Inferring dated genealogies from modern and ancient genomes



Anthony Wilder Lauritano Wohns

Merton College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2021

Acknowledgements

My deepest appreciation to the best supervisors one could hope for: Gil McVean, Yan Wong, and Jerome Kelleher. Thank you to Emma Jones for making it all happen.

Thank you to the members of the McVean group: Xilin Jiang, Marcus Tutert, Chris Eijsbouts, Jason Hendry, Chaimaa Fadil, Yidong Zhang, Patrick Albers and others, as well as Ben Jeffery, for stimulating scientific conversations and many memories in Oxford.

I am also indebted to my mentors and collaborators at the Universities of Cambridge, Tartu, and Leuven: Toomas Kivisild, Freddi Scheib, and Ruoyun Hui. Thank you to David Reich for scientific guidance during our collaboration and for supporting my visit to his lab at Harvard. Thanks also to Ali Akbari and Shop Mallick of the Reich group. I greatly appreciate helpful discussions with Eric Lander (Broad Institute), as well as Terry Capellini and Daniel Lieberman (Harvard).

My most profound gratitude to the Rhodes Trust for making it possible for me to study at Oxford and pursue a DPhil, and especially Mary Eaton for her help with absolutely everything during my Fellowship.

I would like to acknowledge the Oxford-Moussouris-Merton Graduate Scholarship for funding the latter years of my DPhil. Merton College provided a wonderful home for me at Oxford and supported my travel to conferences.

The computational aspects of this research were supported by the Wellcome Trust (Core Award 203141/Z/16/Z) and the NIHR Oxford BRC. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. I would like to thank the Oxford Big Data research computing team, specifically Adam Huffman and Robert Esnouf.

The UK Biobank data was accessed under application number 12788. I would like to acknowledge all the human subjects, past and present, who are included in the datasets used in this thesis.

Thanks to Laurel Chandler for her constant support and for walking with me through Chiltern villages and Scottish rivers alike. Thank you to the Chandler family for your gracious hospitality. Thanks to Kirk Smith for sharing the experience of running from Cambridge to Oxford, and to our friends who served as the support crew for us over the 144 kilometres.

My deepest gratitude to my parents and siblings for teaching me to take risks and instilling the desire to never stop learning. Thank you to my grandmother Taeko for leading by example.

Abstract

Tree sequences encode the ancestral history of a sample of genomes, reflecting the sum total of what is theoretically knowable about the sample's evolutionary history. However, to realise the promise of this structure, scalable methods that infer time-resolved tree sequences from both modern and ancient samples are needed. In this thesis, I address this challenge and thus realise the power of tree sequences for population genetic analysis.

First, I investigate a key assumption of the tree sequence inference algorithm `tsinfer`: that allele frequency reflects relative allele age. I derive an exact iterative formula for the expected accuracy of frequency in assigning the relative age of pairs of alleles at unlinked loci under the coalescent. Using coalescent simulations approximating human demographic history, I further show that allele frequency correctly orders 91.6% of the mutation pairs that are relevant for topological inference.

Second, I introduce new statistics to summarise genetic relationships between individuals and populations. *Genealogical Nearest Neighbours* quantifies topological relationships in a tree sequence and is applied to population sequencing datasets. The related *Genomic Descent* statistic quantifies descent from ancient individuals. I use this statistic to reveal links between Neolithic Britons and modern Sardinian individuals, as well as between an Afanasievo family of four and UK Biobank participants from traditionally Celtic regions of Britain and Ireland.

Third, I introduce `tsdate`, an approximate Bayesian method for inferring the age of ancestral haplotypes conditional on a tree sequence topology. I validate the algorithm using simulations and empirical analyses, the latter indicating that 96.3% of `tsdate`-estimated derived allele ages from Chromosome 20 of the 1000 Genomes Project are consistent with radiocarbon-dated ancient samples.

Finally, I use an iterative approach combining `tsinfer` and `tsdate` to infer the largest time-resolved genealogy constructed to date with 3,719 ancient and modern samples from eight datasets. The genealogy provides estimates of the distribution of time to most recent common ancestor for the 215 populations in the datasets, recovers patterns of descent from Archaics, and suggests that up to 17.5% of variant sites could be the result of more than one ancestral mutation. I also introduce a simple, non-parametric estimator of the spatiotemporal properties of inferred ancestors that recapitulates key events in hominin history. These results demonstrate that tree sequences are a powerful means of synthesising genetic data and provide rich insights into human evolution.

Contents

List of Figures **xi**

List of Abbreviations **xv**

1 Introduction **1**

1.1 Gene genealogies . . . 3
1.1.1 Motivating example: *GDF5* . . . 4
1.2 Coalescent theory . . . 5
1.2.1 Basic properties of the coalescent . . . 6
1.2.2 Coalescent waiting times . . . 8
1.2.3 Mutations in a coalescent genealogy . . . 11
1.2.4 The coalescent with recombination . . . 12
1.3 Ancestral recombination graphs and tree sequences . . . 14
1.4 Allele age estimation . . . 19
1.4.1 Frequency-based estimators of allele age . . . 20
1.4.2 Intra-allelic variability-based estimators of allele age . . . 22
1.5 Genome-wide genealogical inference . . . 23
1.5.1 Inferring tree sequence topologies with `tsinfer` . . . 26
1.5.2 `tsinfer` version 0.2 . . . 29
1.6 Ancient DNA . . . 31
1.6.1 Population genetic inference using ancient DNA . . . 32
1.7 Thesis overview . . . 33

2 The Relative Age of Alleles **35**

2.1 Introduction . . . 35
2.2 The importance of relative allele age . . . 37
2.2.1 Relative allele age in `tsinfer` . . . 39
2.3 Probability of allele frequency reflecting relative age . . . 40
2.3.1 Conditional genealogies and allele age . . . 40
2.3.2 Allele frequency and relative age at unlinked loci . . . 42
2.4 Simulating the pairwise time-ordering of alleles . . . 45
2.4.1 Empirically determined genotype error profile . . . 46

2.4.2 Comparison to GEVA 47
2.4.3 "Out of Africa" model 48
2.4.4 Ancestrally connected pairs of mutations 48
2.5 Simulation results 48
2.6 Discussion 54

3 Analysing tree sequence topologies 57
3.1 Introduction 57
3.2 *Genealogical nearest neighbours* 59
3.2.1 1000 Genomes Project and Simons Genome Diversity Project 61
3.3 *Genomic descent* 62
3.3.1 Trumpington Neolithic samples 64
3.3.2 Afanasievo family and UK Biobank 68
3.4 Discussion 73

4 Inferring dated tree sequences from modern and ancient samples 77
4.1 Introduction 77
4.2 `tsdate` 78
4.2.1 Conditional coalescent prior 79
4.2.2 Inside-Outside algorithm 82
4.2.3 Worked examples of `tsdate` 86
4.3 Iterative inference with ancient and modern samples 92
4.4 Discussion 94

5 Evaluating the accuracy of dated tree sequences 97
5.1 Introduction 97
5.2 Evaluating `tsdate` 98
5.2.1 Accuracy of the `tsdate` prior 99
5.2.2 Accuracy of age estimation 99
5.2.3 Computational efficiency 104
5.3 Evaluating the accuracy of incorporating ancient samples 107
5.3.1 Ancient samples and allele age estimation accuracy 107
5.3.2 Archaic descent simulation 109
5.4 Empirical estimates of 1000 Genomes Project variant ages on chromosome 20 115
5.4.1 Materials and Methods 115
5.4.2 Results 116
5.5 Discussion 117

6 A unified genealogy of modern and ancient genomes 121
6.1 Introduction . . . 122
6.2 Integrating and analysing diverse datasets with tree sequences . . . 123
6.3 A unified genealogy of modern and ancient human genomes . . . 124
6.3.1 Materials and methods . . . 125
6.3.2 Results . . . 128
6.4 Fine-scale worldwide population structure . . . 130
6.4.1 Methods . . . 130
6.4.2 Results . . . 131
6.5 Modern and ancient population relationships . . . 132
6.5.1 Methods . . . 132
6.5.2 Results . . . 134
6.6 Non-parametric inference of spatio-temporal dynamics in human history . . . 137
6.6.1 Methods . . . 138
6.6.2 Results . . . 138
6.7 Discussion . . . 140

7 Discussion 143
7.1 *GDF5* and the unified genealogy . . . 145
7.1.1 Marginal tree describing *rs4911178* . . . 145
7.1.2 The age of *rs4911178* . . . 146
7.1.3 Fine-scale genetic relationships at this locus . . . 147
7.1.4 Possible recurrent mutations at *GDF5* 5' UTR . . . 148
7.1.5 Conclusion . . . 150
7.2 Future Directions . . . 151
7.2.1 Data cleaning and annotation . . . 152
7.2.2 Probabilistic ancestor generation algorithm . . . 153
7.2.3 Phasing using a diploid Li and Stephens algorithm . . . 154
7.2.4 Tree sequences and fine-mapping . . . 156
7.2.5 Tree sequences and data sharing . . . 156
7.2.6 Conclusion . . . 157

Appendices

A Supplementary Video and Interactive Figure from Wohns et al. (2021) 161

B Joint Authorship Statement 163

C Lessons Learned 167

References 169

List of Figures

1.1 The genealogy of a sample of three chromosomes taken from a population of six chromosomes 4
1.2 *rs4911178* worldwide allele frequency and signals of positive selection (reproduction of Figure 5 from Capellini et al. (2017)) 6
1.3 Maximum likelihood tree describing *rs4911178* and associated haplotypes (reproduction of Figure 6 from Capellini et al. (2017)) 7
1.4 One possible coalescent tree with five sampled lineages 8
1.5 Recombination as a subtree-prune-and-regraft (SPR) operation when moving spatially along a tree sequence 15
1.6 Two representations of the genealogy of four samples 17
1.7 "Site" and "branch" definitions of the allele frequency spectrum using a tree sequence 18
1.8 Models of *GDF5* haplotype evolution and migration (reproduction of Figure 6 (b and c) from Capellini et al. (2017)) 20
1.9 A biallelic locus with allele age (T_m) and age of the MRCA of samples with the resulting allele (T_{ca}) 21
1.10 Schematic of ancestral haplotype reconstruction (reproduction of Figure 2 from Kelleher et al. (2019)) 28

2.1 Frequency incorrectly reflects the relative age of alleles in a genealogical tree 36
2.2 Frequency incorrectly reflects the relative age of a pair of alleles in adjacent marginal genealogies in a tree sequence 37
2.3 Five sampled lineages at two independent loci 43
2.4 Heatmap of the expectation of the event that an allele carried by i samples is younger than an allele carried by j samples 45
2.5 Two ancestrally connected pairs of mutations at adjacent genealogies separated by a recombination event 49
2.6 Suitability of frequency as a proxy for relative allele age 50
2.7 Accuracy of frequency and `GEVA` as measures of relative allele age . 51
2.8 Accuracy of frequency and `GEVA` as measures of the relative age of alleles at pairs of ancestrally connected mutations 52

2.9 Average frequency of ancestrally connected allele pairs 53
2.10 Accuracy of allele frequency at ancestrally connected alleles of similar frequency 54

3.1 Genealogical nearest neighbours schematic (adapted from (Kelleher et al., 2019) Figure S19) 60
3.2 Tree sequence characterisation of global genome diversity (adapted from Kelleher et al. (2019) Figure 4) 63
3.3 *Genomic descent* from Trumpington samples in the 1000 Genomes Project (adapted from Scheib et al. (2019) Figure 2) 67
3.4 *Genomic descent* from Trumpington samples in the Simons Genome Diversity Project (adapted from Scheib et al. (2019) Figure S2) . . 68
3.5 Evaluating phasing accuracy in the Afanasievo quartet 71
3.6 *Genomic descent* from Afanasievo family haplotypes among "white British" UK Biobank participants on Chromosomes 17-22 73
3.7 *Genomic descent* statistic from Afanasievo family haplotypes in NUTS 2 regions of the United Kingdom and in the Republic of Ireland 74

4.1 Example tree with $n = 3$ samples and one mutation, showing the calculation of a marginal prior distribution 86
4.2 Marginal prior and posterior distributions for tree example 87
4.3 Filling in the Inside Matrix 88
4.4 Filling in the Outside Matrix 89
4.5 Example tree sequence with $n = 3$ samples, one mutation, and two marginal trees, showing the calculation of a marginal prior distribution 90
4.6 Marginal prior and posterior distributions for tree sequence example 91
4.7 Schematic overview of iterative approach 93

5.1 Accuracy of the `tsdate` node age prior distribution 100
5.2 Accuracy of `tsdate` at various mutation rate settings 101
5.3 Comparison of the accuracy of `tsdate`, `GEVA`, and `Relate` on simulated tree sequence topologies 103
5.4 Comparison of the accuracy of `tsdate`, `GEVA`, and `Relate` on simulated tree sequence topologies 104
5.5 Evaluation of `tsdate` on simulated Chromosome 20 data 105
5.6 Accuracy of variant age estimates from `tsdate`, `Relate` and `GEVA` . 106
5.7 Scaling properties of `tsdate` compared to `tsinfer`, `Relate`, and `GEVA` 107
5.8 Accuracy of incorporating ancient samples into iterative inference methodology 109
5.9 Precision-recall curves for recovering relationships between archaic and modern samples using inferred tree sequences 113

5.10 Comparison of mutation age estimates from three methods (`tsdate`, `Relate` and `GEVA`) using 3,734 ancient samples at 76,889 variants on Chromosome 20 . . . 117
5.11 Estimated age of TGP derived alleles from `tsdate`, `GEVA` and `Relate` compared to allele frequency . . . 118
5.12 Distribution of TGP derived allele age estimates from `tsdate`, `GEVA` and `Relate` . . . 118
5.13 Comparison of TGP derived allele age estimates from `tsdate`, `GEVA` and `Relate` . . . 119

6.1 The relationship between metrics of variant-calling error and the number of mutations in the inferred tree sequence . . . 130
6.2 Clustered heatmap showing the average time to the most recent common ancestor (TMRCA) on Chromosome 20 for haplotypes within pairs of the 215 populations in the HGDP, TGP, SGDP, and ancient samples . . . 133
6.3 Descent from Denisovan proxy ancestors on chromosome 20 . . . 135
6.4 Inferred patterns of descent from Afanasievo proxy ancestors on Chromosome 20 . . . 136
6.5 Descent from Vindija Neanderthal proxy haplotypes . . . 137
6.6 Visualisation of the non-parametric estimator of ancestor geographic location for HGDP, SGDP, Neanderthal, Denisovan, and Afanasievo samples . . . 140

7.1 Marginal tree from the unified genealogy describing *rs4911178* . . . 146
7.2 Allele frequency and estimated age from `tsdate` of 1,178,347 derived alleles on Chromosome 20q . . . 149
7.3 Schematic of our proposed probabilistic method of ancestral haplotype inference . . . 154
7.4 Edge breakpoints can be used to identify potential phasing errors. . 155

List of Abbreviations

aDNA . . . . . Ancient DNA

AFS . . . . . Allele frequency spectrum

ARG . . . . . Ancestral recombination graph

CMS . . . . . Composite of multiple signals method (Grossman et al., 2013)

DAG . . . . . Directed acyclic graph

GEVA . . . . Genealogical Estimation of Variant Age (Albers and McVean, 2020)

GNN . . . . . Genealogical nearest neighbours (Kelleher et al., 2019)

HGDP . . . . Human Genome Diversity Project (Bergström et al., 2020)

HMM . . . . . Hidden Markov model

kb . . . . . Kilobase

kya . . . . . Thousand years ago

LD . . . . . Linkage disequilibrium

LS . . . . . Li and Stephens model (Li and Stephens, 2003)

Mb . . . . . Megabase

MCMC . . . . Markov chain Monte Carlo

MRCA . . . . Most recent common ancestor

PCA . . . . . Principal component analysis

RMSLE . . . Root mean squared log error

SGDP . . . . . Simons Genome Diversity Project (Mallick et al., 2016)

SMC . . . . . Sequentially Markovian Coalescent (McVean and Cardin, 2005a)

SPR . . . . . Subtree prune-and-regraft

TGP . . . . . 1000 Genomes Project (1000 Genomes Project Consortium, 2015)

TMRCA . . . Time to most recent common ancestor

UKB . . . . . United Kingdom Biobank (Bycroft et al., 2018)

UTR . . . . . Untranslated region

VCF . . . . . Variant call format

1 Introduction

Contents

1.1 Gene genealogies 3
1.1.1 Motivating example: *GDF5* 4
1.2 Coalescent theory 5
1.2.1 Basic properties of the coalescent 6
1.2.2 Coalescent waiting times 8
1.2.3 Mutations in a coalescent genealogy 11
1.2.4 The coalescent with recombination 12
1.3 Ancestral recombination graphs and tree sequences . . 14
1.4 Allele age estimation 19
1.4.1 Frequency-based estimators of allele age 20
1.4.2 Intra-allelic variability-based estimators of allele age . . 22
1.5 Genome-wide genealogical inference 23
1.5.1 Inferring tree sequence topologies with `tsinfer` 26
1.5.2 `tsinfer` version 0.2 29
1.6 Ancient DNA 31
1.6.1 Population genetic inference using ancient DNA 32
1.7 Thesis overview 33

Trees encode evolutionary relationships among organisms and the genomes that describe them. In the ∼180 years since Darwin sketched an evolutionary tree in his notebook, they have developed from abstractions to indispensable tools in biology (Darwin, 1837). Thanks to methodological advances and a torrent of data, inferring the "Tree of Life," which encodes the ancestral relationships of *all* organisms to

one another, now appears within reach. Indeed, the 2015 draft of the *Open Tree of Life* relates 2.3 million species to one another (Hinchliff et al., 2015), realising Darwin's musing that "The time will come I believe, though I shall not live to see it, when we shall have very fairly true genealogical trees of each great kingdom of nature" (Darwin, 1857; Kapli, Yang, and Telford, 2020).

However, a single tree often cannot fully capture the complexity of ancestral relationships between organisms, even those from different species. For example, according to the well-established phylogeny of great ape species, chimpanzees and bonobos are more closely related to one another than either is to humans (Morin et al., 1994). While this is true in the aggregate, the sequencing of the bonobo genome demonstrated that at 3.3% percent of the human genome, humans share a more recent common ancestor with either bonobos or chimpanzees than these species do with one another (Prüfer et al., 2012). Multiple phenomena can explain this observation, including incomplete lineage sorting, gene transfer, and gene duplication followed by selective loss (Szöllősi et al., 2015), showing that a single tree does not always suffice in representing ancestral relationships. Multiple networks have been used to represent the complexities of ancestral relationships in the presence of such processes (Morrison, 2016).

In this thesis, I am concerned with ancestral genealogies in the presence of recombination: a ubiquitous process which breaks the single tree assumption by shuffling the genetic material transmitted from parent to child. While the impact of recombination on the ancestral history of a sample of genomes has been the subject of research since the 1980s (Hudson, 1983a; Hudson and Kaplan, 1985), the structure now commonly used to model ancestral relationships in this setting was introduced a decade later. This structure is a directed acyclic graph (DAG) called the ancestral recombination graph (ARG) (Griffiths, 1991; Griffiths and Marjoram, 1996). The ARG records coalescence and recombination events with mutations superimposed, from which it is possible to infer selection, migration and demographic processes. ARGs theoretically encode all that is knowable about the history of a sample and are therefore critically important in population genetics

(Arenas, 2013; Gusfield, 2014). Despite this, the practical use of ARGs has been hindered by the difficulty of exploring the overwhelming space of graphs compatible with even comparatively small datasets (Rasmussen et al., 2014). However, in the past two years, new methods have been introduced which allow for the scalable inference of genome-wide genealogies from much larger datasets (Speidel et al., 2019; Kelleher et al., 2019). These advances enable promising new directions in both population and statistical genetics (Harris, 2019; Stern et al., 2021).

In this thesis, I introduce and validate methods for inferring time-resolved tree sequences, an efficient data structure encoding ancestral relationships and haplotypes which has strong links to the ARG. I also present statistics which summarise the population genetic relationships they encode. These inference methods and statistics are applied to infer and analyse tree sequences from modern population sequencing datasets and ancient samples. The remainder of this chapter provides motivation for examining these questions, as well as background information and key results from the relevant literature. I provide an example of the biological and medical importance of genealogies through exploration of the evolutionary history of an enhancer of the *GDF5* gene which has been strongly associated with both human height and osteoarthritis risk (Capellini et al., 2017). I then provide a literature review of key theoretical results from coalescent theory, a survey of efforts to infer allele ages and genome-wide genealogies, and an overview of the use of ancient DNA (aDNA) in population genetic studies. Finally, I provide an overview of the following chapters in this thesis.

1.1 Gene genealogies

The gene genealogy is a structure that describes the ancestral history of fragments of genetic sequence sampled from a population (the word "gene" refers to such fragments in this context). By tracing ancestral relationships back in time, the gene genealogy assumes a tree structure at discrete positions in the genome (see Figure 1.1), and thus has strong links to other trees such as species trees (Felsenstein, 2004). The distinguishing feature of species trees and gene genealogies is that the

former represents the ancestral history of entire species, while the latter represents the ancestral history of particular fragments of genetic material (Szöllősi et al., 2015)

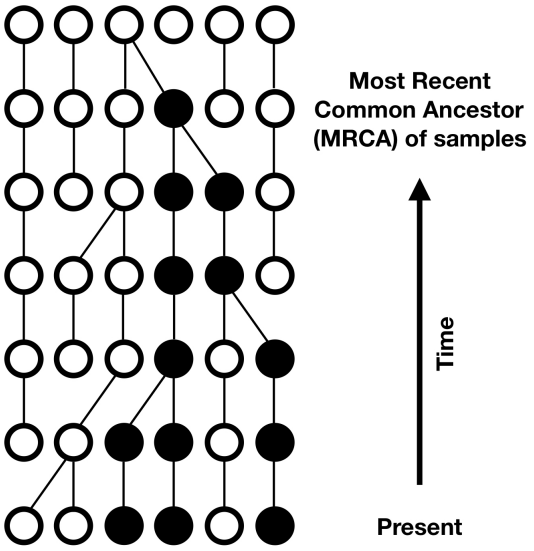


Figure 1.1: The genealogy of a sample of three chromosomes (filled in circles) taken from a population of six chromosomes.

