans (Pongo pygmaeus). *Proceedings of the National Academy of Sciences of the United States of America*, 107(50):21376–81.

[5] Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

[6] Becquet, C. and Przeworski, M. (2007). A new approach to estimate parameters of speciation models with application to apes. *Genome Research*, 17(10):1505–19.

[7] Beltrame, M. H., Rubel, M. A., and Tishkoff, S. A. (2016). Inferences of African evolutionary history from genomic data. *Current Opinion in Genetics and Development*, 41:159–166.

[8] Bhaskar, A. and Song, Y. S. (2014). Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *The Annals of Statistics*, 42(6):2469–2493.

[9] Bird, M. I., Taylor, D., and Hunt, C. (2005). Palaeoenvironments of insular Southeast Asia during the Last Glacial Period: A savanna corridor in Sundaland? *Quaternary Science Reviews*, 24(20-21):2228–2242.

[10] Bjork, A., Liu, W., Wertheim, J. O., Hahn, B. H., and Worobey, M. (2011). Evolutionary history of chimpanzees inferred from complete mitochondrial genomes. *Molecular Biology and Evolution*, 28(1):615–623.

[11] Bradburd, G., Coop, G., and Ralph, P. (2018). Inferring continuous and discrete population genetic structure across space. *Genetics*, 210(1):33–52.

[12] Burgess, R. and Yang, Z. (2008). Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Molecular Biology and Evolution*, 25:1979–1994.

[13] Cann, R. L., Stoneking, M., and Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature*, 325(6099):31–36.

[14] Caswell, J. L., Mallick, S., Richter, D. J., Neubauer, J., Schirmer, C., Gnerre, S., and Reich, D. (2008). Analysis of chimpanzee history based on genome sequence alignments. *PLoS Genetics*, 4(4).

[15] Cavalli-Sforza, L. L., Menozzi, P., and Piazza, A. (1994). *The History and Geography of Human Genes*. Princeton University Press.

[16] Charlesworth, B. (2009). Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(3):195–205.

[17] Charlesworth, B. and Charlesworth, D. (2010). *Elements of Evolutionary Genetics*. Roberts and Company Publishers, Greenwood Village.

[18] Charlesworth, B. and Charlesworth, D. (2017). Population genetics from 1966 to 2016. *Heredity*, 118(1):2–9.

[19] Consortium, T. . G. P. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

[20] Coolidge, H. J. (1933). Pan paniscus. Pigmy chimpanzee from south of the Congo river. *American Journal of Physical Anthropology*, 18(1):1–59.

[21] Coyne, J. A. and Allen, O. H. (2014). *Speciation*. Sinauer Associates, Sunderland, MA.

[22] Crisci, J. L., Poh, Y.-P., Mahajan, S., and Jensen, J. D. (2013). The impact of equilibrium assumptions on tests of selection. *Frontiers in Genetics*, 4:235.

[23] Csilléry, K., Blum, M. G., Gaggiotti, O. E., and François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology and Evolution*, 25(7):410–418.

[24] de Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V. C., Desai, T., Prado-Martinez, J., Hernandez-Rodriguez, J., Dupanloup, I., Lao, O., Hallast, P., Schmidt, J. M., Heredia-Genestar, J. M., Benazzo, A., Barbujani, G., Peter, B. M., Kuderna, L. F. K., Casals, F., Angedakin, S., Arandjelovic, M., Boesch, C., Kühl, H., Vigilant, L., Langergraber, K., Novembre, J., Gut, M., Gut, I., Navarro, A., Carlsen, F., Andrés, A. M., Siegismund, H. R., Scally, A., Excoffier, L., Tyler-Smith, C., Castellano, S., Xue, Y., Hvilsom, C., and Marques-Bonet, T. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*, 354:477–481.

[25] DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498.

[26] Derrien, T., Estellé, J., Marco Sola, S., Knowles, D. G., Raineri, E., Guigó, R., and Ribeca, P. (2012). Fast Computation and Applications of Genome Mappability. *PLoS ONE*, 7(1).

[27] Dillehay, T. D., Ocampo, C., Saavedra, J., Sawakuchi, A. O., Vega, R. M., Pino, M., Collins, M. B., Scott Cummings, L., Arregui, I., Villagran, X. S., Hartmann, G. A., Mella, M., González, A., and Dix, G. (2015). New Archaeological Evidence for an Early Human Presence at Monte Verde, Chile. *PLoS ONE*, 10(11).

[28] Duncan, L., Shen, H., Gelaye, B., Ressler, K., Feldman, M., Peterson, R., and Domingue, B. (2019). Analysis of Polygenic Score Usage and Performance in Diverse Human Populations. *Nature Communications*, 10(3328).

[29] Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, 28(8):2239–2252.

[30] Ewens, W. J. (1982). On the concept of the effective population size. *Theoretical Population Biology*, 21(3):373–378.

[31] Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., and Foll, M. (2013). Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics*, 9(10).

[32] Falush, D., van Dorp, L., and Lawson, D. (2018). A tutorial on how (not) to over-interpret STRUCTURE/ADMIXTURE bar plots. *Nature Communications*, 9(3258).

[33] Freedman, A. H., Schweizer, R. M., Ortega-Del Vecchyo, D., Han, E., Davis, B. W., Gronau, I., Silva, P. M., Galaverni, M., Fan, Z., Marx, P., Lorente-Galdos, B., Ramirez, O., Hormozdiari, F., Alkan, C., Vilà, C., Squire, K., Geffen, E., Kusak, J., Boyko, A. R., Parker, H. G., Lee, C., Tadigotla, V., Siepel, A., Bustamante, C. D., Harkins, T. T., Nelson, S. F., Marques-Bonet, T., Ostrander, E. A., Wayne, R. K., and Novembre, J. (2016). Demographically-based evaluation of genomic regions under selection in domestic dogs. *PLoS Genetics*, 12(3):1–23.

[34] Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., and François, O. (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics*, 196(4):973–83.

[35] Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *Arxiv*.

[36] Gonder, M. K., Locatelli, S., Ghobrial, L., Mitchell, M. W., Kujawski, J. T., Lankester, F. J., Stewart, C.-B., and Tishkoff, S. A. (2011). Evidence from Cameroon reveals differences in the genetic structure and histories of chimpanzee populations. *Proceedings of the National Academy of Sciences of the United States of America*, 108(12):4766–71.

[37] Goodman, M., Porter, C. A., Czelusniak, J., Page, S. L., Schneider, H., Shoshani, J., Gunnell, G., and Groves, C. P. (1998). Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. *Molecular Phylogenetics and Evolution*, 9(3):585–598.