Before introducing theoretical results concerning the properties of gene genealogies, I first give a practical example of their role in understanding human evolution and disease.

1.1.1 Motivating example: *GDF5*

One application of the gene genealogy is in studying variants associated with a phenotype of interest. The evolutionary history of the gene *GDF5*, which encodes the protein growth differentiation factor 5, is particularly illustrative. Common variants in *GDF5* are highly associated with increased risk of osteoarthritis (Miyamoto et al., 2007; Loughlin, 2015) and decreased height (Sanna et al., 2008). Knocking out *GDF5* in mouse models results in a range of morphological deformities including stunted limbs and the loss of bones, tendons, and ligaments (Grüneberg and Lee, 1973; Storm et al., 1994; Mikic et al., 2004; Harada et al., 2007; Daans, Luyten, and Lories, 2011).

Multiple studies have found genetic signatures of positive selection at variants in *GDF5* associated with increased disease risk, hereafter referred to as "risk alleles"

(Voight et al., 2006; Sabeti et al., 2007; Amato et al., 2011; Wu et al., 2012). Capellini et al. (2017) identified a novel downstream enhancer called *GROW1B* that is necessary for normal *GDF5* expression and sought to investigate whether derived variants in this enhancer had been a target of selection. The authors found that the ancestral allele at the sole common risk allele in this locus is highly conserved in placental mammals, but that the derived allele is at high frequency in non-Africans, lower frequency in some African populations (see Figure 1.2a), and was present in Neanderthals and Denisovans. This variant also shows strong signals of positive selection (see Figure 1.2b).

The authors constructed a maximum-likelihood tree using haplotypes from the 1000 Genomes Project (TGP) (1000 Genomes Project Consortium, 2015) together with Neanderthal (Prüfer et al., 2014; Prüfer et al., 2017) and Denisovan (Meyer et al., 2012) haplotypes at the 130 kb regulatory locus of *GDF5* (see Figure 1.3). This approach offers a number of advantages when seeking to understand the evolution of a locus of interest, as trees are a highly intuitive and biologically informative representation of the evolutionary processes that shape genetic diversity.

In Sections 1.4 and 1.5, I describe how the authors use this tree to interpret the evolutionary history of the *GDF5* enhancer. I also highlight some fundamental limitations to the single-tree approach. In Chapter 7, I return to this example to demonstrate that tree sequences provide further insight into the history of this important locus.

1.2 Coalescent theory

The coalescent (Kingman, 1982; Hudson, 1983b; Tajima, 1983) is the most commonly used statistical model of the ancestral history of a sample of genes. Since its inception, the body of work around coalescent theory has grown tremendously; in-depth reviews of the literature are readily available (Hudson, 1990; Tavaré, 2004; Hein, Schierup, and Wiuf, 2005; Wakeley, 2009). In this section, I present results from coalescent theory that are fundamental to the genealogical inference methodologies presented in this thesis.

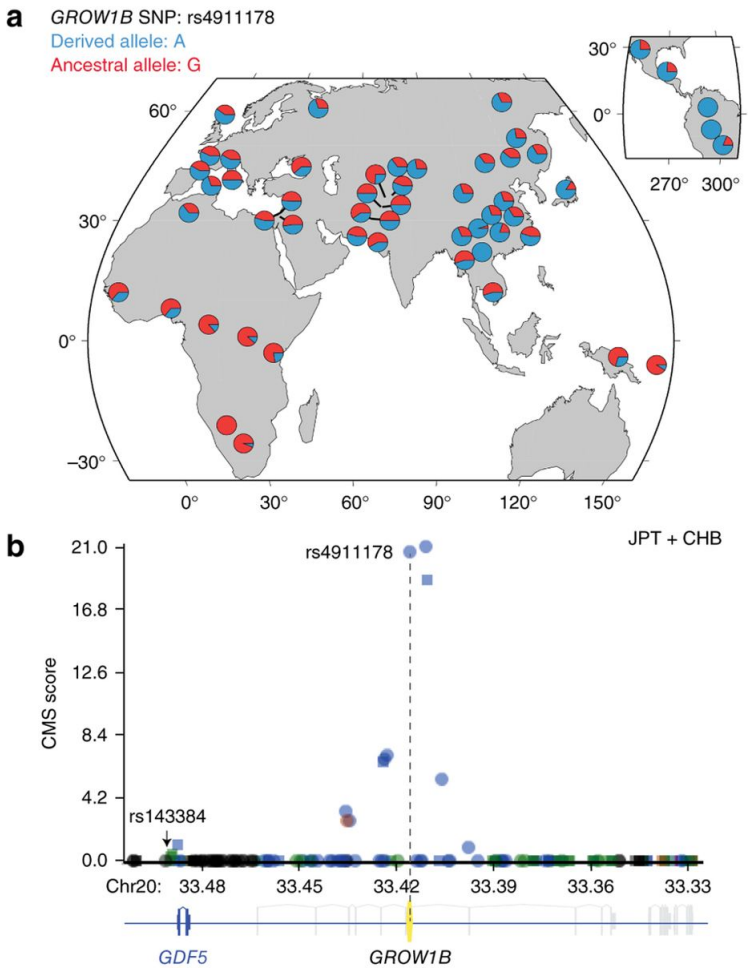


Figure 1.2: Reproduction of Figure 5 from Capellini et al. (2017). Original caption: (a) The derived A (blue) allele at *rs4911178* is present at high frequencies in many out-of-Africa populations, as compared with the ancestral G (red) allele (geographic distribution from Human Genome Diversity Project Selection Browser (Pickrell et al., 2009). (b) CMS (Grossman et al., 2013) selection score in Japanese in Tokyo (JPT) and Han Chinese in Bejing (CHB) populations. The composite of multiple signals of positive selection is centred on the *rs4911178*/*GROW1B* region.

1.2.1 Basic properties of the coalescent

The coalescent is a "backward in time" process that follows n genetic samples as they stochastically merge, or coalesce, $n-1$ times until the most recent common ancestor (MRCA) of the entire sample is found. In each coalescence event, two lineages merge into a common, ancestral lineage, ensuring that the resulting tree is binary.

The coalescent describes both the waiting times, T_i, while i lineages persist in the genealogy and the branching order of ancestral lineages, referred to as the

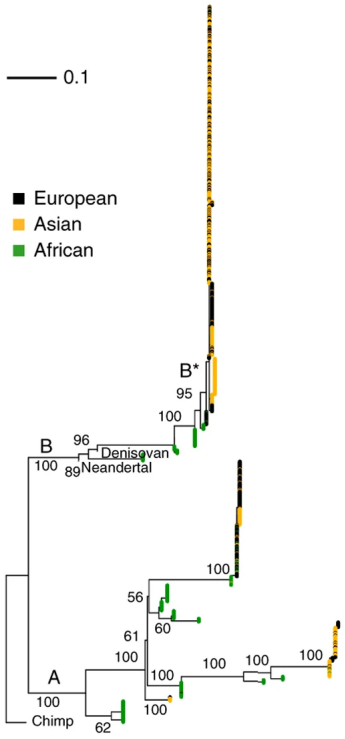


Figure 1.3: Reproduction of Figure 6 from Capellini et al. (2017). Adapted caption: a) Maximum-likelihood tree of phased 1000 Genomes, Neandertal, Denisovan, and chimpanzee (chimp) haplotypes. Two clades (A and B) with strong bootstrap support exist, both containing haplotypes from Europe (black), Asia (orange), and Africa (green). Within clade B, the high-frequency haplotype B* occurs in Eurasian individuals. Within clade B, Neandertal and Denisovan haplotypes are most related to haplotypes only in Africa, and are ancestral to the "short height" B* haplotype found in Eurasians.

tree topology. Figure 1.4 shows one possible genealogy of a sample of $n = 5$ lineages back in time to their MRCA. Under the coalescent, any pair of lineages is equally likely to coalesce. This means that although sample lineages 4 and 5 were observed to coalesce first in this genealogy, it would be just as probable to observe a coalescence between any other pair. The times $T_2, T_3 ... T_i$ show the waiting times while i lineages remain in the genealogy.

The "standard" coalescent assumes a constant-size, panmictic population where each sampled sequence is equally likely to be the parent of a sequence in the next generation. This means that neither genetic variation nor the labelling of tips affects the chance of two lineages coalescing backwards in time or, viewed forwards in time, a sequence's expected number of offspring. Thus, when mutations are neutral, the

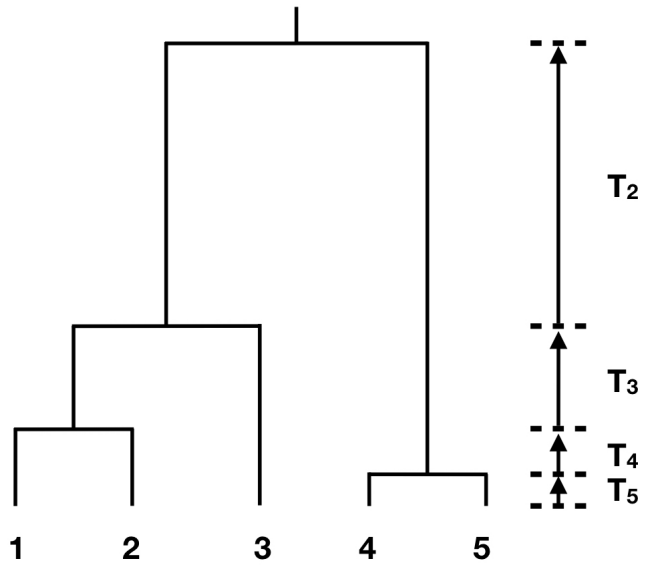


Figure 1.4: One possible coalescent tree with five sampled lineages. The set of times $T_2, T_3 ... T_i$ indicate the period of time where i lineages persist in the genealogy.

mutational and coalescent processes can be treated independently (see Section 1.2.3).

1.2.2 Coalescent waiting times

The distribution of coalescent waiting times is a foundational theoretical result for the work presented in this thesis. Specifically, the prior of `tsdate`, our novel algorithm for inferring the age of ancestral haplotypes presented in Chapter 4, is ultimately based on this result. Here, I provide a brief derivation following Hein, Schierup, and Wiuf (2005).

Consider a sample of 2 sequences at a single locus, drawn from a larger, but still finite, population with a constant size of $2N$. In this formulation $2N$ refers to the number of sequence copies, so in the diploid case there are N individuals each carrying two copies of the sequence. In the haploid case, there are $2N$ individuals each carrying a single copy. Note that the effective population size N_e, or the inverse of the per generation coalescence rate, is used as a parameter in the methods introduced in this thesis. Departure from the assumptions underlying idealised populations, namely panmixia and constant population size, will cause N_e to be smaller than N. This is observed in human populations, where the estimated

N_e of approximately 10,000 is significantly lower than the census population size of humans (Takahata, 1993).

Returning to our example of sampling two sequences from a population, any one of the $2N$ sequences in the population could be the parent of one of the two sequences in the previous generation. Analogous to the well-known "ball into bins problem," the chance of both sequences picking the same parent from the previous generation is thus $1/(2N)$. If two samples pick the same parent, the two sample lineages are said to coalesce. Equally, there is a $1 - 1/(2N)$ chance that the lineages do not coalesce. The waiting time T_2 for the $i = 2$ sequences to coalesce in t generations can thus be described by the geometric distribution,

$$\mathbb{P}(T_2 = t) = \left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N}, \tag{1.1}$$

where the first term is the probability of the two individuals picking different parents for $t - 1$ generations while the second term is the probability of successfully picking the same parent. The expected number of generations until two lineages coalesce is $2N$, equal to the size of the population from which they are drawn.

These results generalise to an arbitrary number of samples in the n coalescent. First consider $G_{i,i}$, the probability that i samples have i ancestors in the previous generation. This can equivalently be thought of as the chance that each sample has a different parent using Equation (1.1):

$$G_{i,i} = \left(\frac{2N-1}{2N}\right)\left(\frac{2N-2}{2N}\right)...\left(\frac{2N-i+1}{2N}\right), \tag{1.2}$$

which can be simplified to

$$G_{i,i} = 1 - \sum_{k=1}^{i-1} \frac{k}{2N} + \mathcal{O}\left(\frac{1}{N^2}\right) = 1 - \binom{i}{2}\frac{1}{2N} + \mathcal{O}\left(\frac{1}{N^2}\right). \tag{1.3}$$

The binomial coefficient reflects that there are now $\binom{i}{2}$ possible pairs of samples, rather than only one pair as in the previous section, each of which is equally likely to coalesce. $\mathcal{O}(1/N^2)$ refers to all terms of order $1/N^2$ or smaller. To safely ignore the $\mathcal{O}(1/N^2)$ term, it is assumed that the sample is much smaller than the population from which it was drawn (without replacement). Wakeley (2009) shows that a

reasonable threshold is sample sizes which are less than 1% of N. Briefly, the reason for this assumption is to ensure that the probability of more than one coalescence event occurring in a generation is negligible.

Thus, only coalescent events where the number of lineages decreases from i to $i-1$ need be considered. From Equation (1.3) the probability $G_{i,i-1}$ is written as

$$G_{i,i-1} = 1 - G_{i,i} = \binom{i}{2}\frac{1}{2N}, \tag{1.4}$$

and from Equations (1.3) and (1.4), the probability that coalescence among a sample of i lineages occurred t generations ago is described by the geometric distribution:

$$\mathbb{P}(T_i = t) \approx \left(1 - \binom{i}{2}\frac{1}{2N}\right)^{t-1}\binom{i}{2}\frac{1}{2N}. \tag{1.5}$$

This geometric distribution can be approximated by the exponential distribution if the population size is large and, as a result, the coalescence rate is small. In this approximation, time is measured continuously by rescaling the time in generations, t, by $2N$,

$$\tau = \frac{t}{2N}. \tag{1.6}$$

The reason for the choice of $2N$ is so that time will be measured in units of the expected time to coalescence for a pair of chromosomes (see Equation (1.1)).

In the limit of continuous time, the waiting time to the next coalescence while i lineages persist is now

$$\mathbb{P}(T_i > \tau) = e^{-\binom{i}{2}\tau}. \tag{1.7}$$

Note that up to this point, T_i referred to the time of coalescence in units of generations. For convenience, the same notation is now used in scaled time.

Since the waiting time is exponentially distributed, the expected amount of time while i ancestors remain is

$$\mathbb{E}[T_i] = \frac{2}{i(i-1)}, \tag{1.8}$$

and the variance is

$$\mathrm{Var}[T_i] = \left(\frac{2}{i(i-1)}\right)^2. \tag{1.9}$$

1.2.3 Mutations in a coalescent genealogy

I now consider a probabilistic model of mutations in genealogical trees which helps us interpret variation observed in real data. Patterns of polymorphism give insight into the unobserved genealogy of our sample and are thus a primary tool for the inference methods introduced in this thesis.

A key assumption of the model to be considered is that mutations are neutral and thus do not affect sequences' expected number of offspring. This assumption allows us to superimpose a mutation process on the coalescent model developed in the previous section.

A commonly used mutation model in population genetics is the infinite sites model (Kimura, 1969). This model underlies much of the inference methodology introduced in this thesis, although deviations are permitted in `tsinfer` version 0.2 (see Section 1.5.2). The model stipulates that every mutation in a population occurs at a novel locus. This means that each mutation generates a new variant site and that recurrent mutations are nonexistent. Biologically, this model is appropriate when the rate of mutation per unit of sequence is much smaller than $1/N$ and the number of sites where mutation can occur is large. In such a setting, the chance of a mutation occurring more than once at any particular location is negligible. This assumption is often not met in humans, where data from the 1000 Genomes Project phase 3 indicate that $\sim 2.3\%$ of sites are multiallelic (Consortium, 2015). The infinite sites model is clearly unsuitable in organisms such as *Anopheles gambiae*, where the *Anopheles gambiae* 1000 Genomes project phase 3 finds that 38% of sites are multiallelic (MalariaGEN, 2021).

Mutations occur on lineages with probability μ, defined as the rate of mutation per generation, per unit of sequence length. When a mutation occurs on a particular lineage, sampled sequences inheriting that mutation carry an allele with the derived state at that variant site, conventionally indicated by "1", while other samples carry the ancestral state, indicated by "0". When time is rescaled in units of $2N$ generations, as in Equation (1.6), the number of mutations occurring on a branch in a single coalescent time unit is $2N\mu$. The parameter θ is used to refer to this "scaled

mutation rate" (or "population mutation rate", since $2N\mu$ is also the expected number of mutations occuring in the population each generation), where θ is defined as $4N\mu$. The additional factor of two is included since θ is also defined as the expected number of mutations in a sample of two. It follows that mutations occur on a single branch in the genealogy with rate $\theta/2$ per coalescent time unit.

This means that the number of mutations, k, on a branch of length τ coalescent time units is Poisson distributed with parameter $\frac{\theta\tau}{2}$:

$$\mathbb{P}(K = k|\tau) = \frac{\left(\frac{\theta\tau}{2}\right)^k}{k!} e^{-\frac{\theta\tau}{2}}. \tag{1.10}$$

Both the expectation and variance of the number of mutations are $\frac{\theta\tau}{2}$. Equation (1.10) is directly used in the mutation-based likelihood function of `tsdate`, the novel algorithm for inferring the age of common ancestors presented in Chapter 4.

1.2.4 The coalescent with recombination

The theory presented so far is applicable to non-recombining sequences where a single tree encapsulates both the ancestral relationships and variation of the sample. In recombining sequences, such as human autosomes, these results only apply to regions where recombination has not played a role in the ancestral history of the sample. However, over longer stretches of sequence, and in a larger samples, it becomes increasingly likely that an ancestral recombination event occurred and a single tree no longer suffices.

Here, I briefly consider two models of recombination in the coalescent. The first, provided by Hudson (1983a), considers recombination backwards in time. This model underlies the coalescent simulations used in this thesis. The second, due to Wiuf and Hein (1999), examines recombination by moving spatially along sequences. This spatial model describes the effects of recombination in tree sequences (see Section 1.3).

In Hudson's algorithm, recombination is modelled in a manner analogous to that of coalescence and mutation. Consider a population with $2N$ gene copies. Equation (1.7) shows that, backwards in time, the waiting time to coalescence is

exponentially distributed with parameter $\binom{i}{2}$. Equation (1.10) shows the number of mutations is Poisson distributed, so the waiting time to the next mutation is exponential with parameter $\theta/2$. Recombination can be thought of as a similar process, where the rate of recombination per locus per generation, r, is small and population size is large. The expected number of recombination events in a coalescent time unit is $2Nr$, and thus ρ is used as the "scaled recombination rate". ρ is defined as $4Nr$ (where a factor of two is again introduced as ρ is equal to twice the expected number of recombination events in a sample of size 2). ρ can be substituted for θ to determine the exponential waiting time to a recombination event as described in Section 1.2.3, and to determine the expected number of recombinations, as in Equation (1.10). When a recombination does occur in this model, the recombined lineage will jump with equal probability to any other lineage in the tree. An advantage of this approach is that it allows recombination to be easily incorporated into coalescent simulations.

Wiuf and Hein (1999) recast the coalescent with recombination as a spatial model along sequences. This approach is more complicated than Hudson's, but is simplified via an approximation known as the sequentially Markovian coalescent (SMC) (McVean and Cardin, 2005a). The following is a brief description of how recombination can be understood spatially under the SMC approximation.

First simulate a marginal tree at the leftmost point in the sequence. Moving to the right, the distance to the break point of the first recombination event is exponentially distributed with parameter $T_0\rho/2$, where T_0 is the sum of the branch lengths of the first marginal tree. The position of the recombination event on the first tree is drawn uniformly, and the branch above the randomly selected position is erased. The pruned, "floating" lineage then coalesces back onto the tree according to the standard coalescent. Note that the coalescence may also occur beyond the time to most recent common ancestor (TMRCA) of the current tree. This process, known as a subtree-prune-and-regraft (SPR), creates a new marginal tree which is highly correlated to the first, as visualised in Figure 1.5. Continuing along the sequence, SPRs corresponding to historical recombination events continue to

create new, but highly correlated, marginal trees until the entire sequence has been traversed. The distinction between the SMC and the coalescent with recombination is that the branch above the recombining lineage is deleted, resulting in a Markovian approximation with substantially improved memory requirements.

The SMC′ is a modified version of the SMC where the branch above the recombining lineage is deleted *after* the floating lineage coalesces back onto the tree (Marjoram and Wall, 2006). As a result, the lineage may recombine onto the branch above the recombination point, forming a "diamond" (Rasmussen et al., 2014). The SMC′ is also Markovian, and offers a slightly better approximation to the coalescent with recombination while retaining the efficiency of the SMC. This model of recombination is key to my explanation of tree sequences in the next section (though the tree sequence structure is not itself dependent on the SMC or SMC′).

1.3 Ancestral recombination graphs and tree sequences

The ARG is a DAG which can be used to represent genetic genealogies in the presence of recombination (Hudson, 1990; Griffiths and Marjoram, 1996; Griffiths, 1997). It encodes a complete record of the history of a sample, including coalescence, mutation, and recombination. Backwards in time, coalescence events are recorded as lineages merging together, while recombination events are lineages splitting in two. The ARG contains marginal trees at every locus in the sequence under consideration. These can be found by tracing the appropriate paths to the left and right of recombination breakpoints in the ARG (see example in Figure 1.6a).

The tree sequence is another structure describing gene genealogies with recombination that has strong links to the ARG. The guiding principle of this structure is that the genealogy of a sample that has experienced recombination can be represented by a sequence of marginal trees along the genome, one for each segment with a unique ancestral history (Hein, 1990). However, a tree sequence is not simply a succession of unrelated trees, as adjacent trees are highly correlated to one another. Indeed, the structure is equivalent to the ARG save for the important

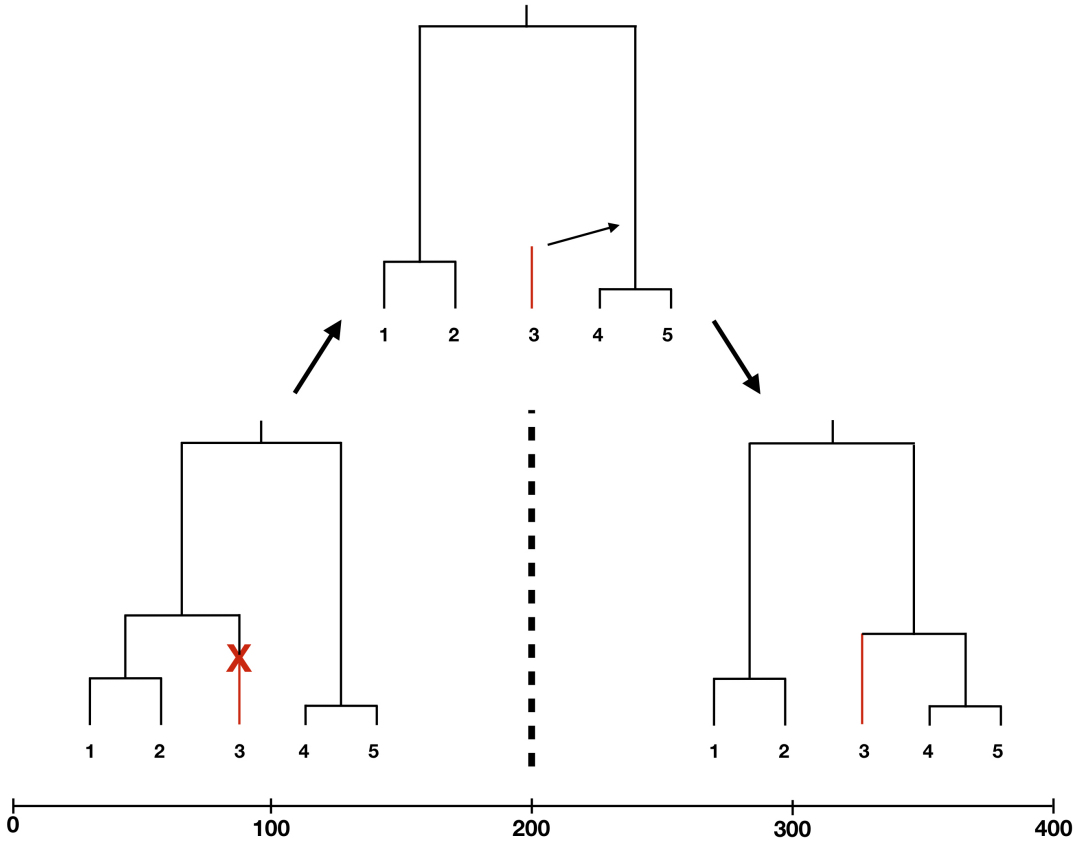


Figure 1.5: Recombination as a subtree-prune-and-regraft (SPR) operation when moving spatially along a tree sequence. In this example, an ancestor of sample three in the marginal genealogy shown on the left undergoes recombination. The subtree containing these samples is pruned from this marginal genealogy and reattached to the edge connecting the MRCA of samples four and five and the MRCA of the entire sample.

distinction that, in its most basic form, it records the effects of recombination, not the (unknowable) positions of recombination nodes in the trees (as shown in Figure 1.6). However, a tree sequence may losslessly represent an ARG if recombination nodes are stored. This is possible when the tree sequence is generated by simulation. A tree sequence without recombination nodes is a more concise representation of the knowable ancestry of the sample it describes, but unlike an ARG, it is not fully explained by parameters.

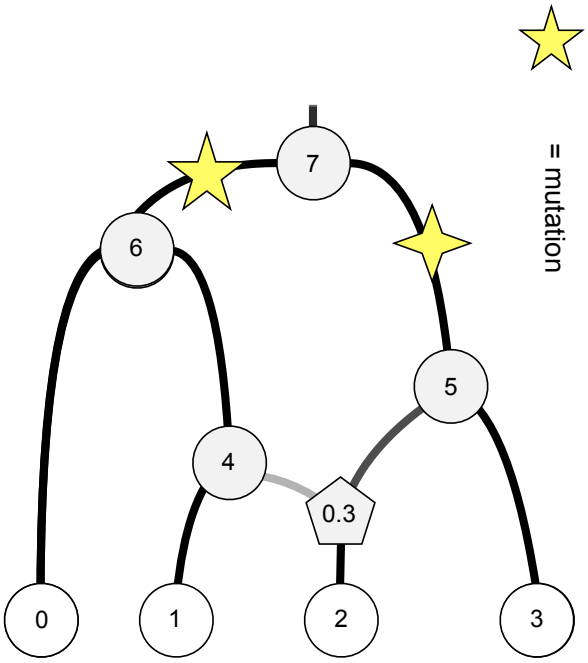
The tree sequence is comprised of nodes, edges, sites, and mutations. Nodes exist at a particular time and refer to ancestral genomes. Edges define the parent-child relationships between nodes over a defined span of sequence. Sites correspond to positions in the genome with variable allelic states. Finally, mutations occur above a

node at a site and are the source of derived alleles. As adjacent trees in a genealogical network are separated by one or more SPRs (see Section 1.2.4), only changes to edges induced by recombination are noted when moving along the sequence.

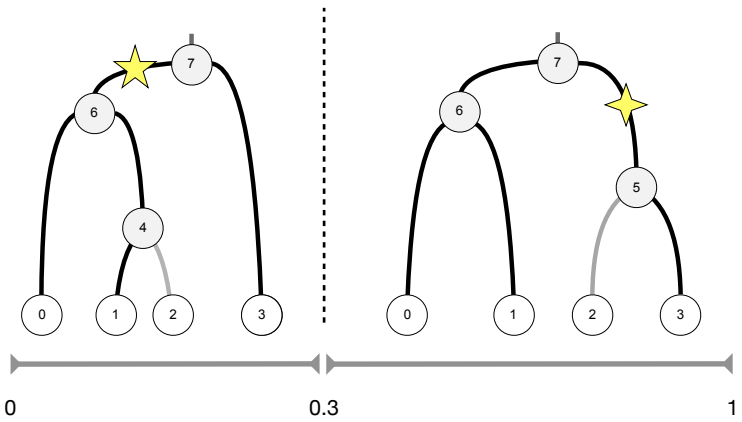
It is important to note that some recombination events are "invisible" in an ARG or tree sequence. The recombination event in Figure 1.6 changes both the topology and branch lengths (discounting unary nodes) across the breakpoint. However, any SPR which prunes a branch and regrafts to the same branch would neither change the topology nor the branch lengths (creating a diamond, see Section 1.2.4). Other recombinations which prune a branch and regraft it to either the parent or child branch of the parent node will change branch lengths, but will not change the topology. These types of recombinations complicate the recovery of accurate recombination rates when using inferred ARGs and tree sequences (Deng, Song, and Nielsen, 2020). I return to this point in Chapter 4 when discussing the challenges of incorporating a recombination-based likelihood in `tsdate`.

One particularly beneficial property of the tree sequence is its space efficiency. While the tree sequence losslessly encodes variation in the samples it describes, it does so much more compactly than the standard text file format used to store gene variant information: the variant call format (VCF). A VCF file uses a genotype matrix with m rows of variant sites and n columns of samples, storing variant data in $\mathcal{O}(m \times n)$. This is less space-efficient than the tree sequence because this matrix fails to capture the correlated nature of ancestry along the sequence. This can intuitively be grasped by considering that a new row is required for each additional variant site, regardless of the frequency of the derived allele. In the tree sequence however, once the set of nodes and edges is stored, mutations can be added in constant space. The resulting storage savings are substantial (Kelleher et al., 2019).

Another advantage of the tree sequence format is that it provides a natural structure for the efficient re-implementation of population genetic statistics. Ralph, Thornton, and Kelleher (2020) show that familiar population genetic statistics calculated using genotype matrices can be redefined on the branches of a tree sequence. For example, the allele frequency spectrum (AFS) is classically found



(a) Ancestral recombination graph



(b) Tree sequence

Figure 1.6: Two representations of the genealogy of four samples (nodes 0-3). Both the ARG and tree sequence contain four common ancestors (nodes 4-7), two mutations, and one recombination event. In 1.6a, a recombination node at position 0.3 is indicated with a pentagon. To obtain the marginal tree from positions 0-0.3 follow the left hand edge above the recombination node, to obtain the marginal tree from positions 0.3-1, follow the right hand edge. In 1.6b, the marginal trees themselves are shown, with the dotted line indicating the location of a recombination event. Transparency of an edge's shading is proportional to its span. Only coalescent nodes are shown in the tree sequence representation.

by computing the frequency of each derived allele and accumulating the number of alleles at each frequency. However, the branches of a tree sequence can be used to compute an "expected" AFS as shown in Figure 1.7, since the distribution of neutral mutations is a function of branch length and span (Equation (1.10)).

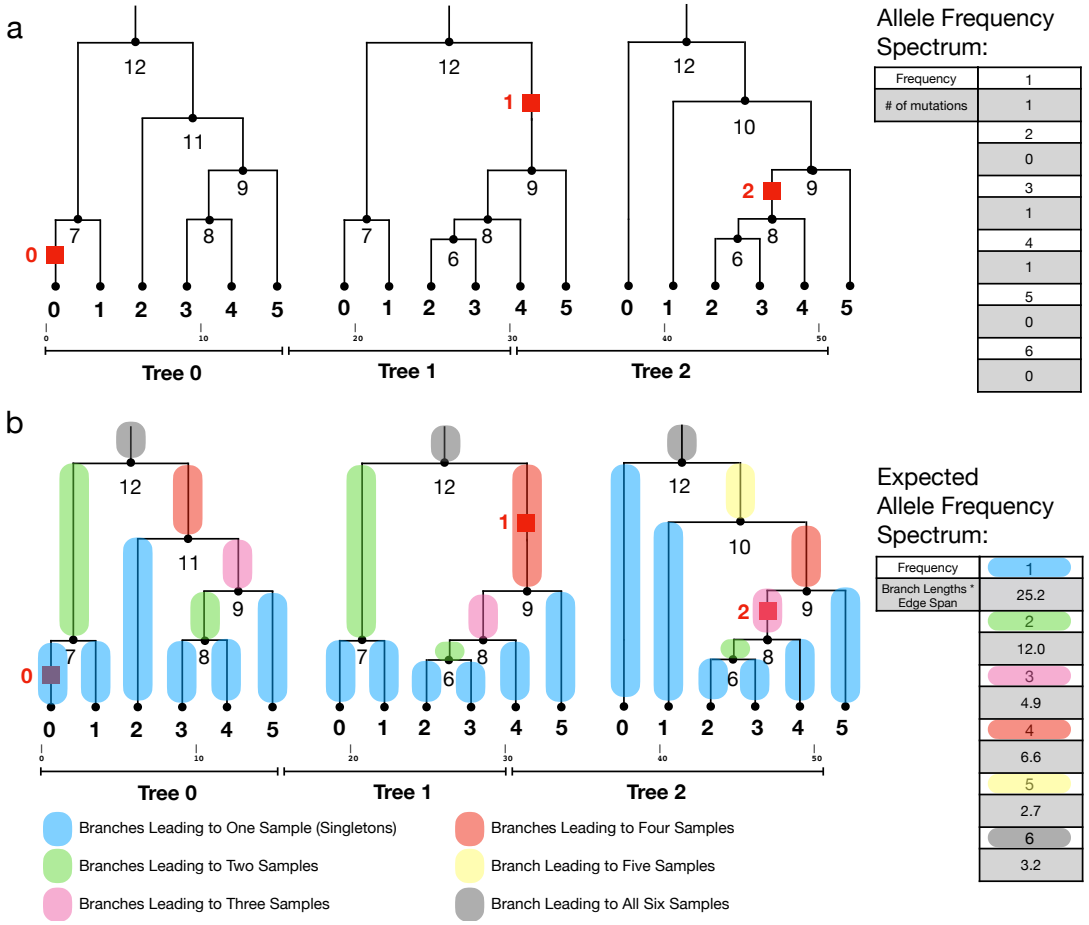


Figure 1.7: (a) "Site" and (b) "branch" definitions of the allele frequency spectrum using a tree sequence. In (a), the frequencies of the three mutations are tabulated and the number of mutations falling into each frequency "bin" are shown in the table. In (b), branches are labelled by the number of samples (tips) descending from them. The expected allele frequency spectrum is then computed as the branch lengths in each "bin" multiplied by the edge span of the branches. The exact values given in the table in (b) are only illustrative.

Tree sequences produced by software including `msprime`, `SLiM`, `tsinfer`, and `Relate` may be manipulated and analysed using `tskit`, the tree sequence toolkit. This robust, community-maintained software ecosystem also includes efficient implementations of a number of population genetic summary statistics, as described

in the previous paragraph. The novel statistics introduced in Chapter 3 are also implemented in `tskit`. The documentation for this software can be found at : `https://tskit.readthedocs.io/en/latest/`.

1.4 Allele age estimation

Now that I have introduced some probabilistic models underlying genealogies and the structures used to represent them, I consider the inference of an important property of these genealogies: how long ago a mutation created a derived allele, here referred to allele age.

The evolutionary history of osteoarthritis risk alleles in *GDF5* provides a illustrative example of the significance of allele age. Capellini et al. (2017) suggest three possible explanations for the heterogeneous pattern of allele sharing between Africans, Eurasians, and archaic humans at the *GDF5* locus. These are: (1) recurrent mutation, (2) archaic introgression, or (3) selection on standing variation in the ancestral population of modern and archaic humans (see Figure 1.8). While detecting recurrent mutations may require another approach, distinguishing between the latter two possibilities would be immensely aided by an accurate estimation of the age of the focal variant.

As shown in Figure 1.9, if all extant sequences carrying the allele descend from a single chromosome hosting the causative mutation, the time since the mutation (T_m in Figure 1.9) can be distinguished from the TMRCA of all lineages carrying that allele (T_{ca}). T_m will necessarily be older than T_{ca}.

According to Slatkin (2000), methods of estimating allele age are based either on the frequency of the allele in a population or on the sequence differences between various allele copies, which he refers to as intra-allelic variability. The former approach is based on the approximation that, assuming neutrality, an older allele will usually be at a higher frequency than a younger allele. The latter uses information resulting from the processes of mutation and recombination.

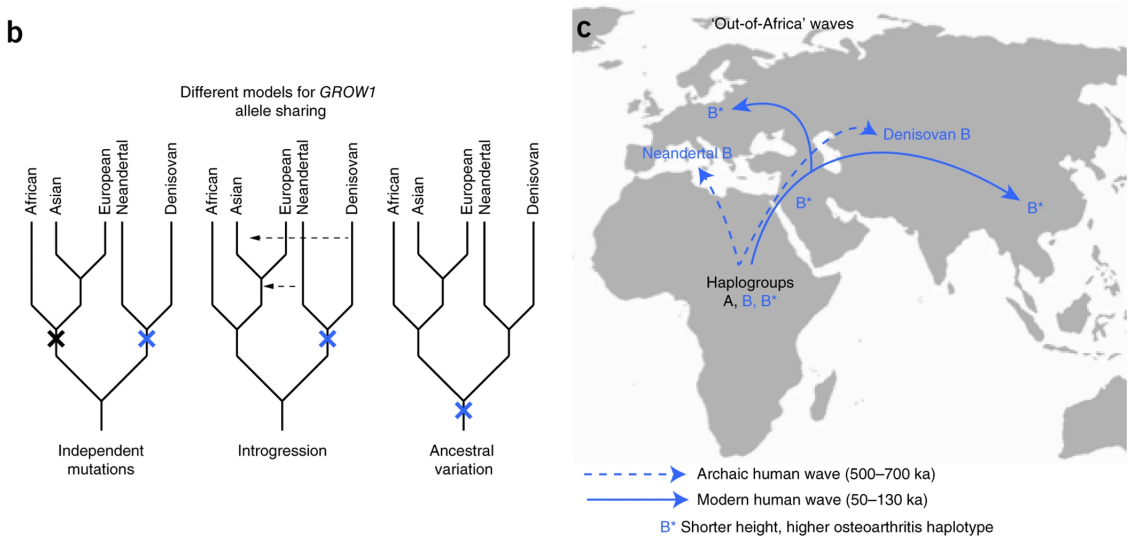


Figure 1.8: Reproduction of Figure 6 (b and c) from Capellini et al. (2017). Adapted caption: (b) Three models of haplotype evolution at the *GDF5* locus. Blue X, shared mutations; black X, independent mutations. (c) Multiple *GDF5* haplotypes were present in Africa in the middle to late Pleistocene, including haplotype B, containing the derived lower-activity growth variant. Neandertal and Denisovan ancestors carried this or a related shorter-height haplotype into northern latitudes (dashed lines). During the late Pleistocene, haplotypes (B*) related to (B) arose and were positively selected during more recent out-of-Africa migration(s). The B* haplotype leads to decreased height via the derived *rs4911178* variant in *GROW1B*.

1.4.1 Frequency-based estimators of allele age

The relationship of allele frequency to allele age was first explored by Kimura and Ohta (1973). Using diffusion equations, the authors found that the expected age of a neutral allele in a large, panmictic population with a constant N is

$$\mathbb{E}[T_m] = \frac{-2p}{1-p}\ln(p), \tag{1.11}$$

where p refers to allele frequency and time has been rescaled in units of $2N$ generations.

The confidence interval of the age of an allele based on its frequency can be derived from the probability distribution of allele age (Slatkin and Rannala, 2000). This distribution, which has a complex mathematical form, is due to Watterson (1976) and Griffiths and Tavaré (1998).

These results offer frequency-based allele age estimates for neutral mutations in a constant-size population. For a review of frequency-based estimators of allele age

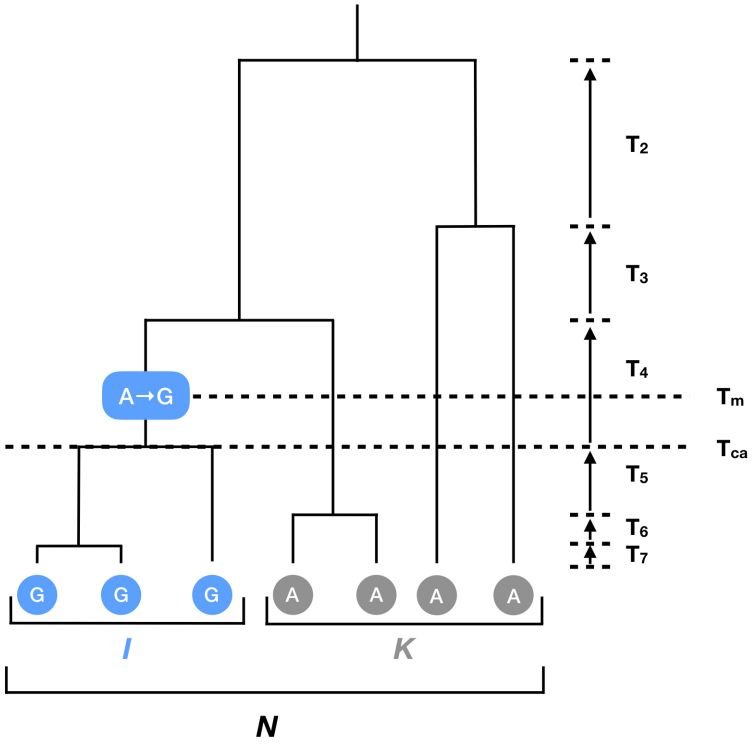


Figure 1.9: A biallelic locus with allele age (T_m) and age of the MRCA of samples with the resulting allele (T_{ca}). The total number of samples is $n = 7$. A subset of samples, here referred to as I, carries the derived allele while the remaining $N - I$ samples are in set K.

in populations that have experienced past population growth or natural selection, see Slatkin (2000). The impact of these factors on the relationship of allele frequency to allele age can be significant. For instance, according to a non-parametric allele age estimation method introduced by Albers and McVean (2020) and discussed in Section 1.4.2, variants found only in African and admixed American populations had a contradictory relationship between frequency and age. Specifically, variants at high frequency in American populations were on average younger than low frequency variants when considering those which were found only in these two continents. The authors argue that these variants are likely of recent African origin and have risen to higher frequency in American populations following admixture with African groups and population bottlenecks.

1.4.2 Intra-allelic variability-based estimators of allele age

The limitations of frequency-based estimators of allele age has led to a comparatively greater interest in methods based on intra-allelic variability, which rely on comparing sequences around the focal mutation of interest (Slatkin, 2000).

The first use of intra-allelic variability in estimating the age of an allele was in Serre et al. (1990). The authors developed a moment estimator based on the frequency of marker alleles and the recombination rate, a method that was also subsequently used by Risch et al. (1995). Slatkin and Rannala (1997) later derived a maximum likelihood estimator of allele age based on an approximation of intra-allelic coalescence times. More recent methods focus on estimating the age of beneficial alleles (Ormond et al., 2015; Nakagome et al., 2016; Smith et al., 2018). Innovations in aDNA technology have also made possible the joint estimation of allele age and selection coefficients from time series data (Malaspinas et al., 2012). `ARGWeaver` and `Relate`, methods discussed in Section 1.5, produce estimates of allele age using intra-allelic variability-based methods.