[38] Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L. F., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., and Pääbo, S. (2010). A draft sequence of the Neandertal genome. *Science (New York, N.Y.)*, 328(5979):710–722.

[39] Griffiths, R. C. (1991). The Two-Locus Ancestral Graph. *Selected Proceedings of the Sheffield Symposium on Applied Probability*, 18:100–117.

[40] Griffiths, R. C. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, 3(4):479–502.

[41] Groves, C. P. (2001). *Primate Taxonomy*. Smithsonian Institution Press.

[42] Hanebuth, T., Stattegger, K., and Grootes., P. M. (2000). Rapid flooding of the Sunda Shelf: a late-glacial sea-level record. *Science*, 288:1033–1035.

[43] Heaney, L. R. (1991). A synopsis of climatic and vegetational change in Southeast Asia. *Climatic Change*, 19(1-2):53–61.

[44] Heather, J. M. and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8.

[45] Henn, B. M., Botigué, L. R., Peischl, S., Dupanloup, I., Lipatov, M., Maples, B. K., Martin, A. R., Musharoff, S., Cann, H., Snyder, M. P., Excoffier, L., Kidd, J. M., and Bustamante, C. D. (2016). Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proceedings of National Academic Sciences of U.S.A.*, 113(4):440–449.

[46] Henn, B. M., Cavalli-Sforza, L. L., and Feldman, M. W. (2019). The great human expansion. *Resonance*, 24(6):711–718.

[47] Hershkovitz, I., Weber, G. W., Quam, R., Duval, M., Grün, R., Kinsley, L., Ayalon, A., Bar-Matthews, M., Valladas, H., Mercier, N., Arsuaga, J. L., Martinón-Torres, M., Bermúdez de Castro, J. M., Fornai, C., Martín-Francés, L., Sarig, R., May, H., Krenn, V. A., Slon, V., Rodríguez, L., García, R., Lorenzo, C., Carretero, J. M., Frumkin, A., Shahack-Gross, R., Bar-Yosef Mayer, D. E., Cui, Y., Wu, X., Peled, N., Groman-Yaroslavski, I., Weissbrod, L., Yeshurun, R., Tsatskin, A., Zaidner, Y., and Weinstein-Evron, M. (2018). The earliest modern humans outside Africa. *Science (New York, N.Y.)*, 359(6374):456–459.

[48] Hey, J. (2001). *Genes, Categories, and Species: The Evolutionary and Cognitive Causes of the Species Problem*. Oxford University Press.

[49] Hodgkinson, A. and Eyre-Walker, A. (2011). Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics*, 12(11):756–766.

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

[51] Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics Applications Note*, 18(2):337–338.

[52] Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., Ni, P., Wang, B., Ou, X., Huasang, Luosang, J., Cuo, Z. X. P., Li, K., Gao, G., Yin, Y., Wang, W., Zhang, X., Xu, X., Yang, H., Li, Y., Wang, J., Wang, J., and Nielsen, R. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, 512(7513):194–197.

[53] Hvilsom, C., Carlsen, F., Heller, R., Jaffré, N., and Siegismund, H. R. (2014). Contrasting demographic histories of the neighboring bonobo and chimpanzee. *Primates*, 55(1):101–112.

[54] International Union for Conservation of Nature and Natural Resources. (2000). The IUCN red list of threatened species.

[55] Joseph, T. A. and Pe'er, I. (2018). Inference of population structure from ancient DNA. *bioRxiv*.

[56] Kaplan, N. L., Darden, T., and Hudson, R. R. (1988). The coalescent process in models with selection. *Genetics*, 120(3):819–29.

[57] Kawamoto, Y., Takemoto, H., Higuchi, S., Sakamaki, T., Hart, J. A., Hart, T. B., Tokuyama, N., Reinartz, G. E., Guislain, P., Dupain, J., Cobden, A. K., Mulavwa, M. N., Yangozene, K., Darroze, S., Devos, C., and Furuichi, T. (2013). Genetic Structure of Wild Bonobo Populations: Diversity of Mitochondrial DNA and Geographical Distribution. *PLoS ONE*, 8(3).

[58] Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Computational Biology*, 12(5):1–22.

[59] Kim, J., Mossel, E., Rácz, M. Z., and Ross, N. (2015). Can one hear the shape of a population history? *Theoretical Population Biology*, 100:26–38.

[60] Kimura, M. (1953). 'Stepping Stone' model of population. *Annual Report of the National Institute of Genetics Japan*, 3:62–63.

[61] Kimura, M. (1971). Theoretical foundation of population genetics at the molecular level. *Theoretical Population Biology*, 2(2):174–208.

[62] Kingman, J. F. C. (1982). On the Genealogy of Large Populations. *Journal of Applied Probability*, 19:27–43.

[63] Kingman, J. F. C. (2000). Origins of the Coalescent: 1974–1982. *Genetics*, 156:1461–1463.

[64] Langergraber, K. E. (2012). Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *PNAS*, 109(39):15716–15721.

[65] Lawson, D. J. (2015). Populations in statistical genetic modelling and inference. In *Population in the Human Sciences: Concepts, Models, Evidence*. Oxford University Press.

[66] Lawson, D. J. and Falush, D. (2012). Population identification using genetic data. *Annual Review of Genomics and Human Genetics*, 13:337–61.

[67] Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics*, 8(1):e1002453.

[68] Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P. H., Schraiber, J. G., Castellano, S., Lipson, M., Berger, B., Economou, C., Bollongino, R., Fu, Q., Bos, K. I., Nordenfelt, S., Li, H., de Filippo, C., Prüfer, K., Sawyer, S., Posth, C., Haak, W., Hallgren, F., Fornander, E., Rohland, N., Delsate, D., Francken, M., Guinet, J.-M., Wahl, J., Ayodo, G., Babiker, H. A., Bailliet, G., Balanovska, E., Balanovsky, O., Barrantes, R., Bedoya, G., Ben-Ami, H., Bene, J., Berrada, F., Bravi, C. M., Brisighelli, F., Busby, G. B. J., Cali, F., Churnosov, M., Cole, D. E. C., Corach, D., Damba, L., van Driem, G., Dryomov, S., Dugoujon, J.-M., Fedorova, S. A., Gallego Romero, I., Gubina, M., Hammer, M., Henn, B. M., Hervig, T., Hodoglugil, U., Jha, A. R., Karachanak-Yankova, S., Khusainova, R., Khusnutdinova, E., Kittles, R., Kivisild, T., Klitz, W., Kučinskas, V., Kushniarevich, A., Laredj, L., Litvinov, S., Loukidis, T., Mahley, R. W., Melegh,