A recently introduced method for allele age estimation based on intra-allelic variability is Genealogical Estimation of Variant Age (`GEVA`) (Albers and McVean, 2020). This probabilistic algorithm estimates the absolute age of a variant based on statistical inference of the TMRCAs between pairs of haplotypes, which are then combined to obtain a posterior distribution on allele age.

The `GEVA` algorithm assumes that sampled sequences surrounding a variant of interest can be divided into the subset I of samples that carry the derived allele and the subset K that carry the ancestral allele, as shown in Figure 1.9. The samples in set I are assumed to coalesce with one another before any coalescence with a sample from set K. The coalescence at the MRCA of the samples in I with the first ancestral lineage of samples in K is informally defined as the upper bound on T_m, while T_{ca} is the lower bound. Similarly, a pair of sample lineages taken from I (called concordant pairs) will coalesce before a pair where one sample lineage is from I and one from K (discordant pairs).

The intra-allelic variability observed in samples at a locus provides a means of estimating the age of alleles. All samples in I have inherited some part of the ancestral haplotype on which the mutation occurred. This can be thought of as the shared haplotype segment that has been inherited from the MRCA of a given pair. However, each sample's copy of the ancestral haplotype will have both accumulated mutations and been broken down by recombination. `GEVA` uses a Hidden Markov model (HMM) to determine the length of the shared haplotype. Then a probabilistic model based on the length of shared sequence (i.e. a "recombination clock") and number of segregating sites (i.e. a "mutation clock") between concordant pairs of haplotypes gives posterior distributions for T_{ca}, while the discordant pairs give the age of the coalescence of samples carrying the derived allele with the rest of the sample. T_m is then described by a composite posterior distribution of the lower and upper bounds. Note that `GEVA` assumes a constant effective population size.

1.5 Genome-wide genealogical inference

In the previous section I surveyed methodology for inferring one property of a genealogy: the age of alleles. I now consider the larger problem of inferring the genealogies themselves. As we will see in Chapter 4, these inferred genealogies will also provide a platform for inferring allele age.

Consider again the motivating example of *GDF5* from Section 1.1.1. Sections 1.2.4 and 1.3 make clear that representing the ancestral history of a sample with a single tree can only be accomplished in the absence of recombination. Since TGP haplotypes indicate that recombination has occurred at the locus of interest, the authors built their maximum likelihood phylogenetic tree using a 130 kb haplotype block they defined based on linkage disequilibrium (LD) between variants in the region (see Figure 1.3). The authors also used a heuristic algorithm to remove 395 low frequency haplotypes that had likely experienced recombination before inferring the maximum-likelihood tree. While limiting the region under consideration and removing haplotypes attenuates the effect of recombination, these

measures discard information that is potentially useful for genealogical inference and potentially introduce biases.

An ARG or tree sequence built over the entire chromosome containing *GDF5* would include multiple marginal trees in the 130 kb target region, reflecting the recombination events that created the low frequency haplotypes. Furthermore, given the correlated nature of genealogies described in Section 1.3, the sequences flanking this 130 kb region would provide additional insight into the ancestral history of *GDF5*. An inferred ARG would thus serve as a more accurate depiction of the evolutionary history of this region of interest.

ARG inference is, however, a very computationally demanding problem. Inferring the "min-ARG," which is guaranteed to have the minimum number of recombination events necessary to describe observed variation, has been shown to be an NP-hard problem (Wang, Zhang, and Zhang, 2001; Bordewich and Semple, 2005). A larger number of methods seek to infer ARGs without this guarantee. Some of these algorithms are based on parsimony (Gusfield, 2005; Song and Hein, 2005; Minichiello and Durbin, 2006; Wu, 2008; Melé et al., 2010) or other heuristics (Griffiths and Marjoram, 1996; Song, Wu, and Gusfield, 2005; Parida et al., 2008; O'Fallon, 2013; Mirzaei and Wu, 2017). However, most of these methods are unfeasible beyond approximately 100 samples or 1000 sites. I will now briefly describe three methods for inferring ARGs, or similar genealogical structures, as well some of their applications.

`Margarita` is a heuristic method for inferring ARGs which supports unphased and missing data and scales to datasets of thousands of samples over hundreds of SNPs (Minichiello and Durbin, 2006). The method does not sample ARGs from the coalescent with recombination, but rather relies on a set of heuristics to determine the set of coalescence, recombination, and mutation events in the ancestral history of the sample. The method is not deterministic, as randomness is introduced when selecting the order of coalescence and mutation events which are inferred to be contemporaneous as well as in choosing recombination breakpoints. The authors' main application of genealogies inferred using `Margarita` was in fine-mapping disease loci. Their approach was based on the intuition that individuals with a

disease-causing variant will appear to cluster beneath branches hosting a causal mutation. The authors thus perform χ^2 tests at marginal trees to determine if there is an association between allelic state and disease status. They note that an advantage of this method is that the correlation between marginal trees in their genealogy only depends on the effects of recombination, unlike the LD matrix used to control for the non-independence of loci in single-marker tests.

`ARGweaver` explicitly samples ARGs from an approximation of the Sequentially Markovian Coalescent (McVean and Cardin, 2005a) using Markov Chain Monte Carlo (MCMC) (Rasmussen et al., 2014). This method is considered to be the state of the art in ARG inference; however, is is computationally intensive, and only scales to a few tens of sequences (Kelleher et al., 2019; Hubisz and Siepel, 2020). Rasmussen et al. (2014) used inferred ARGs from `ARGweaver` to explore patterns of natural selection in the human genome. The authors find that areas of the genome with particularly ancient TMRCAs show patterns consistent with balancing selection and that allele age is depressed at sites thought to be under directional selection.

`Relate` is an efficient algorithm for inferring genome-wide genealogies from thousands of samples (Speidel et al., 2019). The method utilises an HMM based on the Li and Stephens (LS) model (Li and Stephens, 2003) and a hierarchical clustering algorithm to infer marginal trees from the inferred distance matrix between sampled sequences. Branch lengths are then inferred using MCMC with a coalescent prior. While the inferred marginal trees capture local relationships, ancestral haplotypes (nodes persisting across marginal trees) are not explicitly inferred. Instead, `Relate` uses heuristics to identify branches which persist across trees in order to recover some of the correlation between marginal trees encoded in ancestral recombination graphs and tree sequences.

Genome-wide genealogies constructed using `Relate` have been used to study population relationships, effective population size and mutation rates over time, and the effects of directional selection. One study has used trees inferred by `Relate` as the basis of a full-likelihood method for estimating the strength of directional selection on polygenic traits (Stern et al., 2021). Additionally, a recent extension

of `Relate` supports the inclusion of high-coverage ancient samples in inferred genealogies and uses relationships of low-coverage samples through time to estimate coalescence rates (Speidel et al., 2021).

1.5.1 Inferring tree sequence topologies with `tsinfer`

The `tsinfer` software package implements a deterministic algorithm for tree sequence inference (Kelleher et al., 2019). Since this algorithm is fundamental to work presented in the subsequent chapters, a brief explanation of the initial implementation follows. I also discuss an updated version of `tsinfer` that supports missing data and inexact matching in the copying process.

The central premise of `tsinfer` is that a sequence of DNA sampled from a population can be thought of as a collection of segments inherited from ancestral haplotypes. If the ancestral haplotypes and lineages traversed by these segments of genetic material back in time were known, the tree sequence could be reconstructed. There are thus two steps to the algorithm. First, we estimate the ancestral haplotypes upon which derived alleles arose. Second, the copying paths from ancestors to older ancestors and from samples to ancestors are inferred using a HMM based on the LS model.

`tsinfer` Step One: Estimating Ancestral Haplotypes

The first step in the algorithm is to create a set of ancestral haplotypes describing a given genotype matrix. The oldest ancestral haplotype carries the ancestral state at each variant. All other ancestral haplotypes are built using biallelic, non-singleton variant sites with known ancestral state. Other sites are still represented in the output tree sequence but are not used to build inferred ancestral haplotypes. Each derived allele is assumed to be the result of a unique mutation which occurred on an ancestral haplotype at some point in the past (i.e. the infinite sites assumption; see Section 1.2.3). All sampled sequences carrying the derived allele must descend from the original chromosome hosting the mutation. All sequences in I, the set of

samples with the derived allele at the focal variant, are thus assumed to form a clade in the marginal genealogy describing the focal variant site (see Figure 1.9).

Once each derived allele is assigned to a unique ancestor, the state of the ancestral haplotype surrounding the focal variant can be inferred, as shown in Figure 1.10. The ancestral haplotype is extended to either side of the focal variant until it reaches another variant site. If the derived allele at this adjacent variant is at a higher frequency than the focal variant, the algorithm computes whether the majority of samples in I contain the derived allele at the adjacent site. If so, the derived allele is included in the growing ancestral haplotype. If the adjacent, higher frequency variant exhibits an allele sharing pattern that is incompatible with the assumption that all samples in I form a clade, a recombination event is invoked and conflicting samples are removed from I. For example, if one sample in I carries a derived allele and the others do not, the sample is removed from I. If, on the other hand, the derived allele at this adjacent variant is at a lower frequency than the focal variant, the algorithm appends the *ancestral* state of the adjacent variant to the developing ancestral haplotype. The ancestral haplotype generation process is terminated when fewer than half of the original samples in I remain. A key approximation used by `tsinfer` is that frequency is an accurate local proxy of relative allele age, a heuristic that is critical both for constructing valid ancestral haplotypes and for finding pathways of descent (described below). The validity of this heuristic is the subject of the next chapter.

`tsinfer` Step Two: Inferring copying paths between haplotypes

Once ancestral haplotypes have been estimated, the most likely path of sampled genetic material through the ancestors is determined using the LS model (Li and Stephens, 2003). The intuition behind this approach is that both samples and ancestral haplotypes are imperfect copies of their ancestors. In the `tsinfer` algorithm, the LS model is used to determine the likelihoods of the possible "copying" paths (or pathways of descent) of a sample through its possible ancestors. The most likely path is determined by the Viterbi algorithm (Forney, 1973).

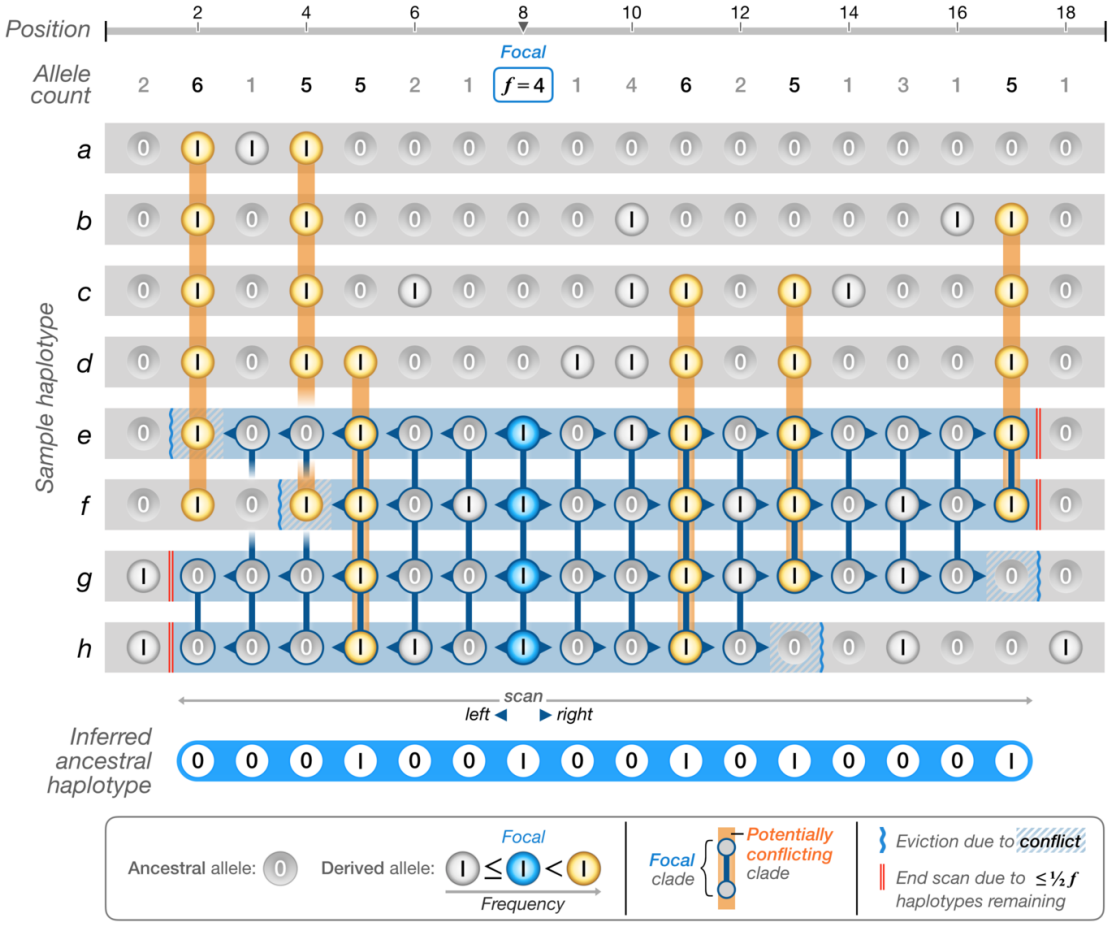


Figure 1.10: Reproduction of Figure 2 from Kelleher et al. (2019). Original caption: Schematic of ancestral haplotype reconstruction. We are constructing an ancestral haplotype for the ancestor corresponding to site 8. The time of each site is approximated by its corresponding frequency value (referred to as allele count here). Note that this is a simplified schematic: the actual implementation includes a dampening effect such that haplotypes are evicted from S if they disagree with the consensus call at two adjacent sites.

The matching process in `tsinfer` occurs in two steps: "match-ancestors" and "match-samples". In the former, inferred ancestral haplotypes are matched against older haplotypes. As mentioned in the previous section, the oldest ancestral haplotype extends for the entire sequence length and carries the ancestral state at all positions. Matching thus begins with the next oldest inferred ancestral haplotype. Note that the relative age of ancestors is again determined by the frequency of the "focal" derived alleles used to build the ancestral haplotypes. Moving from left to right across ancestral haplotypes, the likelihood that each of these ancestors

descends (or copies) from older ancestors existing at the same genomic location is calculated. Once all likelihoods are computed, the Viterbi algorithm is used to find the most likely paths at each location. These paths become the relevant edges in the tree sequence. Inference then proceeds on ancestors in the next youngest epoch. Note that the algorithm can be executed in parallel for multiple ancestors existing in the same epoch. When the copying paths of all ancestors has been found in the "match-ancestors" phase, copying paths from samples to ancestors can be determined in the "match-samples" phase.

The accuracy of tree sequence reconstruction with `tsinfer` is comparable to the state of the art method, `ARGweaver`. Furthermore, due to the use of simple heuristics and the inherent efficiencies of the tree sequence data structure, `tsinfer` scales to the largest datasets currently available (as shown by inference of the UK Biobank tree sequence in Chapter 3).

1.5.2 `tsinfer` version 0.2

In the previous section I described the initial implementation of the `tsinfer` algorithm for inferring tree sequence topologies (Kelleher et al., 2019). This version of the algorithm was used to infer the tree sequences analysed in Chapter 3. The latest version of the algorithm, `tsinfer` version 0.2, incorporates two new features: provision for inexact matching in the copying process and support for missing data (Wohns et al., 2021). All inferred tree sequences in Chapters 5 and 6 were built using this new version of `tsinfer`.

Mismatch, error, and recurrent mutation

As described in Section 1.5.1, in the second step of `tsinfer` an HMM based on the Li and Stephens (LS) copying model (Li and Stephens, 2003) is employed from left to right along each haplotype to infer the mosaic of closest-matching haplotype segments among the array of older haplotypes. In the initial implementation of `tsinfer`, we supported only exact haplotype matching – if a haplotype matched perfectly against an ancestor up to a certain position, but then a mismatch occurred,

it could only be explained by switching to a different ancestor via recombination. The new algorithm now supports the full LS model including a mismatch term that allows for inexact matching, where mismatches between a target haplotype and its inferred ancestor are explained by additional mutations. The relative probabilities of recombination versus mismatch are tuned via a "mismatch ratio" parameter: high ratios lead to fewer inferred recombination events and more additional mutations, while low ratios result in more recombination events and fewer additional mutations. A more extensive description of the full Li and Stephens model can be found in Donnelly and Leslie (2010) and Lunter (2018).

Probabilities of recombination between adjacent sites can be provided in the form of a genetic map. However, optimal mismatch probabilities will depend on factors such as sequencing error and variation in the rates of mutation and gene conversion along the genome. To establish suitable mismatch ratios, in Wohns et al. (2021) we evaluated inference performance on both simulated and real data. In all cases, good results are obtained by mismatch ratios close to unity, where the probability of a mismatch is set equal to the median probability of recombination between adjacent inference sites. A mismatch ratio of 1 was therefore used in Wohns et al. (2021) and all analyses using `tsinfer` version 0.2 in this thesis (the results shown in Chapters 5 and 6).

Missing data

Missing data is accommodated in `tsinfer` version 0.2 by using older ancestors as a "reference panel" for imputation. In the core `tsinfer` HMM, samples and ancestors copy from older ancestors. If the sample or ancestor contains missing data at a site, the missing genotypes are imputed from the most recent ancestor without missing data. The approach provides a principled approach to imputing missing data for both contemporary and ancient genomes, as only older ancestral haplotypes are used.

1.6 Ancient DNA

The inference methods introduced in the previous sections exclusively use genetic data from modern samples. However, in the past two decades the sequencing of genetic material from organisms who lived many thousands of years ago has revolutionised our understanding of the evolution of humans and other species (Botigué et al., 2017; Gutaker and Burbano, 2017; Reich, 2018). The number of ancient human samples processed has increased from a trickle of mitochondrial genomes in the late 1990s (Krings et al., 1997; Krings et al., 1999) to the current norm of whole-genome sequencing or genotyping datasets from hundreds of samples (Narasimhan et al., 2019; Margaryan et al., 2020). Ancient DNA (aDNA) has elucidated mysterious periods and features of human history, not least the relationships between modern and archaic humans. The groundbreaking finding that humans and Neanderthals interbred was provided by autosomal DNA from a Neanderthal individual (Green et al., 2010). Perhaps even more surprising was the revelation of a hitherto unknown lineage of hominins, the Denisovans, which are related to Neanderthals (Meyer et al., 2012). Knowledge of their existence is attributable to the sequencing of a finger bone fragment found in the Altai Mountains of Russia.

The motivating example of *GDF5* demonstrates how aDNA can aid understanding of the evolutionary history of phenotypically significant loci. As noted in sections 1.1.1 and 1.4, Capellini et al. (2017) included Neanderthal and Denisovan genomes in their analysis of the *GDF5* regulatory locus. Notably, these archaic sequences carry the derived variant at *rs4911178* that is common in Eurasians and comparatively less common in Africans. As described in Section 1.4, three scenarios can explain these findings: (1) recurrent mutation, (2) introgression from archaic individuals, or (3) selection on standing variation. Recurrent mutation is ruled out by the authors due to identical patterns of variation in the surrounding region. However, comparative analysis of archaic and modern haplotypes can provide evidence for scenarios (2) vs (3). After examining the data, the authors suggest scenario (3) is more plausible than introgression as they find the archaic haplotype at this locus is relatively distinct from the haplotype observed in Eurasians. Since this variant appears to be under

selection, the authors did not consider a fourth possible scenario: that the mutation occurred in the ancestral population of both modern and archaic humans, but that genetic drift rather than selection is responsible for observed patterns of variation.

However, many other human phenotypes are now thought to be influenced by archaic introgression. These include adaptations for life at high altitude in Tibetan individuals (Huerta-Sánchez et al., 2014), skin and hair development in Eurasians (Sankararaman et al., 2014), and immune function in Europeans (Quach et al., 2016). Recent work shows that the major risk factor for severe COVID-19 disease originates from a 50 kb haplotype that appears to have been introgressed from Neanderthals (Zeberg and Pääbo, 2020). The authors of this study employed similar methods to Capellini et al. (2017), including building a phylogeny of modern humans and Neanderthals at the locus and examining derived allele frequencies worldwide.

1.6.1 Population genetic inference using ancient DNA

Scanning the genome for evidence of introgression from archaic individuals is just one of the many methods that have been used to gain insight from ancient genomes. All methods seeking to use aDNA must contend with the small average size of aDNA sequencing reads (less than 500bp) (Pääbo et al., 2004), characteristic error patterns (Hofreiter et al., 2001), contamination from other sources of DNA (Gilbert et al., 2005), and the difficulties of phasing and imputing individuals. The latter issue is due in part to a lack of satisfactory reference panels for many ancient samples (Hui et al., 2020).

One of the most commonly used tools for understanding the patterns of variation in ancient genomes is principal component analysis (PCA), a dimensionality reduction technique that provides a non-parametric summary of variation data (Menozzi, Piazza, and Cavalli-Sforza, 1978; McVean, 2009). Ancient samples are often projected onto axes built from modern data (Patterson, Price, and Reich, 2006; François and Jay, 2020). Other commonly used statistics include measures of treeness such as f-statistics (Reich et al., 2009; Patterson et al., 2012) and models that assign genetic material to its supposed population of origin (Pritchard,

Stephens, and Donnelly, 2000; Pickrell and Pritchard, 2012). Efforts have also been made to utilise the time-series nature of aDNA for inferring selection (Loog et al., 2017; Mathieson, 2020).

Recent work has sought to infer ARGs from modern and ancient samples using an updated version of the `ARGweaver` algorithm called `ARGweaver-D` (Hubisz, Williams, and Siepel, 2020). This approach samples ARGs instead of providing a deterministic estimate of gene trees (as in `Relate`) or the tree sequence (as in `tsinfer`). Another advantage of `ARGweaver-D` is that unphased genomes are able to be used in this algorithm using a method that samples phased states. The algorithm is highly parametric, as ARGs are sampled conditional on a user-specified demographic model relating populations to one another. In the authors' application, this involved specifying the timing of interbreeding between human ancestors and archaic individuals, as well as conditioning on a lack of gene flow from archaics to the ancestors of Africans.

A recent extension to the `Relate` algorithm supports the inference of genome-wide genealogies from high-coverage ancient data (Speidel et al., 2021). The authors also provide a new method called "Colate", which infers coalescence rates between low-coverage, unphased, and unimputed ancient genomes.

1.7 Thesis overview

In the following chapters, I introduce, validate, and apply novel methods for tree sequence inference using modern and ancient samples. In Chapter 2, I examine the problem of establishing the correct relative ordering of alleles. This is essential for `tsinfer`, the most scalable method for inferring tree sequence topologies. In Chapter 3, I show how tree sequences inferred by `tsinfer` can be used to understand genetic relationships in population-sequencing datasets, with particular emphasis on descent from ancient samples. In Chapter 4, I introduce `tsdate`, an approximate Bayesian method for inferring the age of ancestral haplotypes conditional on a tree sequence topology. I show how this method can be used in conjunction with `tsinfer` to infer genealogies combining modern and ancient individuals. These approaches

are validated in Chapter 5 using simulations and empirical comparisons with ancient genomes. In Chapter 6, I use these methods to infer the most comprehensive representation of human ancestry to date using eight modern and ancient datasets. Examining this structure provides a fine-scale understanding of worldwide genetic variation and is suggestive of the extent of sequencing error in widely-used datasets. I also provide a simple estimator of the spatiotemporal properties of inferred ancestors that recapitulates key events in hominin evolutionary history. In Chapter 7, I return to the example of *GDF5*, showing that the methods I have introduced can provide new insights at this important locus. I also discuss further improvements and extensions to these inference methods as well as prospects for future tree sequence-based analyses of population sequencing data.

2

The Relative Age of Alleles

Contents

2.1 Introduction 35
2.2 The importance of relative allele age 37
2.2.1 Relative allele age in `tsinfer` 39
2.3 Probability of allele frequency reflecting relative age . 40
2.3.1 Conditional genealogies and allele age 40
2.3.2 Allele frequency and relative age at unlinked loci 42
2.4 Simulating the pairwise time-ordering of alleles 45
2.4.1 Empirically determined genotype error profile 46
2.4.2 Comparison to `GEVA` 47
2.4.3 "Out of Africa" model 48
2.4.4 Ancestrally connected pairs of mutations 48
2.5 Simulation results 48
2.6 Discussion 54

2.1 Introduction

While multiple methods have been introduced to infer the absolute age of alleles (Kimura and Ohta, 1973; Rasmussen et al., 2014; Mathieson and McVean, 2014; Palamara et al., 2018; Speidel et al., 2019; Albers and McVean, 2020), inferring the *relative* age of alleles is a largely unexplored problem in population genetics. This problem has relevance for a key heuristic used by `tsinfer`: that allele frequency

accurately reflects allele age. If this relationship generally holds, then knowledge of the site frequency spectrum would be sufficient to determine the relative age of mutations. In this chapter, I investigate the validity of this heuristic by determining the accuracy of allele frequency in reflecting the relative age of pairs of alleles.

Section 1.4 describes various methods for inferring allele age, of which allele frequency is the simplest. While estimating the relative age of alleles using their frequency is trivial, "misorderings" may occur at pairs of alleles appearing in the same marginal tree (Figure 2.1) or pairs appearing in different marginal trees in a tree sequence (Figure 2.2). This chapter examines how often these "misorderings" occur.

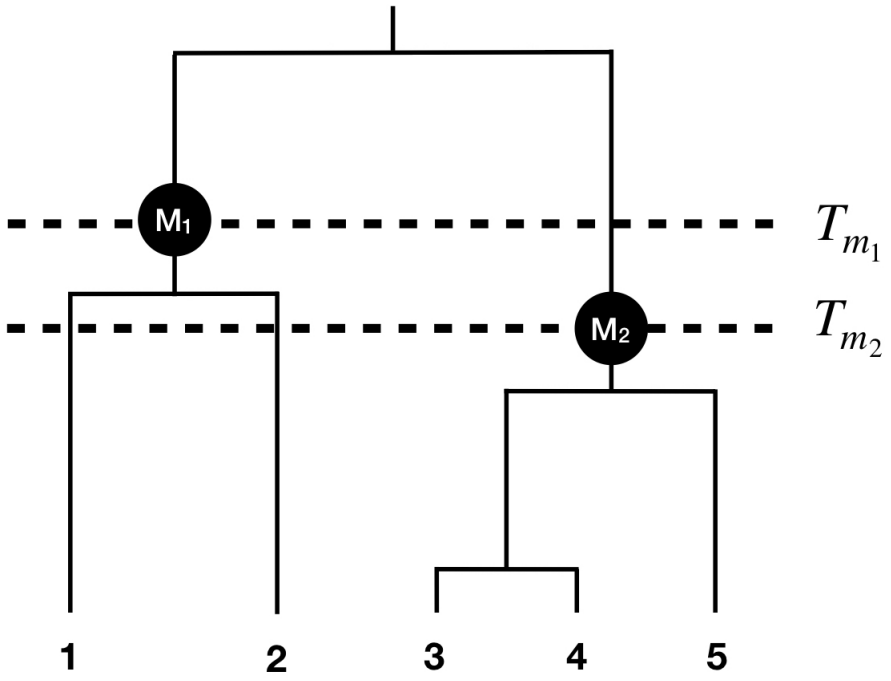


Figure 2.1: Frequency incorrectly reflects the relative age of alleles in a genealogical tree. Mutation 1 is at frequency 2/5 while mutation 2 is at frequency 3/5. However, $T_{m_1} > T_{m_2}$.

In the following sections, I first motivate this work by examining the significance of relative allele age to problems in evolutionary biology, oncology, and population genetics. I then describe how allele frequency is used as a proxy for relative allele age in `tsinfer`. Next, I derive a novel recursion equation for the expected accuracy of allele frequency in reflecting the relative age of alleles at unlinked loci. `msprime` simulations are then used to evaluate this proxy measure of relative allele age when

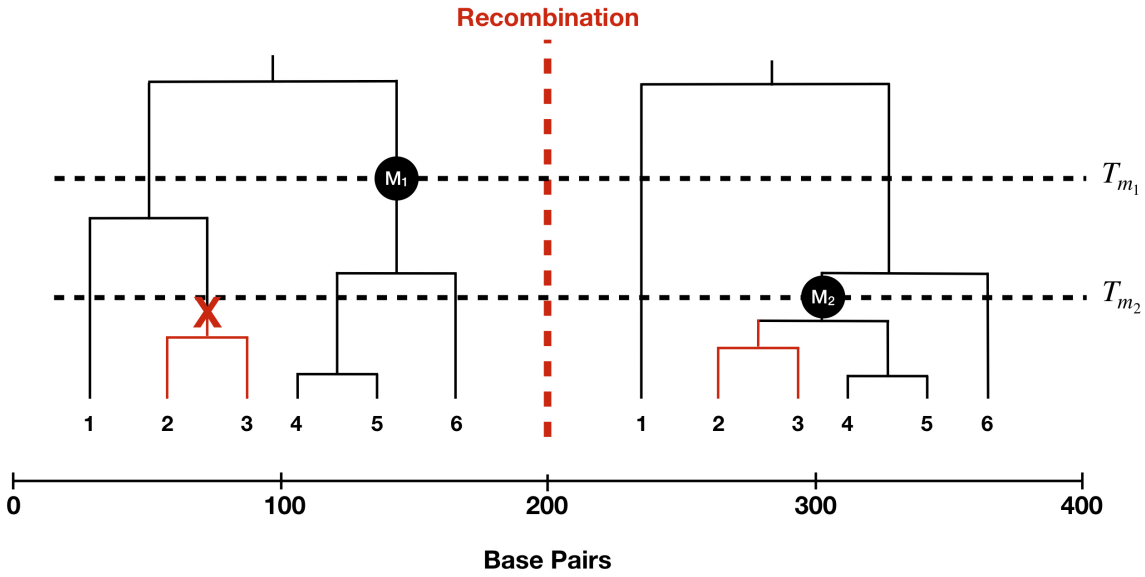


Figure 2.2: Frequency incorrectly reflects the relative age of a pair of alleles in adjacent marginal genealogies in a tree sequence. While $T_{m_1} > T_{m_2}$, mutation 1 is at frequency 1/2 while mutation 2 is at frequency 2/3. Note that these mutations are "ancestrally connected" (see Section 2.4.4).

comparing all pairs of alleles and the subset of "ancestrally connected" allele pairs where the haplotype on which one allele arose is the ancestor of the haplotype on which the other allele arose. I also consider the effects of genotype error and a demographic model approximating recent human evolutionary history. These results are compared to the accuracy of the allele age estimation method `GEVA` in determining relative allele age.

I produced a version of Figure 2.6 as Supplementary Figure 1 of Kelleher et al. (2019). My other contributions to this publication are included in Chapter 3.

2.2 The importance of relative allele age

Inferring the relative age of alleles has relevance to a variety of biological problems. Here, I briefly consider three topics where mutation order has important ramifications: "mutation-order" speciation, somatic mutation order in oncogenesis, and LD patterns resulting from a selective sweep. Note that these examples describe situations that are not well modelled by the neutral coalescent and are thus somewhat tangential to the novel results in this chapter. I then explain the role

of relative allele age in the `tsinfer` algorithm, a setting which directly motivates the work presented in the next section.

Two main models of speciation have been proposed: "ecological" and "mutation-order" speciation. In ecological speciation, different environments provide different evolutionary pressures, driving localised adaptation and reproductive isolation (Schluter, 2009). In "mutation-order" speciation, incompatible alleles arise in response to similar evolutionary pressures and achieve fixations in different populations (Mani and Clarke, 1990). The circumstances under which the latter occurs are complex, as the influence of drift, the selective advantage of alleles, gene flow, and the degree of incompatibility between alleles all play a role. It has been argued that the order in which mutations arise is an important factor as well (Mani and Clarke, 1990). Nosil and Flaxman (2011) show that if two incompatible mutations with different selective advantages arise in populations with gene flow between them, speciation is more likely to occur if the less advantageous mutation occurs slightly earlier. Inference of the relative ordering of incompatible alleles could thus provide insight into the dynamics of ongoing divergence under this model.

The time order of somatic mutations has been shown to play an important role in tumour progression (Kent and Green, 2017; Levine, Jenkins, and Copeland, 2019). The first evidence of this process was provided by time series tumour samples which revealed a stepwise accumulation of mutations in human colorectal carcinomas (Nigro et al., 1989; Baker et al., 1990a; Baker et al., 1990b). These studies showed that the development of these carcinomas is usually initiated by a "gatekeeping" mutation in the gene *APC*, followed by mutations in other genes. The discovery that oncogenesis is driven by the emergence of cellular clones which outcompete their neighbours and is thus governed by evolutionary principles provides a coherent framework for explaining these findings (Gerlinger et al., 2012; Levine, Jenkins, and Copeland, 2019). In this evolutionary context, "gatekeeper" mutations can play an important role in influencing the fitness of future clones and the emergence of subsequent mutations. This principle is conceptually similar to arguments for the

role of epistasis in speciation (Schluter, 2009), whereby early mutations influence which subsequent mutations are favoured.

The order in which alleles occur also has important consequences for LD between neutral variants following strong selective sweeps. McVean (2007) showed that LD between variants which occur on the same side of the sweep and are closely linked to the selected variant is highly dependant on the age of the neutral variants relative to the onset of sweep. If the variants predate the sweep, LD between them is typically very strong. However, variants younger than the sweep typically are in weak or nonexistent LD with one another. Thus, if the relative age of alleles were known, patterns of LD could be examined conditional on their age relationships to the selective sweep and could potentially aid detection of selection. Furthermore, drawing inspiration from the previous two examples, examining epistatic effects between time-ordered mutations is potentially an intriguing extension to studies of selection.

2.2.1 Relative allele age in `tsinfer`

Relative allele age is used in two steps of the `tsinfer` algorithm: in the estimation of ancestral haplotypes and in ordering these ancestral haplotypes in the model's core Li and Stephens HMM.

As described in the previous chapter and illustrated by Figure 1.10, the relative age of alleles is a key component of the ancestral haplotype estimation algorithm used in `tsinfer`. Note that the following assumes an infinite sites model, where each derived allele is the result of a single ancestral mutation. Consider once again the set of samples, I carrying the derived allele at a focal site. The ancestral haplotype around the focal site is estimated by examining adjacent variant sites to the left and right where samples in I carry a derived allele. At such an adjacent site, only mutations older than the derived allele at the focal site may have existed on the ancestral haplotype and are thus informative of its state. Younger mutations could not have existed on the haplotype and are disregarded. It is assumed that the majority of samples in I represent the original ancestral haplotype, so if the majority

carry an older derived allele at an adjacent site, the derived allele is added to the ancestral haplotype. Any samples carrying the conflicting allele are ejected from I. In the actual implementation we only eject samples if two consecutive conflicts are observed. The intuition behind this heuristic is that samples providing conflicting information must have recombined away from the focal ancestral haplotype and are no longer informative. Higher frequency mutations are assumed to be older, but if this assumption is broken the algorithm will ignore informative, older variants while utilising non-informative, younger variants.

The second use of relative allele age in `tsinfer` is in the Li and Stephens-based HMM which infers copying paths from ancestors to descendants. Inferring these copying paths requires the correct relative ordering of ancestral haplotypes built in the previous step. Incorrect orderings will lead to topological inaccuracies in the resulting tree sequence which must be resolved by spurious recombination events or, if using `tsinfer` version 0.2, by additional mutations (see Section 1.5.2).

Clearly, an incorrect relative ordering of alleles breaks the assumptions of both the ancestor generation and ancestor matching steps of `tsinfer`, resulting in less accurate tree inference. To understand how often this occurs, I next investigate how well frequency reflects the relative age of pairs of alleles.

2.3 Probability of allele frequency reflecting relative age

Wiuf and Donnelly (1999) derived the first two moments of the age of a neutral mutation under the coalescent, conditional on the number of samples observed to carry the resulting allele (i.e. allele frequency). I briefly present relevant results from their work in Section 2.3.1 before deriving a novel recursion for the probability of allele frequency reflecting the relative age of alleles at unlinked loci in Section 2.3.2.

2.3.1 Conditional genealogies and allele age

Consider Figure 1.9 from the previous chapter, where seven sampled lineages are divided into sets I and K. The samples in set I all carry the derived "G" allele,

while the samples in set K carry the ancestral "A" allele. Assume that the mutation causing the derived allele occurred only once and that the resulting allele is neutral. This description presupposes that the sample lineages in I coalesce before they share a common ancestor with a sample lineage from set K, an event referred to as E. This event is also assumed in `GEVA`, a method for inferring allele age described in Section 1.4.2. The event E depends only on the topological structure of the coalescent genealogy, not the distribution of the waiting times, and thus conditioning on this event only affects the topology.

Conditioning on E, Wiuf and Donnelly (1999) define a Markov chain whose states are (i, k). These states describe the number of sample lineages in two sets, I and K, where the total number of samples is $i + k = n$. Transitions in the Markov chain reflect the jump chain of the coalescent. Specifically, $q_{ik,i'k'}$ is the probability of a transition from the state with i and k lineages to a state with i' and k' lineages. The conditional probability that coalescence occurs within set I is thus

$$q_{ik,(i-1)k} = \frac{i(i-1)}{n(n-1)},$$

and the probability of coalescence within set K is

$$q_{ik,i(k-1)} = \frac{k(k-1)}{n(n-1)}.$$

The probability that a pair of sample lineages from sets I and K coalesce is

$$1 - q_{ik,(i-1)k} - q_{ik,i(k-1)} = \frac{2ik}{n(n-1)}.$$

These probabilities allowed the authors to write a recursion equation describing the transition rate matrix $Q(i, k)$ that gives the probability of event E:

$$(i+k)(i+k-1)Q(i,k) = i(i-1)Q(i-1,k) + k(k-1)Q(i,k-1).$$

The absorbing state of the recursion is $Q(1, k) = 1$. This refers to reaching the MRCA of I ($i = 1$) when $k \geq 1$ (as the MRCA of K cannot yet have coalesced with a lineage from I). With this absorbing state, the recursion can be simplified to

$$Q(i,k) = \frac{2(i-1)!}{(i+1)(k+1)_{i-1}}, \tag{2.1}$$

where $a_x = a(a+1)..(a+x-1)$.

It is important to note that if population size is not constant, conditioning on the event E which creates a clade of size i is *not* equivalent to conditioning on the existence of a mutation with i descendants. The reason for this is that conditioning on the existence of a mutation with i descendants is dependant on branch lengths in the genealogy, while conditioning on a partition of size i is not. The following results concerning the accuracy of allele frequency reflecting allele age are thus approximate when demography is variable. If population sizes vary significantly, the results presented in the following sections may be a poor approximation to the truth. See Wiuf and Donnelly (1999) for further details on the effect of conditioning on the existence of a mutation, as well as the effect of exponentially increasing population size.

2.3.2 Allele frequency and relative age at unlinked loci

Building on these results from Wiuf and Donnelly (1999), I now derive novel analytical expressions for the expected accuracy of frequency reflecting pairwise allele age at unlinked loci. Consider two variants located distantly from one another along a chromosome such that the genealogies at the two sites are independent, i.e. the sites are unlinked. The first site contains i samples with a derived allele and k with the ancestral allele. The second site has j and m samples with derived and ancestral alleles, respectively. This scenario is represented in Figure 2.3. What is the probability that the mutation at the first variant site occurred before the mutation at the second site, conditional on both allele frequencies?

Using the results from Wiuf and Donnelly (1999) described in Section 2.3.1, we condition on events E_1 and E_2 to derive analytical expressions for the Markov chain transition probabilities for coalescent events in subsamples I, K, J, or M. Event E_1 is that the sample lineages in I share a common ancestor before any coalesce with a sample lineage from K while event E_2 is that sample lineages in J share a common ancestor before any coalesce with a sample lineage from M (Equation (2.1)). Conditional on these events, we can determine the probability that $I = 1$

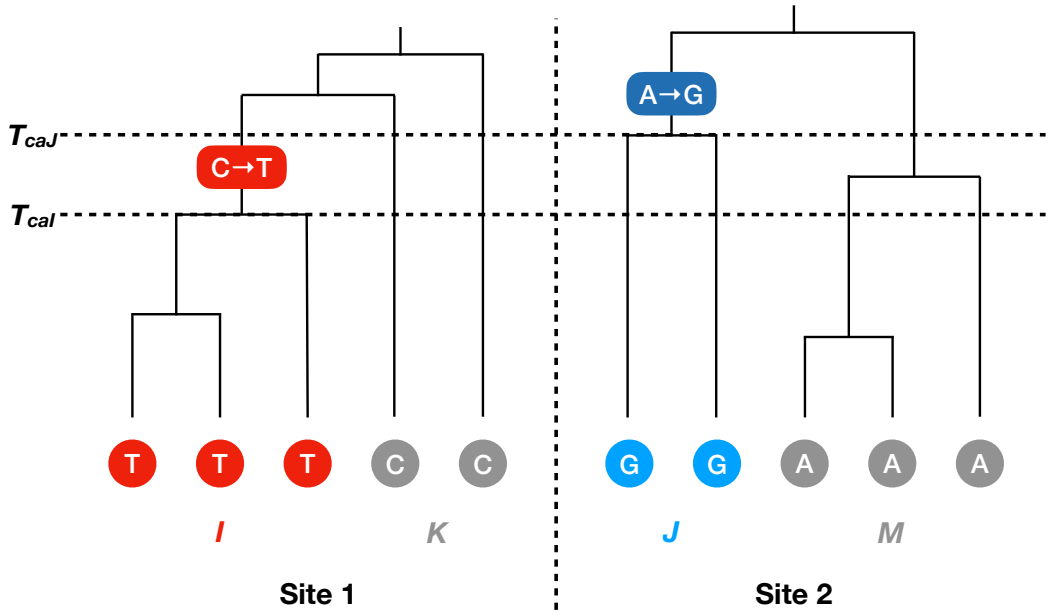


Figure 2.3: Five sampled lineages at two independent loci. Subsamples I and J carry the derived allele at their respective sites. Condition on the events E_1 and E_2 that, at each site, samples carrying the derived allele coalesce before they share a common ancestor with a sample carrying an ancestral allele. At site 1, this means that the samples in I find an MRCA before any ancestor to a sample in I coalesces with an ancestor to a sample in K.

occurs before $J = 1$, or the probability that the common ancestor of the lineages in I is younger than the common ancestor of the lineages in J.

The actual age of the mutation, T_m, is the age of the ancestor in which the mutation occurred. However, here we are interested in the age of the common ancestor of all samples carrying the derived allele, T_{ca}, where $T_{ca} \leq T_m$.