B., Metspalu, E., Molina, J., Mountain, J., Näkkäläjärvi, K., Nesheva, D., Nyambo, T., Osipova, L., Parik, J., Platonov, F., Posukh, O., Romano, V., Rothhammer, F., Rudan, I., Ruizbakiev, R., Sahakyan, H., Sajantila, A., Salas, A., Starikovskaya, E. B., Tarekegn, A., Toncheva, D., Turdikulova, S., Uktveryte, I., Utevska, O., Vasquez, R., Villena, M., Voevoda, M., Winkler, C. A., Yepiskoposyan, L., Zalloua, P., Zemunik, T., Cooper, A., Capelli, C., Thomas, M. G., Ruiz-Linares, A., Tishkoff, S. A., Singh, L., Thangaraj, K., Villems, R., Comas, D., Sukernik, R., Metspalu, M., Meyer, M., Eichler, E. E., Burger, J., Slatkin, M., Pääbo, S., Kelso, J., Reich, D., and Krause, J. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518):409–413.

[69] Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E. C., Cunliffe, B., Lawson, D. J., Falush, D., Freeman, C., Pirinen, M., Myers, S., Robinson, M., Donnelly, P., Bodmer, W., Donnelly, P., and Bodmer, W. (2015). The fine-scale genetic structure of the British population. *Nature*, 519(7543):309–314.

[70] Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496.

[71] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup, . G. P. D. P. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–9.

[72] Li, H., Ruan, J., Durbin, R., Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–1858.

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

[74] Liang, M. and Nielsen, R. (2014). The lengths of admixture tracts. *Genetics*, 197(3):953–967.

[75] Lipson, M. and Reich, D. (2017). A Working Model of the Deep Relationships of Diverse Modern Human Genetic Lineages Outside of Africa. *Molecular Biology and Evolution*, 34(4):889–902.

[76] Lobon, I., Tucci, S., De Manuel, M., Ghirotto, S., Benazzo, A., Prado-Martinez, J., Lorente-Galdos, B., Nam, K., Dabad, M., Hernandez-Rodriguez, J., Comas, D., Navarro, A., Schierup, M. H., Andres, A. M., Barbujani, G., Hvilsom, C., and Marques-Bonet, T. (2016). Demographic history of the genus Pan inferred from whole mitochondrial genome reconstructions. *Genome Biology and Evolution*, 8(6):2020–2030.

[77] Locke, D. P., Hillier, L. W., Warren, W. C., Worley, K. C., Nazareth, L. V., Muzny, D. M., Yang, S.-P., Wang, Z., Chinwalla, A. T., Minx, P., Mitreva, M., Cook, L., Delehaunty, K. D., Fronick, C., Schmidt, H., Fulton, L. a., Fulton, R. S., Nelson, J. O., Magrini, V., Pohl, C., Graves, T. a., Markovic, C., Cree, A., Dinh, H. H., Hume, J., Kovar, C. L., Fowler, G. R., Lunter, G., Meader, S., Heger, A., Ponting, C. P., Marques-Bonet, T., Alkan, C., Chen, L., Cheng, Z., Kidd, J. M., Eichler, E. E., White, S., Searle, S., Vilella, A. J., Chen, Y., Flicek, P., Ma, J., Raney, B., Suh, B., Burhans, R., Herrero, J., Haussler, D., Faria, R., Fernando, O., Darré, F., Farré, D., Gazave, E., Oliva, M., Navarro, A., Roberto, R., Capozzi, O., Archidiacono, N., Della Valle, G., Purgato, S., Rocchi, M., Konkel, M. K., Walker, J. a., Ullmer, B., Batzer, M. a., Smit, A. F. a., Hubley, R., Casola, C., Schrider,

D. R., Hahn, M. W., Quesada, V., Puente, X. S., Ordoñez, G. R., López-Otín, C., Vinar, T., Brejova, B., Ratan, A., Harris, R. S., Miller, W., Kosiol, C., Lawson, H. a., Taliwal, V., Martins, A. L., Siepel, A., Roychoudhury, A., Ma, X., Degenhardt, J., Bustamante, C. D., Gutenkunst, R. N., Mailund, T., Dutheil, J. Y., Hobolth, A., Schierup, M. H., Ryder, O. a., Yoshinaga, Y., de Jong, P. J., Weinstock, G. M., Rogers, J., Mardis, E. R., Gibbs, R. a., and Wilson, R. K. (2011). Comparative and demographic analysis of orang-utan genomes. *Nature*, 469(7331):529–33.

[78] Ma, X., Kelley, J. L., Eilertson, K., Musharoff, S., Degenhardt, J. D., Martins, A. L., Vinar, T., Kosiol, C., Siepel, A., Gutenkunst, R. N., and Bustamante, C. D. (2013). Population Genomic Analysis Reveals a Rich Speciation and Demographic History of Orang-utans (Pongo pygmaeus and Pongo abelii). *PLoS ONE*, 8(10):e77175.

[79] Malaspinas, A.-S., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., Bergström, A., Athanasiadis, G., Cheng, J. Y., Crawford, J. E., Heupink, T. H., Macholdt, E., Peischl, S., Rasmussen, S., Schiffels, S., Subramanian, S., Wright, J. L., Albrechtsen, A., Barbieri, C., Dupanloup, I., Eriksson, A., Margaryan, A., Moltke, I., Pugach, I., Korneliussen, T. S., Levkivskyi, I. P., Moreno-Mayar, J. V., Ni, S., Racimo, F., Sikora, M., Xue, Y., Aghakhanian, F. A., Brucato, N., Brunak, S., Campos, P. F., Clark, W., Ellingvåg, S., Fourmile, G., Gerbault, P., Injie, D., Koki, G., Leavesley, M., Logan, B., Lynch, A., Matisoo-Smith, E. A., McAllister, P. J., Mentzer, A. J., Metspalu, M., Migliano, A. B., Murgha, L., Phipps, M. E., Pomat, W., Reynolds, D., Ricaut, F.-X., Siba, P., Thomas, M. G., Wales, T., Wall, C. M., Oppenheimer, S. J., Tyler-Smith, C., Durbin, R., Dortch, J., Manica, A., Schierup, M. H., Foley, R. A., Lahr, M. M., Bowern, C., Wall, J. D., Mailund, T., Stoneking, M., Nielsen, R., Sandhu, M. S., Excoffier, L., Lambert, D. M., and Willerslev, E. (2016). A genomic history of Aboriginal Australia. *Nature*, 538(7624):207–214.