The conditional coalescent process can be formalised as an absorbing Markov Chain with state space $S = (i, k, j, m)$. We seek to find the probability that all samples in I coalesce before the samples in J, or $T_{ca_I} < T_{ca_J}$:

$$\mathbb{P}_{i,k,j,m} = \mathbb{P}(T_{ca_I} < T_{ca_J} | i, k, j, m). \tag{2.2}$$

We also have the following boundary conditions:

$$\begin{aligned} \mathbb{P}_{1,k,j>1,m} &= 1, \\ \mathbb{P}_{i>1,k,1,m} &= 0, \\ \mathbb{P}_{1,k=m,1,m=k} &= 0.5. \end{aligned} \tag{2.3}$$

The transition probabilities $q_{ss'}$ of moving from state s to s' are the product of the probabilities of two independent process. The first is whether the event occurs at site 1 or 2. This is simply the number of possible coalescent events at one site as a proportion of the total number of possible coalescent events at either site. The second process is the probability of each of the two possible events at the chosen site: coalescence within I or K at site 1 or within J or M at site 2.

$$\begin{aligned}
q_{ikjm,(i-1)kjm} &: \frac{\binom{i+k}{2}}{\binom{i+k}{2}+\binom{j+m}{2}}\frac{i+1}{i+k},\\
q_{ikjm,i(k-1)jm} &: \frac{\binom{i+k}{2}}{\binom{i+k}{2}+\binom{j+m}{2}}\frac{k-1}{i+k},\\
q_{ikjm,ik(j-1)m} &: \frac{\binom{j+m}{2}}{\binom{i+k}{2}+\binom{j+m}{2}}\frac{j+1}{m+j},\\
q_{ikjm,ikj(m-1)} &: \frac{\binom{j+m}{2}}{\binom{i+k}{2}+\binom{j+m}{2}}\frac{m-1}{m+j}.
\end{aligned} \tag{2.4}$$

In each term in Equation (2.4), the left side of the product determines whether the event occurs at site 1 or 2 and the right side determines the type of event at the chosen site.

With these transition probabilities and boundary conditions, we can determine $s = \sum_{s'} q_{ss'}\mathbb{P}_{s'}$ for each value of s. I numerically solved this recursion with sample size 50 and integrated over parameters k and m to create Figure 2.4, which shows $\mathbb{E}[\mathbb{P}(T_{ca_I} < T_{ca_J}|i,j)]$ at values of i and j from 1 to 50.

By integrating over all values of i, j, k and m, I determine that the average probability that an allele at lower frequency is younger than an allele at higher frequency is 82.41% at unlinked loci under a standard coalescent model for a sample size of 50. The expected accuracy of frequency is dependant on sample size. For instance, the corresponding value at a sample size of 20 is 80.32%.

The code to produce Figure 2.4 and the values presented here can be found at `https://github.com/awohns/relative-allele-age`.

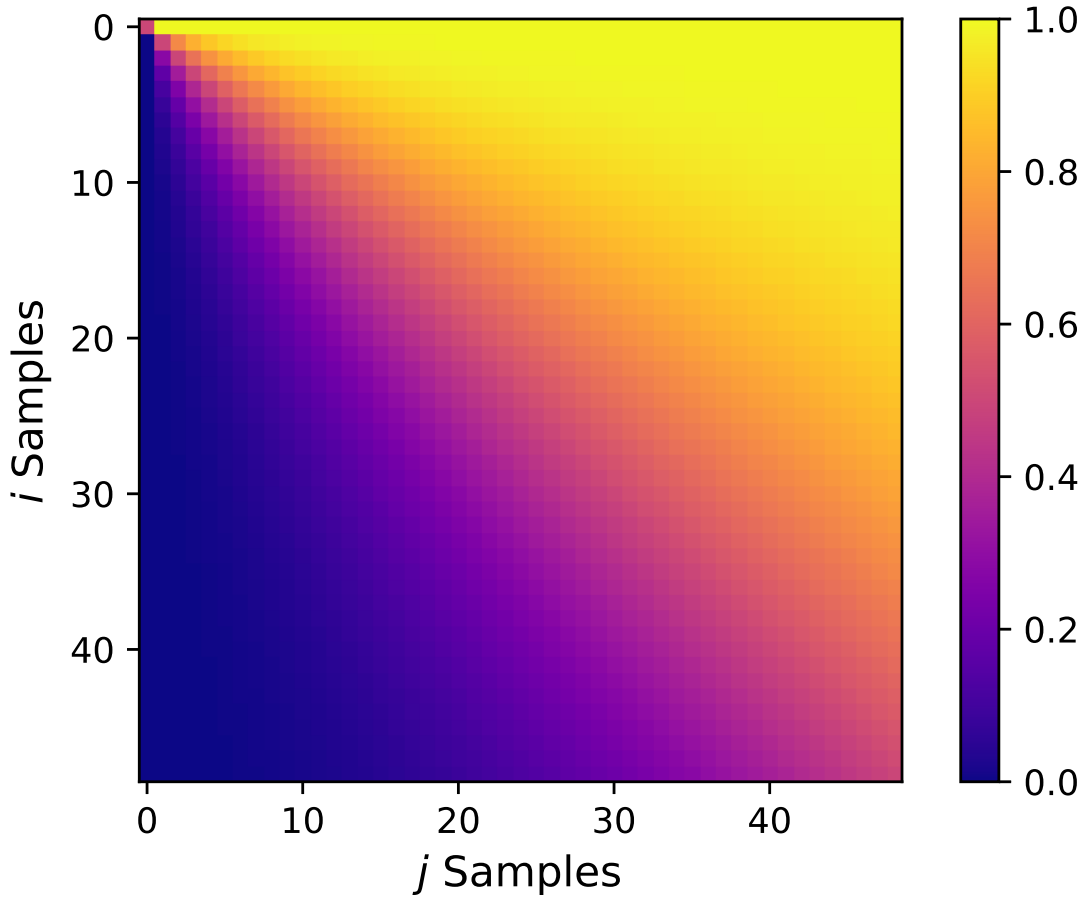


Figure 2.4: Heatmap showing $\mathbb{E}[\mathbb{P}(T_{ca_I} < T_{ca_J}|i,j)]$, the expectation of the event that an allele carried by i samples is younger than an allele carried by j samples with a sample size of 50. Frequency is an increasingly accurate proxy for relative age as the difference between i and j increases.

2.4 Simulating the pairwise time-ordering of alleles

The above described numerical solution to the exact iterative formula gives the expected accuracy of frequency at unlinked loci. Coalescent simulations were conducted to determine the accuracy of frequency reflecting the relative age of alleles under a variety of conditions and parameter schemes. In particular, I explored how accuracy was affected by the physical distance separating alleles, realistic rates of genotype error, a demographic model approximating recent human evolution, and whether allele pairs were "ancestrally connected". The accuracy of `GEVA` in determining relative allele age was also considered.

The coalescent simulator `msprime` (Kelleher, Etheridge, and McVean, 2016) was used with the following parameters: 50 sequences of 100 kb were simulated with $N_e = 10,000$ and $\mu = r = 10^{-8}$. Different demographic parameters were used in the "Out of Africa" simulations detailed in Section 2.4.3. As in the previous section, T_{ca} is used as a proxy for the age of an allele.

Variants were only considered if both the ancestral and derived alleles were carried by more than one individual in the sample. This was motivated by the fact that `tsinfer` does not use singletons to infer ancestral haplotypes and `GEVA` excludes both singletons and variants with frequency $(n-1)$. Furthermore, pairs where both alleles were the same age were excluded from consideration. For all pairs where both variants pass these restrictions, the allele ages were compared to the differences in allele frequency. Frequency correctly predicts relative age if $T(a_1) < T(a_2)$ and $f(a_1) < f(a_2)$, where $T(a_1)$ and $T(a_2)$ are the ages of the derived allele at sites 1 and 2 respectively and $f(a_1)$ and $f(a_2)$ are their current frequencies in the population. For instance, $T(a_1) = 2000; f(a_1) = 0.3$ and $T(a_2) = 1000; f(a_2) = 0.2$, counts as a single correct comparison. Figures 2.1 and 2.2 are examples of comparisons marked as failures. If variants are at the same frequency, the comparison is randomly assigned as a success or failure with equal probability.

2.4.1 Empirically determined genotype error profile

To approximate the error inherent in all genotyping and next-generation sequencing methods, simulated genotypes were resampled using the empirically constructed genotype error profile from Albers and McVean (2020).

Given a biallelic locus with ancestral alleles indicated by 0 and derived alleles by 1, three distinguishable diploid genotypes exist: 00, 01/10, and 11, referred to here as 0, 1, and 2. Genotype error may differentiate the observed genotype, g_i, from the true genotype g_j. The probability of observing g_i given g_j is ϵ_{ij}. For instance, ϵ_{21} is the probability of observing a homozygous derived genotype given that the true genotype is heterozygous. The nine possible values of ϵ_{ij} can be represented in a 3x3 matrix:

$$\epsilon = \begin{bmatrix} \epsilon_{00} & \epsilon_{01} & \epsilon_{02} \\ \epsilon_{10} & \epsilon_{11} & \epsilon_{12} \\ \epsilon_{20} & \epsilon_{21} & \epsilon_{22} \end{bmatrix}$$

The authors determined frequency-specific values of ϵ by comparing low-coverage sequencing data of CEPH pedigree 1463 from the phase 3 release of the 1000 Genomes Project (Consortium, 2015) to corresponding high-coverage genomic data of the same pedigree from the Illumina Platinum Genomes Project (Eberle et al., 2017).

The resulting empirical error profile was used to resample genotypes produced by `msprime` simulations. For each variant in the simulated sample, the frequency of the derived allele was determined and matched to the closest frequency value from the empirical error profile. Individual chromosomes from the simulation were paired to make them pseudo-diploid and the resulting genotypes were resampled with probabilities dependant on the frequency-specific values of ϵ. For instance, a simulated heterozygote would be resampled to a homozygous ancestral variant with probability ϵ_{01}, where the value of ϵ_{01} is dependant on the frequency of the variant.

When genotypes are resampled according to the empirical error profile, the number of eligible variant pairs in each simulation changes. For example, some singletons become doubletons or tripletons and are thus eligible for comparison while other variants become singletons. The following inclusion criteria was used to determine eligible pairs. In the no error case, the accuracy metrics are calculated using eligible pairs determined by the true genotypes; in the error case, the accuracy metrics are calculated using eligible pairs determined by the observed genotypes.

2.4.2 Comparison to `GEVA`

`GEVA`, an allele age estimation method briefly described in Section 1.4.2, was used to estimate the age of simulated variants (Albers and McVean, 2020). `GEVA` was run using the binary downloaded from `https://github.com/pkalbers/geva` using default parameters. When singletons and sites with frequency $(n-1)$ were excluded, `GEVA` provided allele age estimates at 99.4% of the variable sites generated in standard `msprime` simulations. The `GEVA` accuracy statistics reported in Section 2.5 are based on all eligible sites, not only the 99.4% of sites for which estimates were made.

2.4.3 "Out of Africa" model

The only demographic parameter passed to the previously described `msprime` simulations was a constant N_e of 10,000. In order to determine the accuracy of frequency and `GEVA` under a demographic model approximating recent human evolutionary history, simulations were also performed under the Gutenkunst et al. (2009) "Out of Africa" model. This is equivalent to the "OutOfAfrica_3G09" model implemented in `stdpopsim` (Adrion et al., 2020), though this analysis was performed before `stdpopsim` was available. 100 kb of sequence was simulated with 16 chromosomes sampled from the African, European, and Asian populations.

2.4.4 Ancestrally connected pairs of mutations

The previously described simulations compared all variants with more than one ancestral and derived allele. I now consider the subset of mutation pairs which are "ancestrally connected". These are pairs where the haplotype on which one mutation arose is an ancestor of the haplotype on which the other mutation arose (see Figure 2.5).

These pairs are particularly significant for ancestral inference because it is impossible to reverse the temporal order of the child nodes on the edges where these mutations appear while maintaining the true topology of the tree sequence.

2.5 Simulation results

Figure 2.6 shows the accuracy of frequency plotted against the physical distance separating allele pairs. The average accuracy when error is not considered is 83.13%. Accuracy is higher when variants are physically close to one another and decays as the distance between variants grows due to the correlation of adjacent tree topologies. Notably, accuracy in the simulations without error converges to the expected value under a standard neutral coalescent model where alleles are found at independent loci. This value, 82.4%, is reached when alleles are roughly 75 kb apart. With a constant recombination rate of 10^{-8}, 75 kb is equivalent to 0.075 centiMorgans (cM).

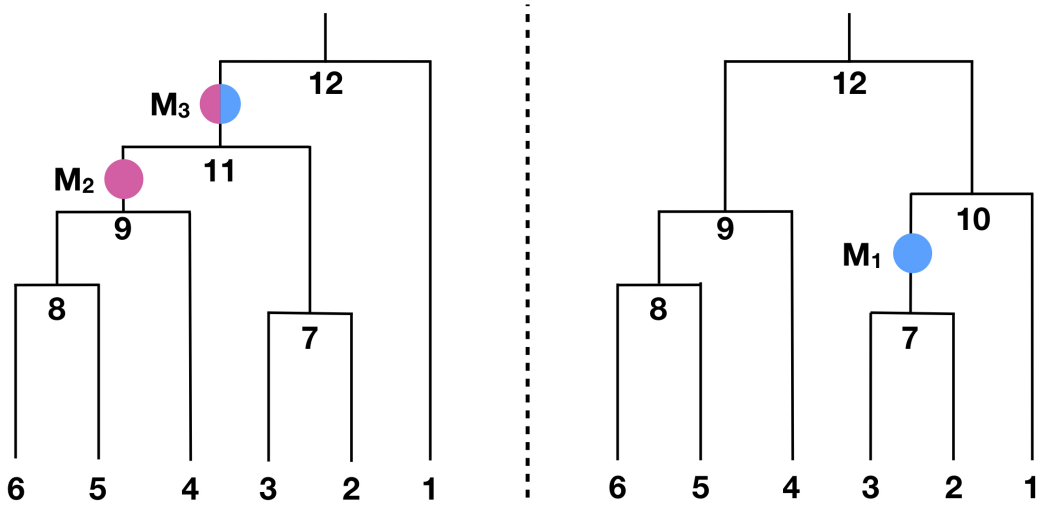


Figure 2.5: Two ancestrally connected pairs of mutations at adjacent genealogies separated by a recombination event. Assigning incorrect relative ages to either the blue or purple pairs of mutations would violate the tree topology. For example, since node eleven is an ancestor of node nine, M_2 cannot be older than M_3. Though the blue mutations appear on different genealogies, since node eleven is an ancestor of node seven, M_1 cannot be older than M_3. However, since M_1 and M_2 are not ancestrally connected, they may be either younger or older than one another provided that they are both younger than M_3.

Multiplying the population scaled mutation rate, $4N_e r$, by the distance separating the sites indicates that 30 recombinations are expected to occur between these sites.

The accuracy of allele frequency is only slightly affected by resampling simulated genotypes under the empirically constructed error profile. The average accuracy under the error model is 83.06%, which differs from the "no error" case by less than 1%. Singletons and variants with frequency $(n-1)$ were excluded based on the specifications of `tsinfer` and `GEVA`, but the accuracy of frequency is approximately 5% higher when singletons are included.

The left panel of Figure 2.7 shows the accuracy of both frequency and `GEVA` plotted against the physical distance separating allele pairs. The code producing the values for frequency with and without error in this figure is the same as in Figure 2.6. However, since `GEVA` is significantly more computationally expensive than determining relative age with frequency alone, results are averaged over 10,000 replicates instead of 100,000 replicates as in Figure 2.6. `GEVA` proves to be more accurate than frequency in the no-error case, with an average accuracy of 84.29%.

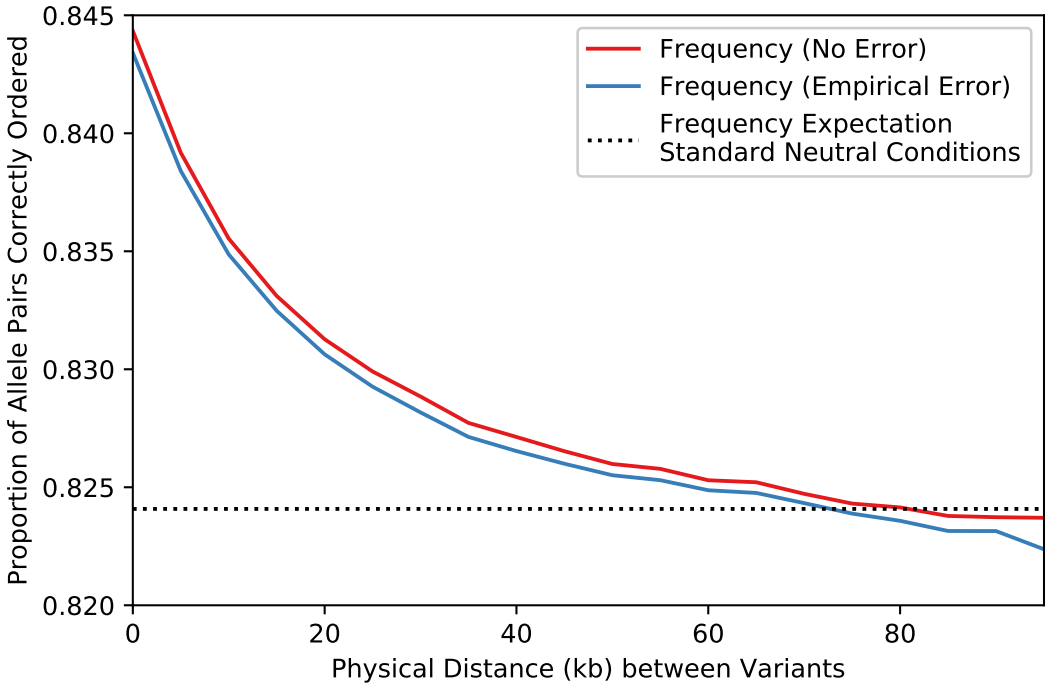


Figure 2.6: Suitability of frequency as a proxy for the relative age of alleles. 50 sequences of 100 kb were simulated using `msprime` with $N_e = 10,000$ and $\mu = r = 10^{-8}$. For all pairs, except those where alleles are the same age or where a variant has only 1 copy of the derived or ancestral allele, the difference in allele age was compared to the difference in variant frequency. The proportion of pairs correctly ordered by frequency is shown. Results were averaged over 5 kb bins and 100,000 replicates. Correlation of adjacent tree topologies increases the accuracy of frequency when variants are physically close to one another. Including singletons and variants at frequency $(n-1)$ increases the average accuracy of frequency by roughly 5%.

Interestingly, `GEVA` appears relatively less robust to error than frequency. The accuracy was 83.26% when estimating ages using genotypes resampled according to the empirical error profile. This value is only marginally higher than the corresponding value for frequency in the error setting.

The right panel of Figure 2.7 shows the same results for simulations based on the "Out of Africa" demographic model. The accuracy of both frequency and `GEVA` is significantly lower in this setting, though `GEVA` appears more robust to demographic

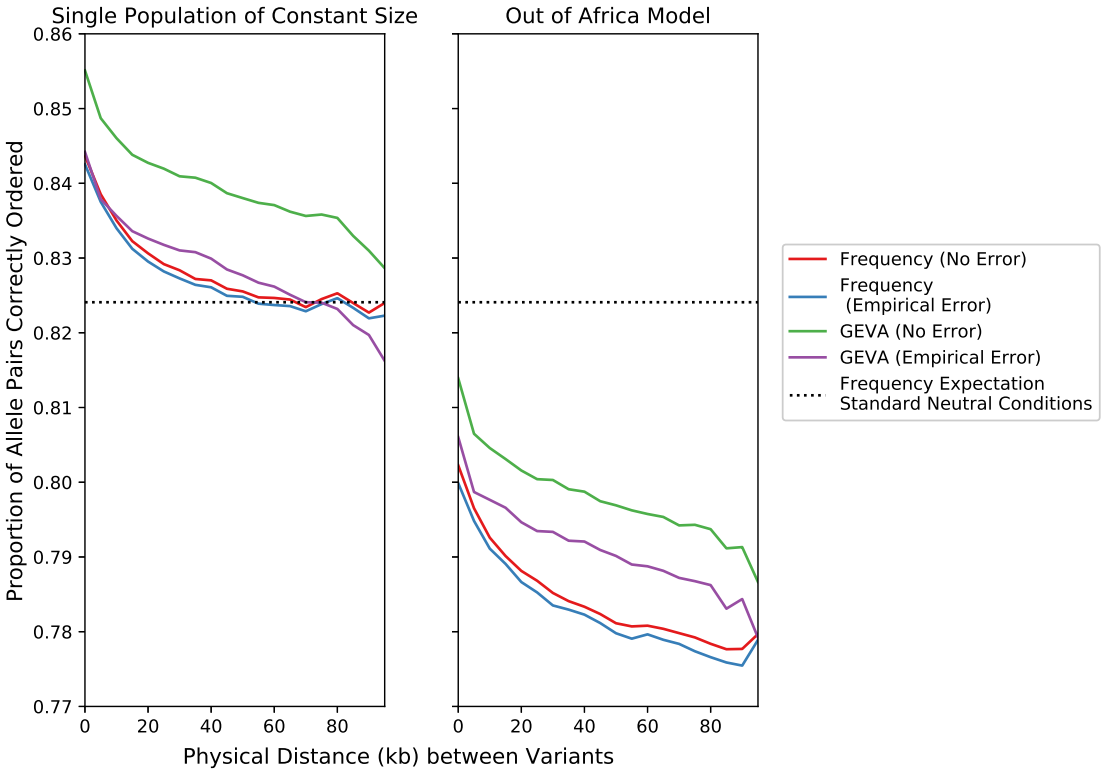


Figure 2.7: Accuracy of frequency and `GEVA` as measures of the relative age of alleles under a standard `msprime` simulation with a single population of constant size $N_e = 10,000$ (left panel) and the Gutenkunst et al. (2009) "Out of Africa" demographic model (right panel). In both cases, 50 sequences of 100 kb were simulated using `msprime` with $\mu = r = 10^{-8}$. Results were averaged over 5 kb bins and 10,000 replicates. Within both simulations, all four statistics shown are based on the same data.

effects than frequency. Frequency correctly ordered 78.77% of allele pairs correctly without error and 78.62% of variants when genotypes were resampled with error. The corresponding values for `GEVA` were 80.20% and 79.44%.