[80] Malinsky, M., Svardal, H., Tyers, A. M., Miska, E. A., Genner, M. J., Turner, G. F., and Durbin, R. (2018). Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature Ecology & Evolution*, 2(12):1940–1955.

[81] Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., Spence, J. P., Song, Y. S., Poletti, G., Balloux, F., Van Driem, G., De Knijff, P., Romero, I. G., Jha, A. R., Behar, D. M., Bravi, C. M., Capelli, C., Hervig, T., Moreno-Estrada, A., Posukh, O. L., Balanovska, E., Balanovsky, O., Karachanak-Yankova, S., Sahakyan, H., Toncheva, D., Yepiskoposyan, L., Tyler-Smith, C., Xue, Y., Abdullah, M. S., Ruiz-Linares, A., Beall, C. M., Di Rienzo, A., Jeong, C., Starikovskaya, E. B., Metspalu, E., Parik, J., Villems, R., Henn, B. M., Hodoglugil, U., Mahley, R., Sajantila, A., Stamatoyannopoulos, G., Wee, J. T., Khusainova, R., Khusnutdinova, E., Litvinov, S., Ayodo, G., Comas, D., Hammer, M. F., Kivisild, T., Klitz, W., Winkler, C. A., Labuda, D., Bamshad, M., Jorde, L. B., Tishkoff, S. A., Watkins, W. S., Metspalu, M., Dryomov, S., Sukernik, R., Singh, L., Thangaraj, K., Paäbo, S., Kelso, J., Patterson, N., and Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206.

[82] Marjoram, P. and Wall, J. D. (2006). Fast "coalescent" simulation. *BMC Genetics*, 7(1):16.

[83] Marques-Bonet, T., Ryder, O. A., and Eichler, E. E. (2009). Sequencing primate genomes: What have we learned? *Annual Review of Genomics and Human Genetics*, 10(1):355–386.

[84] Marshall, A. J., Ancrenaz, M., Brearley, F. Q., Fredriksson, G. M., Ghaffar, N., Heydon, M., Husson, S. J., Leighton, M., McConkey, K. R., Morrogh-Bernard, H. C., Proctor, J., van Schaik, C. P., Yeager, C. P., and Wich, S. A. (2009). The effects of forest phenology and floristics on populations of Bornean and Sumatran orangutans. In *Orangutans : Geographic Variation in Behavioral Ecology and Conservation by Wich, Serge A. Utami-Atmoko, S.Suci et al*, pages 97–117. Oxford University Press.

[85] Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., Daly, M. J., Bustamante, C. D., and Kenny, E. E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, 100(4):635–649.

[86] Mattle-Greminger, M. P., Sonay, T. B., Nater, A., Pybus, M., Desai, T., de Valles, G., Casals, F., Scally, A., Bertanpetit, J., Marques-Bonet, T., van Schaik, C. P., Anisimova, M., and Krützen, M. (2018). Genomes reveal marked differences in the adaptive evolution between orangutan species. *Genome Biology*, 6:1–13.

[87] Mazet, O., Rodríguez, W., Grusea, S., Boitard, S., and Chikhi, L. (2016). On the importance of being structured: instantaneous coalescence rates and human evolution–lessons for ancestral population size inference? *Heredity*, 116(4):362–71.

[88] McBrearty, S. and Jablonski, N. G. (2005). First fossil chimpanzee. *Nature*, 437(7055):105.

[89] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297–303.

[90] McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genetics*, 5(10).

[91] McVean, G. and Cardin, N. (2005). Approximating the coalescent with recombination. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360(1459):1387–93.

[92] Menozzi, P., Piazza, A., and Cavalli-Sforza, L. (1978). Synthetic maps of human gene frequencies in Europeans. *Science*, 201(4358):786–92.

[93] Mérot, C., Salazar, C., Merrill, R. M., Jiggins, C. D., and Joron, M. (2017). What shapes the continuum of reproductive isolation? Lessons from Heliconius butterflies. *Proceedings of the Royal Society B: Biological Sciences*, 284(1856):20170335.

[94] Mirazón Lahr, M. (2016). The shaping of human diversity: filters, boundaries and transitions. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 371(1698):20150241.

[95] Miyata, T., Hayashida, H., Kuma, K., Mitsuyasu, K., and Yasunaga, T. (1987). Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harbor Symposia on Quantitative Biology*, 52:863–7.

[96] Möhle, M. (2006). On Sampling Distributions for Coalescent Processes with Simultaneous Multiple Collisions. *Bernoulli*, 12(1):35–53.

[97] Montinaro, F., Busby, G. B. J., Gonzalez-Santos, M., Oosthuitzen, O., Oosthuitzen, E., Anagnostou, P., Destro-Bisol, G., Pascali, V. L., and Capelli, C. (2017). Complex ancient genetic structure and cultural transitions in southern African populations. *Genetics*, 205(1):303–316.

[98] Moorjani, P., Amorim, C. E. G., Arndt, P. F., and Przeworski, M. (2016). Variation in the molecular clock of primates. *Proceedings of the National Academy of Sciences*, 113(38):10607–10612.

[99] Myers, S., Fefferman, C., and Patterson, N. (2008). Can one learn history from the allelic spectrum? *Theoretical Population Biology*, 73(3):342–348.

[100] Narasimhan, V. M., Patterson, N. J., Moorjani, P., Lazaridis, I., Mark, L., Mallick, S., Rohland, N., Bernardos, R., Kim, A. M., Nakatsuka, N., Olalde, I., Coppa, A., Mallory, J., Moiseyev, V., Monge, J., Olivieri, L. M., Adamski, N., Broomandkhoshbacht, N., Candilio, F., Cheronet, O., Culleton, B. J., Ferry, M., Fernandes, D., Gamarra, B., Gaudio, D., Hajdinjak, M., Harney, E., Harper, T. K., Keating, D., Lawson, A.-M., Michel, M., Novak, M., Oppenheimer, J., Rai, N., Sirak, K., Slon, V., Stewardson, K., Zhang, Z., Akhatov, G., Bagashev, A. N., Baitanayev, B., Bonora, G. L., Chikisheva, T., Derevianko, A., Dmitry, E., Douka, K., Dubova, N., Epimakhov, A., Freilich, S., Fuller, D., Goryachev, A., Gromov, A., Hanks, B., Judd, M., Kazizov, E., Khokhlov, A., Kitov, E., Kupriyanova, E., Kuznetsov, P., Luiselli, D., Maksudov, F., Meiklejohn, C., Merrett, D. C., Micheli, R., Mochalov, O., Muhammed, Z., Mustafakulov, S., Nayak, A., Petrovna, R. M., Pettner, D., Potts, R., Razhev, D., Sarno, S., Sikhymbaevae, K., Slepchenko, S. M., Stepanova, N., Svyatko, S., Vasilyev, S., Vidale, M., Voyakin, D., Yermolayeva, A., Zubova, A., Shinde, V. S., Lalueza-Fox, C., Meyer, M., Anthony, D., Boivin, N., Thangaraj, K., Kennett, D., Frachetti, M., Pinhasi, R., and Reich, D. (2018). The genomic formation of South and Central Asia. *bioRxiv*, 292581.

[101] Nater, A., Arora, N., Greminger, M. P., van Schaik, C. P., Singleton, I., Wich, S. A., Fredriksson, G., Perwitasari-Farajallah, D., Pamungkas, J., and Krützen, M. (2013). Marked population structure and recent migration in the critically endangered Sumatran orangutan (Pongo abelii). *Journal of Heredity*, 104(1):2–13.

[102] Nater, A., Mattle-Greminger, M. P., Nurcahyo, A., Nowak, M. G., de Manuel, M., Desai, T., Groves, C., Pybus, M., Sonay, T. B., Roos, C., Lameira, A. R., Wich, S. A., Askew, J., Davila-Ross, M., Fredriksson, G., de Valles, G., Casals, F., Prado-Martinez, J., Goossens, B., Verschoor, E. J., Warren, K. S., Singleton, I., Marques, D. A., Pamungkas, J., Perwitasari-Farajallah, D., Rianti, P., Tuuga, A., Gut, I. G., Gut, M., Orozco-terWengel, P., van Schaik, C. P., Bertranpetit, J., Anisimova, M., Scally, A., Marques-Bonet, T., Meijaard, E., and Krützen, M. (2017). Morphometric, behavioral, and genomic evidence for a new orangutan species. *Current Biology*, 27(22):3576–3577.

[103] Nater, A., Nietlisbach, P., Arora, N., van Schaik, C. P., van Noordwijk, M. A., Willems, E. P., Singleton, I., Wich, S. A., Goossens, B., Warren, K. S., Verschoor, E. J., Perwitasari-Farajallah, D., Pamungkas, J., and Krützen, M. (2011). Sex-biased dispersal and volcanic activities shaped phylogeographic patterns of extant orangutans (genus: Pongo). *Molecular Biology and Evolution*, 28(8):2275–2288.

[104] Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., and Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature*, 541(7637):302.

[105] Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., and Clark, A. G. (2007). Recent and ongoing selection in the human genome. *Nature Reviews Genetics*, 8(11):857–868.

[106] Nordborg, M. (1997). Structured coalescent processes on different time scales. *Genetics*, 146(4):1501–1514.

[107] Nordborg, M. and Krone, S. M. (2002). Separation of time scales and convergence to the coalescent in structured populations. *Modern Developments in Theoretical Populations Genetics: The Legacy of Gustave Malécot*, pages 194–232.

[108] Novembre, J. and Barton, N. H. (2018). Tread lightly interpreting polygenic tests of selection. *Genetics*, 208(4):1351–1355.

[109] Novembre, J. and Peter, B. M. (2016). Recent advances in the study of fine-scale population structure in humans. *Current Opinion in Genetics and Development*, 41:98–105.

[110] Novembre, J. and Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40(5):646–649.

[111] Oates, J. F., Groves, C. P., and Jenkins, P. D. (2009). The type locality of Pan troglodytes vellerosus (Gray, 1862), and implications for the nomenclature of West African chimpanzees. *Primates*, 50(1):78–80.

[112] Ostrander, E. A., Wayne, R. K., Freedman, A. H., and Davis, B. W. (2017). Demographic history, selection and functional diversity of the canine genome. *Nature Reviews Genetics*, 18(12):705–720.

[113] Pachter, L. (2014). What is principal component analysis? *Bits of DNA*.

[114] Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. *Genetics*, 192(3):1065–1093.

[115] Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, 2(12).

[116] Pedersen, M. W., Ruter, A., Schweger, C., Friebe, H., Staff, R. A., Kjeldsen, K. K., Mendoza, M. L., Beaudoin, A. B., Zutter, C., Larsen, N. K., Potter, B. A., Nielsen, R., Rainville, R. A., Orlando, L., Meltzer, D. J., Kjær, K. H., and Willerslev, E. (2016). Postglacial viability and colonization in North America's ice-free corridor. *Nature*, 537(7618):45–49.

[117] Peter, B. M. (2016). Admixture, population structure, and F-statistics. *Genetics*, 202(4):1485–501.

[118] Pickrell, J. K. and Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, 8(11).

[119] Pouyet, F., Aeschbacher, S., Thiéry, A., and Excoffier, L. (2018). Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *eLife*, 7:e36317.

[120] Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'Connor, T. D., Santpere, G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., Munch, K., Hobolth, A., Halager, A. E., Malig, M., Hernandez-Rodriguez, J., Hernando-Herraez, I., Prüfer, K., Pybus, M., Johnstone, L., Lachmann, M., Alkan, C., Twigg, D., Petit, N., Baker, C., Hormozdiari, F., Fernandez-Callejo, M., Dabad, M., Wilson, M. L., Stevison, L., Camprubí, C., Carvalho, T., Ruiz-Herrera, A., Vives, L., Mele, M., Abello, T., Kondova, I., Bontrop, R. E., Pusey, A., Lankester, F., Kiyang, J. a.,

Bergl, R. a., Lonsdorf, E., Myers, S., Ventura, M., Gagneux, P., Comas, D., Siegismund, H., Blanc, J., Agueda-Calpena, L., Gut, M., Fulton, L., Tishkoff, S. a., Mullikin, J. C., Wilson, R. K., Gut, I. G., Gonder, M. K., Ryder, O. a., Hahn, B. H., Navarro, A., Akey, J. M., Bertranpetit, J., Reich, D., Mailund, T., Schierup, M. H., Hvilsom, C., Andrés, A. M., Wall, J. D., Bustamante, C. D., Hammer, M. F., Eichler, E. E., and Marques-Bonet, T. (2013). Great ape genetic diversity and population history. *Nature*, 499(7459):471–5.