Finally, Figure 2.8 shows the accuracy of both frequency and `GEVA` plotted against the distance separating ancestrally connected allele pairs. Here the relationship between the accuracy of frequency and `GEVA` is reversed, with frequency appearing more accurate than `GEVA` in both the no error and the error settings. The average accuracy of frequency at this subset of pairs when error is not considered is 91.64% and 91.51% when error is injected. The average accuracy of `GEVA` is 89.55% when error is not considered and 88.63% with error.

In Figure 2.8, accuracy first decreases until alleles are separated by approximately 10 kb and then steadily increases. The initial decrease may be the result of increased

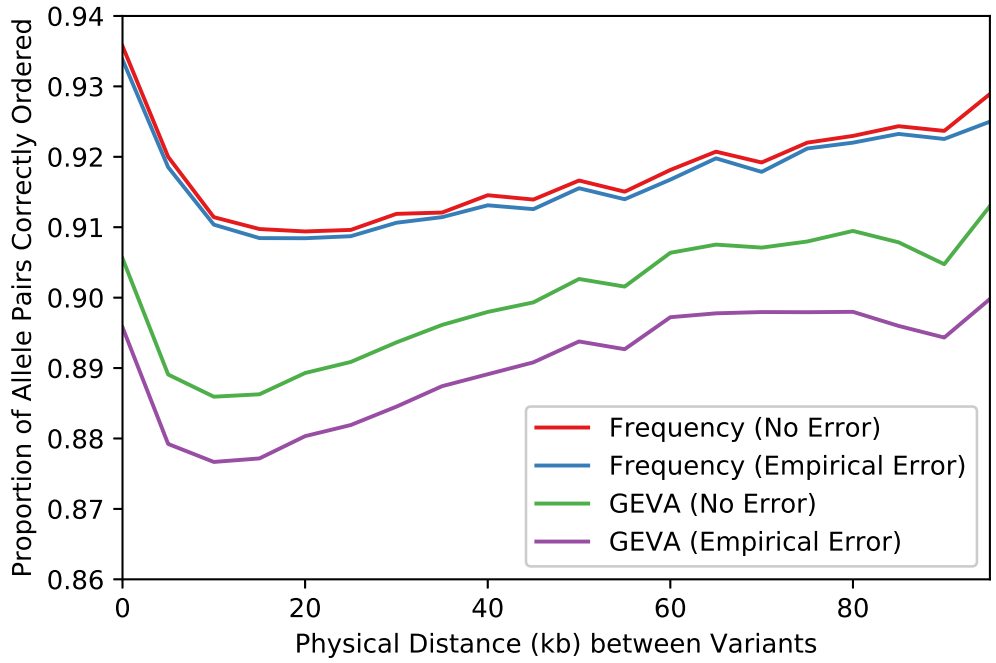


Figure 2.8: Accuracy of frequency and `GEVA` as measures of the relative age of alleles at pairs of ancestrally connected mutations. 50 sequences of 100 kb were simulated using `msprime` with $\mu = r = 10^{-8}$. Results were averaged over 5 kb bins and 10,000 replicates.

accuracy when mutations are in linked genealogies. The subsequent increase may be due to the lower average frequency of ancestrally connected allele pairs situated further apart, since recombination is more likely to break apart ancestrally connected pairs where one or both mutations occurred on longer branches higher in the tree.

To investigate this hypothesis, I plotted the average frequency of pairs of ancestrally connected alleles against the physical distance separating them. The results are shown in Figure 2.9, confirming that the average frequency of these pairs steadily decreases as they are located further from one another. However, the average *difference* in frequency between the allele pairs stayed relatively constant at 35%. According to Figure 2.4, two allele pairs differing by 35% in frequency should be ordered more accurately than another pair differing by 35% but with a higher average frequency. For instance, if one pair has alleles at frequencies of 17% and 52% (the average of which is approximately the highest value in Figure 2.9), the allele at the lower frequency is expected to be younger 75.7% of the time.

Alternatively, if one allele is at frequency 10% and the other is at 45% (the average of which is approximately the lowest value in Figure 2.9), the expected accuracy of frequency in reflecting allele age is 82.7%.

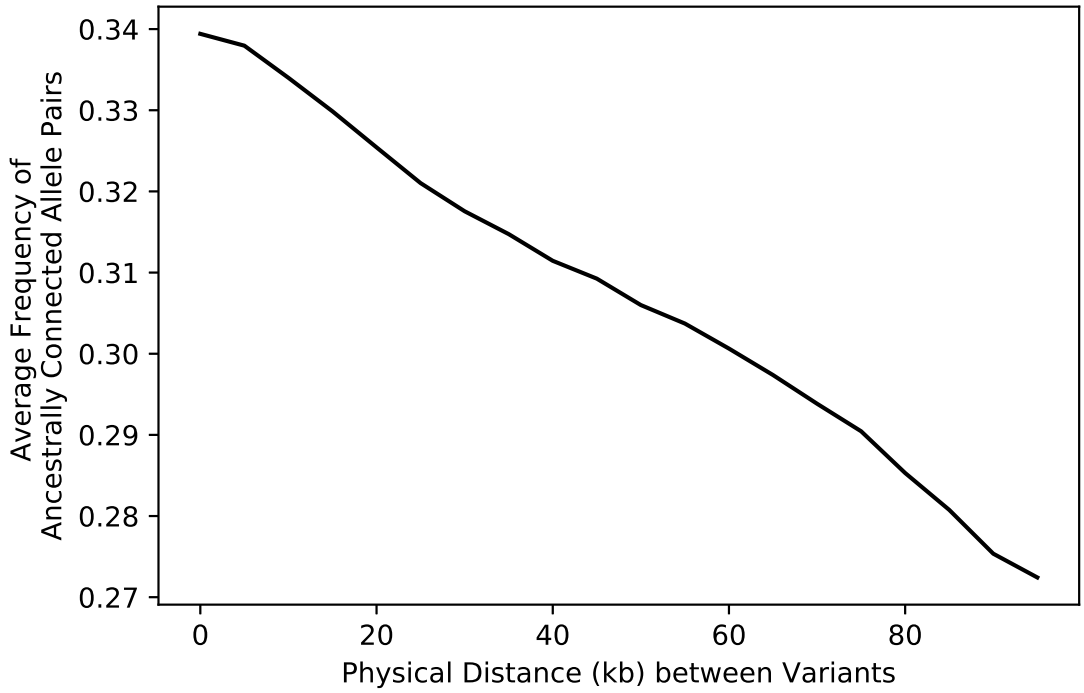


Figure 2.9: Average frequency of ancestrally connected allele pairs. 50 sequences of 100 kb were simulated using `msprime` with $\mu = r = 10^{-8}$. Results were averaged over 5 kb bins and 10,000 replicates.

Finally, I sought to investigate a setting particularly relevant to `tsinfer`: ancestrally connected pairs of mutations with similar frequencies. Figure 2.10 shows the results of this simulation, where 9,142 mutations have 15.5 million ancestrally connected pairs. Of these, 5.5 million pairs differ by a frequency of less than 20%. Frequency is a highly accurate proxy in this setting, as the more frequent allele is older 86.1% of the time.

The code to run these simulations and produce these plots can be found at `https://github.com/awohns/relative-allele-age`. Note that important considerations when producing the simulation results were how to handle alleles of the same frequency or age, whether to include singletons and variants with frequency $(n-1)$, and which pairs of alleles to use as a baseline for the accuracy

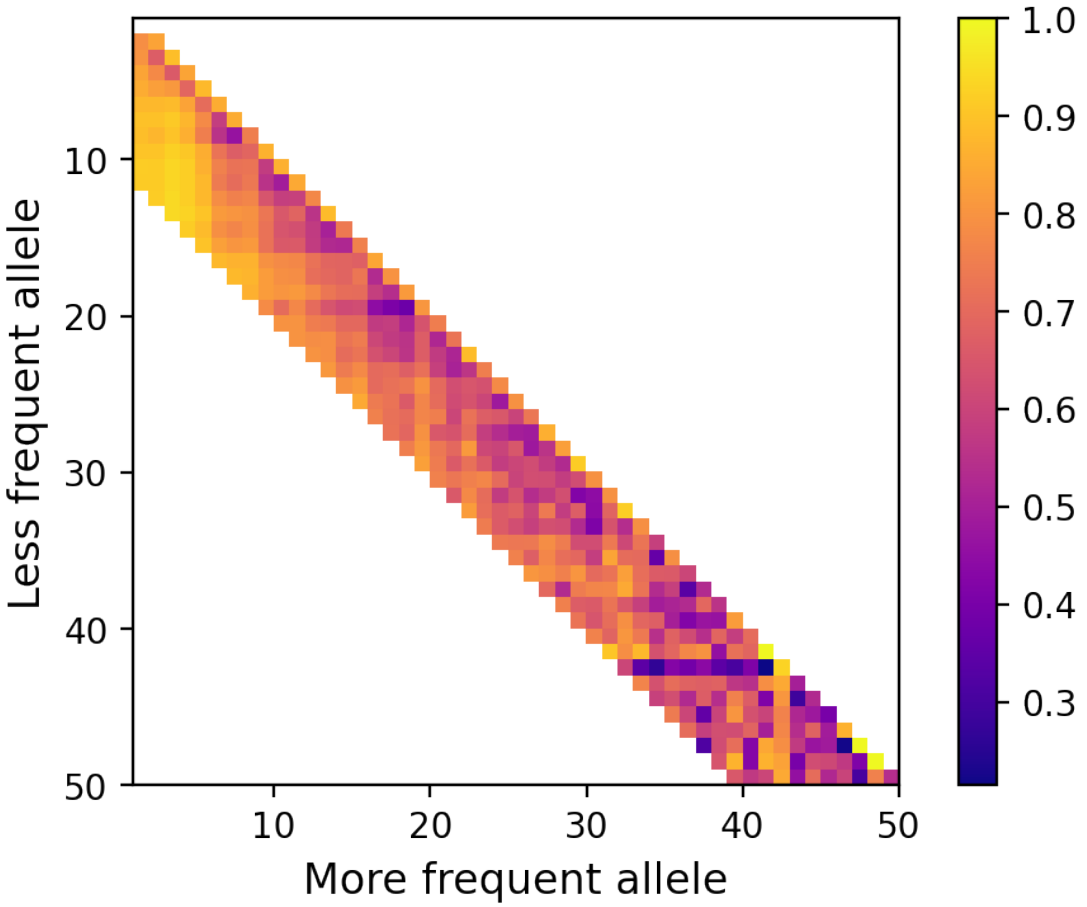


Figure 2.10: Heatmap showing the empirical accuracy of allele frequency at ancestrally connected alleles of similar frequency. Each cell shows the probability that the more frequent allele of an ancestrally connected pair is older than the less frequent allele. The number of samples descending from the more frequent allele is shown on the x axis, while the number of samples descending from the less frequent allele is shown on the y axis. The data shown was generated by an `msprime` simulation with 50 sequences of 5 Mb, $N_e = 10,000$ and $\mu = r = 10^{-8}$.

statistics. In each case, choosing different rules for which allele pairs are included in the analysis would lead to different accuracy statistics.

2.6 Discussion

The results presented in this chapter support `tsinfer`'s use of frequency as a heuristic measure of relative age. Allele frequency can accurately determine the relative age of alleles with greater than 78% accuracy in all situations considered. Average accuracy rises to more than 90% at ancestrally connected alleles: the

most important subset of allele pairs for the purposes of genealogical inference. The relevance of this subset is due to the fact that incorrectly ordering these alleles would cause topological inaccuracy in an inferred tree sequence. Notably, frequency was more accurate than `GEVA` in assigning relative times at this subset. It should be emphasised, however, that `GEVA` is designed to estimate *absolute*, not relative, allele ages.

The accuracy of frequency was observed to be highly robust to the effects of resampling simulated genotypes by the empirically constructed error profile from Albers and McVean (2020). The difference between the no error and error cases in each simulation was approximately 1%, which was only slightly higher than the overall error rate in the empirical error profile of 0.079%. `GEVA` was much more susceptible to error, which implies this method may have only a minor advantage over frequency when using real genetic data. However, `GEVA` performs relatively better on data simulated under the "Out of Africa" model. Accuracy in datasets produced with this simulation is likely to be affected by demographic factors such as population structure and variable N_e. To examine this further, future work could examine the performance of frequency compared to `GEVA` within the three populations simulated, rather than across all three populations.

A practical concern in the genealogical inference setting is the expense of determining relative age. Frequency is a much more efficient approach than `GEVA`. In the standard simulation parameter scheme, each simulation replicate and analysis of allele pairs took only 1.24 seconds when only frequency was used to determine relative age, while the simulations and analyses where both `GEVA` and frequency were used took 27.8 seconds each.

Though the results presented here support `tsinfer`'s use of frequency as a heuristic measure of relative age, measuring the actual effect of misordering variants on the accuracy of inferred tree sequences using the metrics presented in Kelleher et al. (2019) could provide insight into the downstream effects of this heuristic. Specifically, determining whether the slight advantage in accuracy offered by `GEVA` at all allele pairs meaningfully affects the accuracy of genealogical inference may be

an interesting avenue for future work. However, any potential benefit should be weighed against the increased computational burden of `GEVA` or a similar approach.

Having validated a key assumption in `tsinfer`, the method is used in the following chapters to infer tree sequences from a variety of population sequencing datasets.

3

Analysing tree sequence topologies

Contents

3.1 Introduction 57
3.2 *Genealogical nearest neighbours* 59
3.2.1 1000 Genomes Project and Simons Genome Diversity Project 61
3.3 *Genomic descent* 62
3.3.1 Trumpington Neolithic samples 64
3.3.2 Afanasievo family and UK Biobank 68
3.4 Discussion 73

3.1 Introduction

Summary statistics such as Tajima's D (Tajima, 1989) and Patterson's f-statistics (Patterson et al., 2012) provide insight into population genetic processes including selection, population growth, and migration. These statistics are commonly calculated using the genotype matrix, but can also be found using the branches of a tree sequence, as demonstrated in Section 1.7 (Ralph, Thornton, and Kelleher, 2020). Such genealogy-based statistics are of interest as they potentially offer greater power than equivalent, site-based statistics (Harris, 2019; Ralph, Thornton, and Kelleher, 2020).

In this chapter, I discuss two summary statistics defined exclusively on a tree sequence topology that provide insight into genetic relationships between individuals and populations. The first is *genealogical nearest neighbours* (*GNN*), introduced in Kelleher et al. (2019), which quantifies the topological relationships of samples to one another. As a co-author of this paper, I was involved in the formulation and application of this statistic. The second statistic is genomic descent, a restricted version of *GNN* that quantifies descent from ancestral haplotypes, which I introduced in Scheib et al. (2019). The relevant sections of these two publications appear in modified form in this chapter, see joint authorship statement for details.

In this chapter, I show how *GNN* and *genomic descent* can be used to understand population genetic relationships in tree sequences inferred from a variety of modern and ancient genetic datasets using `tsinfer` version 0.1 (see Section 1.5.1). These datasets are the 1000 Genomes Project (1000 Genomes Project Consortium, 2015), Simons Genome Diversity Project (Mallick et al., 2016), UK Biobank (Bycroft et al., 2018), and two previously unpublished ancient datasets. The first is whole genome data from two brothers (with average coverage of 1.6x and 1.2x) who lived in the vicinity of Trumpington, England during the Neolithic period ~ 5.7 thousand years ago (kya) (Scheib et al., 2019). The second is high coverage whole genome data from a nuclear family of four (a mother, a father, and their two sons with average coverage of 21.2x, 25.3x, 10.8x, and 25.8x) from the Afanasievo Culture, ~ 4.1 kya. This latter dataset allows for more reliable phasing, a major challenge when working with ancient data.

GNN reveals known patterns of worldwide genetic relatedness in the TGP and SGDP, as well as patterns of admixture in populations and individual chromosomes. The *genomic descent* statistic shows that modern populations in the TGP and SGDP have differential affinities to haplotypes carried by the Trumpington brothers. Finally, an inferred tree sequence from the Afanasievo family, TGP, and UKB suggests an intriguing gradient of relatedness to the Afanasievo family in Great Britain and Ireland, with higher affinities in historically Celtic regions.

Since allele frequency is used to provide a relative order to mutations and ancestral haplotypes in `tsinfer` (as described in Chapter 2), it is not possible to order ancient samples in relation to estimated ancestral haplotypes. In this chapter, ancient samples are thus included as the most recent ancestral haplotypes in the second step of `tsinfer` (see Section 1.5.1). An iterative approach placing ancient samples at their archaeologically-dated age in the tree sequence is introduced in Chapter 4.

3.2 *Genealogical nearest neighbours*

The *GNN* statistic summarises the identity of the "nearest neighbours" of a focal node based on the topological properties of trees, defined with respect to a collection of reference sets (lists of nodes in the tree sequence). Figure 3.1 provides a schematic representation of the statistic.

For a given tree and focal node u, *GNN* values are computed by traversing upwards from u until we find a target node v that has one or more descendants (not including u) present in any of the reference sets. The *GNN*s are then the proportions of each reference set among the descendants of v, not including u. Note that this is conceptually similar information to the information provided by the software `Chromopainter` (Lawson et al., 2012).

More formally, let R be a list of K sets of "reference nodes", and let $C^t_{v,k}$ be the number of nodes descending from (and including) v in tree t from the set R_k, with $C^t_v = \sum_{k=1}^K C^t_{v,k}$. For a given focal node u and tree t, the *GNN* statistic is then defined as

$$G^t_{u,k} = \frac{C^t_{v,k} - [u \in R_k]}{C^t_v - \delta(u)} \tag{3.1}$$

where $[\mathbf{x}]$ is an Iversonian condition such that $[\mathbf{x}] = 0$ if $\mathbf{x}$ is false and $[\mathbf{x}] = 1$ otherwise; $\delta(u) = [u \in \bigcup_{j=1}^K R_j]$; and v is the first node on the path from u to root in t such that $C^t_v > \delta(u)$. The average *GNN* along a tree sequence $\mathbb{T}$ is

$$G_{u,k} = \frac{1}{L} \sum_{t \in \mathbb{T}} L^t G^t_{u,k}, \tag{3.2}$$

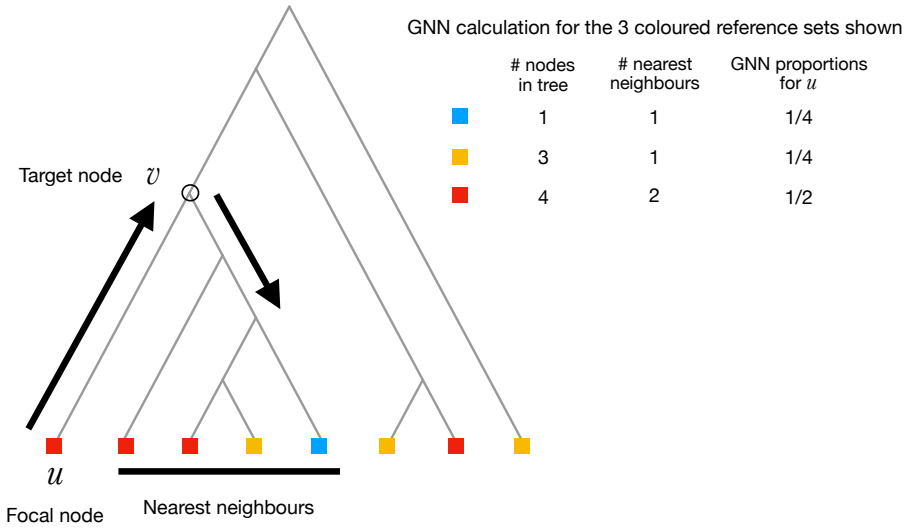


Figure 3.1: Adapted from (Kelleher et al., 2019) Figure S19: *Genealogical nearest neighbours* (*GNN*) example for a focal node u with respect to three reference sets (indicated by colours). To compute *GNN* values we first traverse upwards from focal node u to find the target node v. This node is the first node on the path from u to root which counts among its descendants at least one member of a reference set (not including u). The *GNN* values are then computed as the proportion of descendants of v (not including u) from each of the reference sets.

where each tree $t \in \mathbb{T}$ covers a span of L^t units of genetic sequence and $L = \sum_{t \in \mathbb{T}} L^t$. Thus, the average *GNN*s are equal to the treewise *GNN* values weighted by each tree's genomic span.

Note that if the reference sets do not include all leaf nodes in a tree, the sister clade to a leaf u may not contain any individuals in a reference set. In this case we will ascend to an older target node v, higher in the tree, and the *GNN* statistic will represent a more distant relationship.

In dated genealogies, the age of the target node v is also informative of the nature of genetic relationships. To capture this information, I implemented a modified version of *GNN* called *windowed-genealogical nearest neighbours* which assigns nearest neighbours to user-defined epochs ("time windows") as well as regions of the genome ("span windows"). While the output of *GNN* is a n by m matrix, with n equal to the number of focal nodes and m equal to the number of reference sets,

windowed-*GNN* returns a r by s by n by m matrix, where r is the number of span windows and s is the number of time windows. Note that this definition expands the "windowed statistics" from Ralph, Thornton, and Kelleher (2020) to include time-based windows as well. While I do not report results using the windowed *GNN* in this thesis, it represents a promising means of understanding more granular genetic relationships in a dated tree sequence.

Both *GNN* and *windowed-GNN* are implemented in `tskit` (see Section 1.3).

3.2.1 1000 Genomes Project and Simons Genome Diversity Project

In Kelleher et al. (2019), my co-authors and I used *GNN* to characterise ancestral relationships in tree sequences inferred using `tsinfer`. We used three datasets on human chromosome 20: the 1000 Genomes Project (TGP), consisting of low-coverage whole genome sequencing data from 2504 individuals across 26 worldwide populations (1000 Genomes Project Consortium, 2015); the Simons Genome Diversity Project (SGDP), consisting of high coverage sequencing data from 278 individuals from 142 worldwide populations (Mallick et al., 2016); and the UK Biobank (UKB), consisting of SNP array data from 487,327 individuals within the UK (Bycroft et al., 2018). In this section, I describe tree sequences inferred from the TGP and SGDP and show results from the corresponding *GNN* analyses. See Kelleher et al. (2019) for results from the UKB.