[121] Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

[122] Provine, W. B. (2001). *The origins of theoretical population genetics: With a new afterword*. University of Chicago Press.

[123] Prüfer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J. R., Walenz, B., Koren, S., Sutton, G., Kodira, C., Winer, R., Knight, J. R., Mullikin, J. C., Meader, S. J., Ponting, C. P., Lunter, G., Higashino, S., Hobolth, A., Dutheil, J., Karakoç, E., Alkan, C., Sajjadian, S., Catacchio, C. R., Ventura, M., Marques-Bonet, T., Eichler, E. E., André, C., Atencia, R., Mugisha, L., Junhold, J., Patterson, N., Siebauer, M., Good, J. M., Fischer, A., Ptak, S. E., Lachmann, M., Symer, D. E., Mailund, T., Schierup, M. H., Andrés, A. M., Kelso, J., and Pääbo, S. (2012). The bonobo genome compared with the chimpanzee and human genomes. *Nature*, 486(7404):527.

[124] Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

[125] Racimo, F., Gokhman, D., Fumagalli, M., Ko, A., Hansen, T., Moltke, I., Albrechtsen, A., Carmel, L., Huerta-Sánchez, E., and Nielsen, R. (2017). Archaic adaptive introgression in TBX15/WARS2. *Molecular Biology and Evolution*, 34(3):509–524.

[126] Raghavan, M., Skoglund, P., Graf, K. E., Metspalu, M., Albrechtsen, A., Moltke, I., Rasmussen, S., Stafford, T. W., Orlando, L., Metspalu, E., Karmin, M., Tambets, K., Rootsi, S., Mägi, R., Campos, P. F., Balanovska, E., Balanovsky, O., Khusnutdinova, E., Litvinov, S., Osipova, L. P., Fedorova, S. A., Voevoda, M. I., Degiorgio, M., Sicheritz-Ponten, T., Brunak, S., Demeshchenko, S., Kivisild, T., Villems, R., Nielsen, R., Jakobsson, M., and Willerslev, E. (2014). Upper palaeolithic Siberian genome reveals dual ancestry of native Americans. *Nature*, 505(7481):87–91.

[127] Raghavan, M., Steinrücken, M., Harris, K., Schiffels, S., Rasmussen, S., DeGiorgio, M., Albrechtsen, A., Valdiosera, C., Ávila-Arcos, M. C., Malaspinas, A.-S., Eriksson, A., Moltke, I., Metspalu, M., Homburger, J. R., Wall, J., Cornejo, O. E., Moreno-Mayar, J. V., Korneliussen, T. S., Pierre, T., Rasmussen, M., Campos, P. F., de Barros Damgaard, P., Allentoft, M. E., Lindo, J., Metspalu, E., Rodríguez-Varela, R., Mansilla, J., Henrickson, C., Seguin-Orlando, A., Malmström, H., Stafford, T., Shringarpure, S. S., Moreno-Estrada, A., Karmin, M., Tambets, K., Bergström, A., Xue, Y., Warmuth, V., Friend, A. D., Singarayer, J., Valdes, P., Balloux, F., Leboreiro, I., Vera, J. L., Rangel-Villalobos, H., Pettener, D., Luiselli, D., Davis, L. G., Heyer, E., Zollikofer, C. P. E., Ponce de León, M. S., Smith, C. I., Grimes, V., Pike, K.-A., Deal, M., Fuller, B. T., Arriaza, B., Standen, V., Luz, M. F., Ricaut, F., Guidon, N., Osipova, L., Voevoda, M. I., Posukh, O. L., Balanovsky, O., Lavryashina, M., Bogunov, Y., Khusnutdinova, E., Gubina, M., Balanovska, E., Fedorova, S., Litvinov, S., Malyarchuk, B., Derenko, M., Mosher, M. J., Archer, D., Cybulski, J., Petzelt, B., Mitchell, J., Worl, R., Norman, P. J., Parham, P., Kemp, B. M., Kivisild, T., Tyler-Smith, C., Sandhu, M. S., Crawford, M., Villems, R., Smith, D. G., Waters, M. R., Goebel, T., Johnson, J. R., Malhi, R. S., Jakobsson, M., Meltzer, D. J., Manica,

A., Durbin, R., Bustamante, C. D., Song, Y. S., Nielsen, R., and Willerslev, E. (2015). Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science*, 349(6250).

[128] Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., and Cavalli-Sforza, L. L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 102(44):15942–7.

[129] Rasmussen, M., Anzick, S. L., Waters, M. R., Skoglund, P., DeGiorgio, M., Stafford Jr, T. W., Rasmussen, S., Moltke, I., Albrechtsen, A., Doyle, S. M., David Poznik, G., Gudmundsdottir, V., Yadav, R., Malaspinas, A.-S., Stockton White, S. V., Allentoft, M. E., Cornejo, O. E., Korneliussen, S., Meltzer, D. J., Pierre, T. L., Stenderup, J., Saag, L., Warmuth, V., Cabrita Lopes, M., Malhi, R. S., Brunak, S., Sicheritz-Ponten, T., Barnes, I., and Collins, M. (2014a). The genome of a late Pleistocene human from a Clovis burial site in western Montana. *Nature*, 506:225–229.

[130] Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. (2014b). Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, 10(5).

[131] Reich, D. (2018). *Who We Are and How We Got Here: Ancient DNA and the new science of the human past*. Oxford University Press.

[132] Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L. F., Maricic, T., Good, J. M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E. E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M. V., Derevianko, A. P., Hublin, J.-J., Kelso, J., Slatkin, M., and Pääbo, S. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468(7327):1053–1060.

[133] Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M. V., Rojas, W., Duque, C., Mesa, N., García, L. F., Triana, O., Blair, S., Maestre, A., Dib, J. C., Bravi, C. M., Bailliet, G., Corach, D., Hünemeier, T., Bortolini, M. C., Salzano, F. M., Petzl-Erler, M. L., Acuña-Alonzo, V., Aguilar-Salinas, C., Canizales-Quinteros, S., Tusié-Luna, T., Riba, L., Rodríguez-Cruz, M., Lopez-Alarcón, M., Coral-Vazquez, R., Canto-Cetina, T., Silva-Zolezzi, I., Fernandez-Lopez, J. C., Contreras, A. V., Jimenez-Sanchez, G., Gómez-Vázquez, M. J., Molina, J., Carracedo, Á., Salas, A., Gallo, C., Poletti, G., Witonsky, D. B., Alkorta-Aranburu, G., Sukernik, R. I., Osipova, L., Fedorova, S. A., Vasquez, R., Villena, M., Moreau, C., Barrantes, R., Pauls, D., Excoffier, L., Bedoya, G., Rothhammer, F., Dugoujon, J. M., Larrouy, G., Klitz, W., Labuda, D., Kidd, J., Kidd, K., Di Rienzo, A., Freimer, N. B., Price, A. L., and Ruiz-Linares, A. (2012). Reconstructing Native American population history. *Nature*, 488(7411):370–374.

[134] Reich, D., Thangaraj, K., Patterson, N., Price, A. L., and Singh, L. (2009). Reconstructing Indian population history. *Nature*, 461(7263):489–494.

[135] Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology*, 26(3):303–304.

[136] Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., Feldman, M. W., Bodmer, W. F., Bonne-Tamir, B., Cambon-Thomsen, A., Chen, Z., Chu, J., Carcassi, C., Contu, L., Du, R., Excoffier, L., Ferrara, G. B., Friedlaender, J. S., Groot, H., Gurwitz, D., Jenkins, T., Herrera, R. J., Huang, X., Kidd, J., Kidd, K. K., Langaney, A., Lin, A. A., Mehdi, S. Q., Parham, P., Piazza, A., Pistillo, M. P., Qian,

Y., Shu, Q., Xu, J., Zhu, S., Weber, J. L., Greely, H. T., Feldman, M. W., Thomas, G., Dausset, J., and Cavalli-Sforza, L. L. (2002). Genetic structure of human populations. *Science*, 298(5602):2381–5.

[137] Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N., and Bierne, N. (2016). Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS Biology*, 14(12).

[138] Runge, J. (2007). The Congo River, Central Africa. In *Large Rivers: Geomorphology and Management*, pages 293–301. John Wiley & Sons.

[139] Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., and Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, 507(7492):354–357.

[140] Sankararaman, S., Mallick, S., Patterson, N., Sankararaman, S., Mallick, S., Patterson, N., and Reich, D. (2016). The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Current Biology*, 26(9):1241–1247.

[141] Santiago, E. and Caballero, A. (2005). Variation after a selective sweep in a subdivided population. *Genetics*, 169(1):475–83.

[142] Scally, A. (2016). Mutation rates and the evolution of germline structure. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 371(1699):20150137.

[143] Scheib, C. L., Li, H., Desai, T., Link, V., Kendall, C., Dewar, G., Griffith, P. W., Mörseburg, A., Johnson, J. R., Potter, A., Kerr, S. L., Endicott, P., Lindo, J., Haber, M., Xue, Y., Tyler-Smith, C., Sandhu, M. S., Lorenz, J. G., Randall, T. D., Faltyskova, Z., Pagani, L., Danecek, P., O'Connell, T. C., Martz, P., Boraas, A. S., Byrd, B. F., Leventhal, A., Cambra, R., Williamson, R., Lesage, L., Holguin, B., Soto, E. Y.-D., Rosas, J., Metspalu, M., Stock, J. T., Manica, A., Scally, A., Wegmann, D., Malhi, R. S., and Kivisild, T. (2018). Ancient human parallel lineages within North America contributed to a coastal expansion. *Science*, 360(6392):1024–1027.

[144] Schiffels, S. and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8):919–925.

[145] Schrider, D. R., Shanku, A. G., and Kern, A. D. (2016). Effects of linked selective sweeps on demographic inference and model selection. *Genetics*, 204(3):1207–1223.

[146] Simons, Y. B., Bullaughey, K., Hudson, R. R., and Sella, G. (2018). A population genetic interpretation of GWAS findings for human quantitative traits. *PLoS Biology*, 16(3).

[147] Sjödin, P., Kaj, I., Krone, S., Lascoux, M., and Nordborg, M. (2005). On the meaning and existence of an effective population size. *Genetics*, 169(2):1061–1070.

[148] Skoglund, P., Mallick, S., Bortolini, M. C., Chennagiri, N., Hünemeier, T., Petzl-Erler, M. L., Salzano, F. M., Patterson, N., and Reich, D. (2015). Genetic evidence for two founding populations of the Americas. *Nature*, 525(7567):104.

[149] Skoglund, P. and Reich, D. (2016). A genomic view of the peopling of the Americas. *Current Opinion in Genetics and Development*, 41:27–35.

[150] Skoglund, P., Thompson, J. C., Prendergast, M. E., Mittnik, A., Sirak, K., Hajdinjak, M., Salie, T., Rohland, N., Mallick, S., Peltzer, A., Heinze, A., Olalde, I., Ferry, M., Harney, E., Michel, M., Stewardson, K., Cerezo-Román, J. I., Chiumia, C., Crowther, A., Gomani-Chindebvu, E., Gidna, A. O., Grillo, K. M., Helenius, I. T., Hellenthal, G., Helm, R., Horton, M., López, S., Mabulla, A. Z., Parkington, J., Shipton, C., Thomas, M. G., Tibesasa, R., Welling, M., Hayes, V. M., Kennett, D. J., Ramesar, R., Meyer, M., Pääbo, S., Patterson, N., Morris, A. G., Boivin, N., Pinhasi, R., Krause, J., and Reich, D. (2017). Reconstructing prehistoric African population structure. *Cell*, 171(1):59–71.e21.

[151] Slatkin, M. (1991). Inbreeding coefficients and coalescence times. *Genetical research*, 58(2):167–75.

[152] Slatkin, M. (2005). Seeing ghosts: The effect of unsampled populations on migration rates estimated for sampled populations. *Molecular Ecology*, 14(1):67–73.

[153] Slatkin, M. and Hudson, R. R. (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, 129(2):555–62.

[154] Slatkin, M. and Voelm, L. (1991). FST in a hierarchical island model. *Genetics*, 127(3):627–9.

[155] Slatkin, M. and Wiehe, T. (1998). Genetic hitch-hiking in a subdivided population. *Genetical research*, 71(2):155–60.

[156] Spehar, S. N., Sheil, D., Harrison, T., Louys, J., Ancrenaz, M., Marshall, A. J., Wich, S. A., Bruford, M. W., and Meijaard, E. (2018). Orangutans venture out of the rainforest and into the anthropocene. *Science Advances*, 4(6):1–14.

[157] Staab, P. R., Zhu, S., Metzler, D., and Lunter, G. (2015). Scrm: Efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*, 31(10):1680–1682.

[158] Stankiewicz, J. and de Wit, M. J. (2006). A proposed drainage evolution model for Central Africa—Did the Congo flow east? *Journal of African Earth Sciences*, 44(1):75–84.

[159] Steinrücken, M., Kamm, J. A., and Song, Y. S. (2019). Inference of complex population histories using whole-genome sequences from multiple populations. *Proceedings of the National Academy of Sciences*, 116(34):17115–17120.

[160] Sul, J. H., Martin, L. S., and Eskin, E. (2018). Population structure in genetic studies: Confounding factors and mixed models. *PLoS Genetics*, 14(12).

[161] Takemoto, H., Kawamoto, Y., and Furuichi, T. (2015). How did bonobos come to range south of the congo river? Reconsideration of the divergence of Pan paniscus from other Pan populations. *Evolutionary Anthropology*, 24(5):170–184.

[162] Takemoto, H., Kawamoto, Y., Higuchi, S., Makinose, E., Hart, J. A., Hart, T. B., Sakamaki, T., Tokuyama, N., Reinartz, G. E., Guislain, P., Dupain, J., Cobden, A. K., Mulavwa, M. N., Yangozene, K., Darroze, S., Devos, C., and Furuichi, T. (2017). The mitochondrial ancestor of bonobos and the origin of their major haplogroups. *PLoS ONE*, 12(5):1–14.

[163] Tang, H., Peng, J., Wang, P., and Risch, N. J. (2005). Estimation of individual admixture: Analytical and study design considerations. *Genetic Epidemiology*, 28:289–301.

[164] Thompson, J. A. M. (2003). A model of the biogeographical journey from Proto-pan to Pan paniscus. *Primates*, 44(2):191–197.

[165] Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J.-M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J. H., Mortensen, H., Nyambo, T. B., Omar, S. A., Powell, K., Pretorius, G. S., Smith, M. W., Thera, M. A., Wambebe, C., Weber, J. L., and Williams, S. M. (2009). The genetic structure and history of Africans and African Americans. *Science*, 324(5930):1035–44.

[166] Venn, O., Turner, I., Mathieson, I., De Groot, N., Bontrop, R., and McVean, G. (2014). Strong male bias drives germline mutation in chimpanzees. *Science*, 344(6189):1272–1275.

[167] Vervaecke, H. and Vanelsacker, L. (1992). Hybrids between common chimpanzees (Pan-Troglodytes) and pygmy chimpanzees (Pan-Paniscus) in captivity. *Mammalia*, 56(4):667–669.

[168] Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., Mccarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 years of GWAS discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101:5–22.

[169] Voris, H. K. (2000). Maps of Pleistocene sea levels in Southeast Asia: Shorelines, river systems and time durations. *Journal of Biogeography*, 27(5):1153–1167.

[170] Wakeley, J. (1999). Nonequilibrium migration in human history. *Genetics*, 153(4):1863–71.

[171] Wakeley, J. (2009). *Coalescent Theory: An Introduction*. Macmillan Learning.

[172] Wang, K., Mathieson, I., Schiffels, S., and View, M. (2019). Tracking human population structure through time from whole genome sequences. *bioRxiv*.

[173] Waterson, R. H., Lander, E. S., and Wilson, R. K. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87.

[174] Westaway, K. E., Louys, J., Awe, R. D., Morwood, M. J., Price, G. J., Zhao, J.-x., Aubert, M., Joannes-Boyau, R., Smith, T. M., Skinner, M. M., Compton, T., Bailey, R. M., van den Bergh, G. D., de Vos, J., Pike, A. W. G., Stringer, C., Saptomo, E. W., Rizal, Y., Zaim, J., Santoso, W. D., Trihascaryo, A., Kinsley, L., and Sulistyanto, B. (2017). An early modern human presence in Sumatra 73,000–63,000 years ago. *Nature*, 548(7667):322–325.

[175] Whitlock, M. C. and Barton, N. H. (1997). The effective size of subdivided populations. *Genetics*, 146(1):427–441.

[176] Wich, S. A., Meijaard, E., Marshall, A. J., Husson, S., Ancrenaz, M., Lacy, R. C., Van Schaik, C. P., Sugardjito, J., Simorangkir, T., Traylor-Holzer, K., Doughty, M., Supriatna, J., Dennis, R., Gumal, M., Knott, C. D., and Singleton, I. (2008). Distribution and conservation status of the orang-utan (Pongo spp.) on Borneo and Sumatra: How many remain? *Oryx*, 42(3):329–339.

[177] Wich, S. A., Usher, G., Peters, H. H., Khakim, M. F. R., Nowak, M. G., and Frederiksson, G. M. (2014). Preliminary Data on the Highland Sumatran Orangutans (Pongo abelii) of Batang Toru. In Grow, N., Gursky-doyen, S., and Krzton, A., editors, *High Altitude Primates*, pages 265–284. Springer.

[178] Wich, S. A., Utami Atmoko, S. S., Setia, T. M., and van Schaik, C. P. (2009). *Orangutans: Geographic Variation in Behavioral Ecology and Conservation*. Oxford University Press.

[179] Wich, S. A., Vogel, E. R., Larsen, M. D., Fredriksson, G., Leighton, M., Yeager, C. P., Brearley, F. Q., van Schaik, C. P., and Marshall, A. J. (2011). Forest fruit production is higher on Sumatra than on Borneo. *PLoS ONE*, 6(6):36–38.

[180] Wiuf, C. and Hein, J. (1999). Recombination as a point process along sequences. *Theoretical population biology*, 55(3):248–259.

[181] Wright, S. (1943). Isolation by distance. *Genetics*, 28(2):114.

[182] Wright, S. (1949). The Genetical Structure of Populations. *Annals of Eugenics*, 15(1):323–354.

[183] Xue, Y., Prado-martinez, J., Sudmant, P. H., Narasimhan, V., Ayub, Q., Szpak, M., Frandsen, P., Chen, Y., Yngvadottir, B., Cooper, D. N., Manuel, M. D., Hernandez-rodriguez, J., Lobon, I., Siegismund, H. R., Pagani, L., Quail, M. A., Hvilsom, C., Mudakikwa, A., Eichler, E. E., Cranfield, M. R., Marques-bonet, T., and Tyler-smith, C. (2015). Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science*, 348(6231):242.

[184] Xue, Y., Wang, Q., Long, Q., Ng, B. L., Swerdlow, H., Burton, J., Skuce, C., Taylor, R., Abdellah, Z., Zhao, Y., Asan, MacArthur, D. G., Quail, M. A., Carter, N. P., Yang, H., and Tyler-Smith, C. (2009). Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Current Biology*, 19(17):1453–7.