Across chromosome 20, the TGP data consisted of 860K sites after filtering (see methods of Kelleher et al. (2019) for details). After inferring ancestors and matching sample haplotypes to these ancestors, we obtained a 297MiB tree sequence. Under the standard neutral model we expect the number of distinct trees to be roughly equal to the number of variable sites within human data, and the number of edges to be roughly four times the number of trees (Kelleher, Etheridge, and McVean, 2016; Kelleher et al., 2018). We observe similar numbers of distinct trees (550K) as filtered sites (860K, from 1.8M unfiltered), but the number of edges is 7.3M, which is greater than expected and suggests that the differences between

adjacent inferred trees are larger than expected under idealised conditions. The SGDP data consisted of 348K sites after filtering, and this resulted in an 83MiB tree sequence with 1.7M edges and 196K distinct trees.

We next computed the *GNN* statistic for each individual in the TGP and SGDP using the inferred tree sequences. We find that, despite the noise generated by uncertainty in tree reconstruction (manifest as polytomies), the chance nature of the genealogical process and data error, the tree sequences can characterise global population structure (Fig. 3.2a,b), within-population relatedness (Fig. 3.2c) and identify regions of differential ancestry within an individual (Fig. 3.2d). These analyses demonstrate the potential of interrogating the inferred genealogical structure at different resolutions to describe both broad and fine-scale patterns in contemporary human genomic diversity.

3.3 *Genomic descent*

The *genomic descent* statistic quantifies the proportion of the total genetic material in a reference set descending from focal ancestral nodes in a tree sequence. This statistic is useful for analysing tree sequences with ancient samples.

Let R be a list of k sets of nodes. In our application this refers to the k populations in TGP or SGDP with each node in R_k representing a sampled chromosome. For node i in reference set R_k, define A^c_{iu} as the total span of node i which descends from node u on chromosome c. The statistic is averaged over A^c_u, the portion of the genome where the node is ancestral to any sample from the sets in R. Finally, it is normalized by the size of reference set $|R_k|$. Thus, the *genomic descent* statistic of node u in reference set k on chromosome c is:

$$D_{c,R_k}(u) = \frac{\sum_{i \in R_k} A^c_{iu}}{A^c_u |R_k|}. \tag{3.3}$$

The *mean_descendants* function in `tskit` implements a version of *genomic descent* which is unnormalised by sample set size. This is equivalent to Equation 3.3 without the $|R_k|$ term in the denominator.

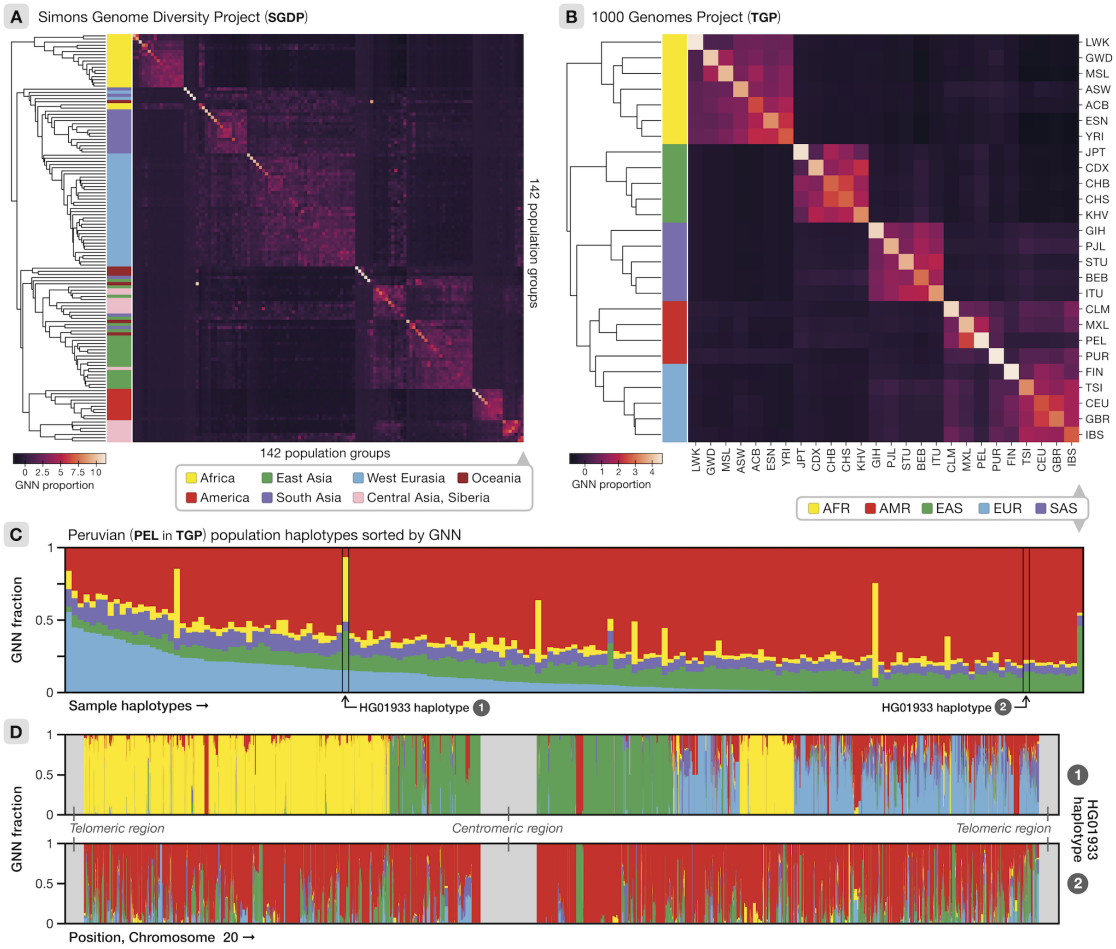


Figure 3.2: Caption adapted from Kelleher et al. (2019) Figure 4: Tree sequence characterisation of global genome diversity. a). Z-score normalised *GNN* proportions for SGDP by population ($n = 278$ individuals). The *GNN* matrix was first z-score normalised by column and the rows then hierarchically clustered. b). As for (a), but for the TGP data ($n = 2504$ individuals). c). Average *GNN* proportions for all individuals within the PEL population in TGP. Colours indicate continental-level groupings. d). The *GNN* proportions across the chromosome for the two haplotypes of HG01933, from the PEL population in TGP. HG01933 was chosen as an example of an individual showing evidence of very recent admixture from multiple source populations. We note that apparent short tracts of different ancestry most likely do not reflect true changes in recent ancestry, but arise through the stochastic nature of genealogical processes and errors in inference.

To combine information from the autosomes for a given population, the total *genomic descent* statistic is calculated by adjusting the value of Equation 3.3 for each chromosome by the ratio of the size of each chromosome, L_c, to the total genome length, L_g:

$$D_{g,R_k}(u) = \frac{\sum_{c=1}^{22} L_c D_{c,R_k}(u)}{L_g}. \tag{3.4}$$

3.3.1 Trumpington Neolithic samples

In Scheib et al. (2019), I used the *genomic descent* statistic to investigate the population genetic relationships of two Neolithic-era individuals to modern and ancient populations. Recent archaeogenomic research in Britain has mainly focused on the broad genetic affinities of Mesolithic (n=6), Neolithic (n=83) and Bronze Age (n=67) people and population turnover coinciding with material culture change (Olalde et al., 2018; Brace et al., 2019), with one study focusing on the connections within and between megalithic burials along the Atlantic façade (Sánchez-Quinto et al., 2019). To date, Neolithic sites from the easternmost section of southern Britain (East Anglia) remain uncharacterised. At the site of Trumpington Meadows, Cambridgeshire, two adjacent, broadly contemporary monuments, erected as early as 3762–3648 cal. BCE and modified through the Early Bronze Age, were excavated revealing an internment of at least three partially articulated and disarticulated older middle/mature adult males in Monument I (Evans, Lucy, and Patten, 2018). Two of the individuals, Sk.4/799 and Sk.1/880, were buried in close association at the south end of the burial chamber with a third disarticulated Sk.3/800 overlying Sk.1/880 that may be the same individual. All individuals were directly radiocarbon dated (Scheib et al. (2019), Table 1), suggesting usage of the burial chamber over 50–75 years (Evans, Lucy, and Patten, 2018).

My collaborators successfully extracted ancient DNA from the teeth of Sk.1/880 and Sk.4/799. Extracts were built into double-stranded, single-indexed libraries and were shotgun-sequenced on Illumina NextSeq500 using the single-end 75 base pair kit. Variants were called using the ATLAS pipeline (Link et al., 2017) using genotype likelihoods of 2.8 million SNPs with a minor allele count >= 5 in TGP Phase 3. Base quality was recalibrated according to post-mortem damage pattern estimated from heterozygous reads on chromosome X. Phasing and imputation were performed as follows: genotypes were first estimated with BEAGLE 4.1 (Browning and Browning, 2007) from genotype likelihoods produced by ATLAS (Link et al., 2017) in Beagle -gl mode, followed by imputation in Beagle -gt mode with BEAGLE 5 (Browning, Zhou, and Browning, 2018) using the 503 Europeans genomes in

TGP as the reference panel. Variants with a minor allele count below 5 in the full TGP reference panel were filtered out.

To infer descent from the Trumpington haplotypes within the TGP and SGDP, we made the phased and imputed ancient genomes available as potential ancestors of modern samples using `tsinfer`. The analysis conducted in this work "augments" tree sequences generated from the TGP and SGDP datasets with the phased and imputed genomes of the two Trumpington individuals. Augmentation consists of inserting the four chromosomes of the Trumpington individuals at the lowest level of the copying process of `tsinfer` (see Section 1.5.1). This is equivalent to using the ancient samples as the youngest ancestral haplotypes in the copying process, allowing these samples to match against inferred ancestral haplotypes and the samples in the TGP and SGDP to directly match against the ancient individuals.

To augment the previously built tree sequences, it was necessary to find the overlap between the imputed ancient chromosomes (where alleles were filtered as described above) and modern genetic material. The imputed ancient samples included 64.2% of the TGP sites, leaving an overlap of 24,570,629 sites for analysis. In the SGDP dataset 81.4% of sites overlapped, leaving 12,374,480 sites for analysis. After the non-overlapping sites were pruned from the dataset, we augmented the ancestral tree sequence of each TGP and SGDP chromosome with the ancient samples and then matched the modern samples to the ancestral tree sequences.

It is important to note that this copying process does not necessarily mean that the Trumpington individuals were direct genetic ancestors of the modern samples. Instead, it indicates that, over some parts of the genome, individuals genetically indistinguishable from the Trumpington brothers are a better match for modern samples than any of the inferred ancestors.

The resulting tree sequences contain considerable detail about the ancestry of the ancient and modern samples at every site included in the analysis. To gain a high-level understanding of the genomic legacy of the Trumpington individuals we computed the *genomic descent* statistic. This statistic summarises the relative proportion of the aggregated genetic material of each TGP and SGDP population

which "copies," or putatively descends from, the exact haplotypes found in the Trumpington individuals (Figure 3.3b).

The height of the stacked bars in Figure 3.3B reflects the total *genomic descent* statistic for each population (the value of Equation 2). The relative size of each colored section of the stacked bars in Figure 3.3B reflects the value of Equation 1. Thus, while the total height represents the genomic inheritance from the Trumpington haplotypes, the relative size of coloured regions within the bar shows the contribution from each chromosome relative to the length of that chromosome. In other words, long chromosomes such as chromosome 1 will not necessarily show a greater relative contribution than smaller chromosomes. This has been done to highlight any differences between chromosomes in terms of their proportion of *genomic descent* from Trumpington haplotypes. The evenness of the distribution of colours in the different bars indicates that most chromosomes support the same pattern of genomic descent. Large deviations between chromosomes, for example in chromosome 22 in the SGDP Itelman population, may be due to phasing or imputation errors in ancient or SGDP samples (see Kelleher et al. (2019) for evidence of phasing errors in some TGP and SGDP samples). Variance due to small reference set size (for instance, only two Itelman individuals are included in the SGDP) or small chromosome size may also be potential explanatory factors.

Among the TGP super-populations, the greatest amount of genetic material derived from Trumpington haplotypes is found in European populations, followed by admixed American populations. Iberians (IBS) show the greatest amount of genetic material attributable to the Trumpington haplotypes, consistent with recent findings (Olalde et al., 2018). In the SGDP, Sardinians, followed by Basque and Spanish individuals, show the highest levels of genetic descent from the Trumpington haplotypes. Sardinians are thought to have strong links to "Early Neolithic Farmers" (Lazaridis et al., 2014; Marcus et al., 2020). These results also support recent work from Chiang et al. (2018), who used D and f-statistics to suggest shared Neolithic ancestry between Basques and Sardinians.

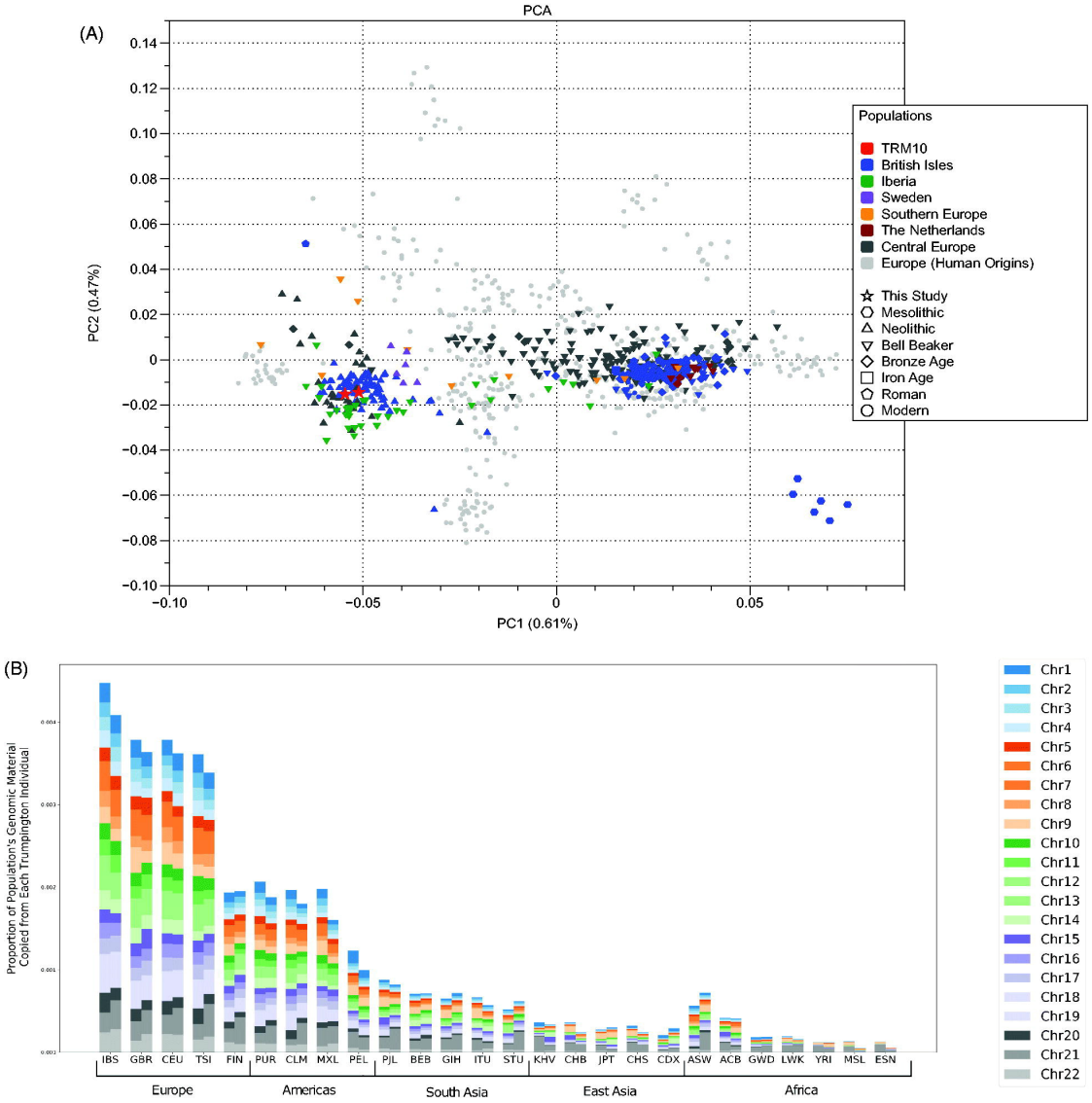


Figure 3.3: Caption adapted from Scheib et al. (2019) Figure 2: (A) Principal component analysis of ancient samples from Olalde et al. (2018), Brace et al. (2019), and Sánchez-Quinto et al. (2019) projected onto modern variation from 1 KGP (1000 Genomes Project Consortium, 2015). (B) *Genomic descent* from the Trumpington samples in the 1000 Genomes Project. Bar height indicates the relative proportion of genomic material in the named population, which is inferred to copy directly from the haplotypes present in Sk.4/799 (left side of each bar) and Sk.1/880 (right side of each bar). Colours give the contribution from each chromosome, relative to chromosome size, to indicate if the pattern differs for different chromosomes. Populations are sorted by super population and decreasing values of the *genomic descent* statistic (averaged between Sk.4/799 and Sk.1/880).

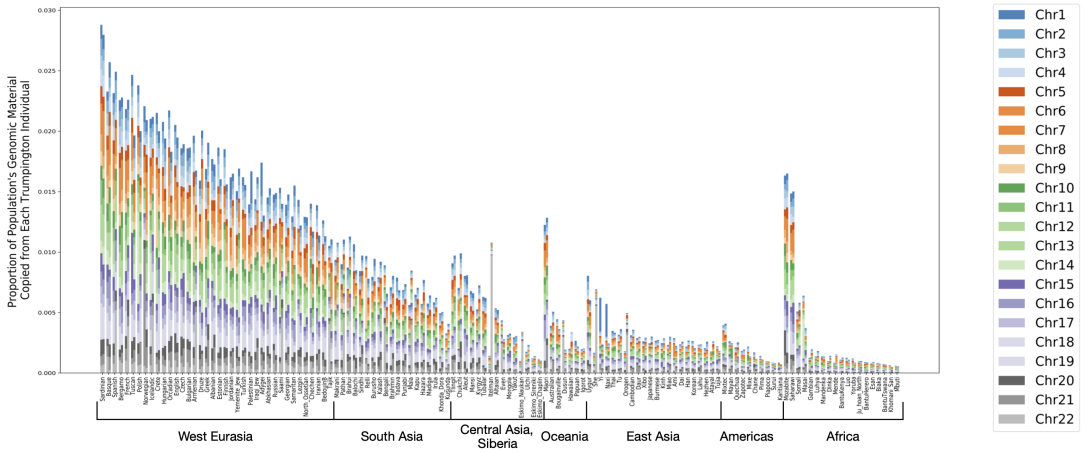


Figure 3.4: Original caption from Scheib et al. (2019) Figure S2: *Genomic descent* from the Trumpington samples in the Simons Genome Diversity Panel. Bar height indicates the relative proportion of genomic material in the named population which is inferred to copy directly from the haplotypes present in Sk.4/799 (left side of each bar) and Sk.1/880 (right side of each bar). Colours give the contribution from each chromosome, relative to chromosome size, to indicate if the pattern differs for different chromosomes. Populations are sorted by super population and decreasing values of the *genomic descent* statistic (averaged between Sk.4/799 and Sk.1/880).

The code implementing this analysis can be found at `https://github.com/awohns/neolithic-tsinfer`.

3.3.2 Afanasievo family and UK Biobank

Ancient DNA studies have revealed that individuals from the Eurasian Steppe appeared in Europe in the third millennium BCE and in South Asia in the second millennium BCE, leading to major changes in genetic diversity and material culture, and possibly introducing Indo-European languages to these regions (Lazaridis et al., 2014; Haak et al., 2015; Narasimhan et al., 2019). An unresolved question in the genetic history of Britain and Ireland is the distribution of this "steppe-like ancestry" in modern individuals across the region. Cassidy et al. (2016) showed that Neolithic and Bronze Age ancient DNA from Ireland suggests a degree of population continuity in modern "Celtic" populations from ancient groups carrying steppe-like ancestry. Intriguingly, the authors also show that a derived allele causing hemochromatosis reaching global maximum frequency in Ireland may have originated in steppe peoples.

Galinsky et al. (2016) used an early release of the UK Biobank and ancient samples from the Eurasian Steppe to argue that Scottish individuals in the UK Biobank have closer affinity with ancient steppe peoples than English or Welsh individuals.

To further investigate these signals of differential steppe-like ancestry in the UK and Ireland, I used the UKB and a novel dataset of a family of four from the Afanasievo culture, consisting of a mother, father, and two sons (Wohns et al., 2021), to investigate links between steppe peoples and modern individuals living in the UK.

Afanasievo quartet dataset

The Afanasievo pastoralist culture existed $\sim 4.5 - 5.1$ kya in the Altai mountains and Yenisey River basin in modern day Russia. The Afanasievo had strong cultural and genetic links to the Yamnaya peoples living on the European steppe thousands of miles to the West. Both groups are thought to be the progenitors of later populations carrying steppe-like ancestry (Narasimhan et al., 2019).

The Afanasievo quartet were sampled from the left bank of the Karasuk tributary of the Yenisey River in the Altai Mountains of Russia. The Reich laboratory generated between 5 to 8 lanes of shotgun sequencing data from these samples using an Illumina HiSeqX10 instrument (2 x 101 cycles and reading out the indices with 2 x 7 cycles), using sequencing libraries for which in-solution enrichment data was previously published in Narasimhan et al. (2019) (Supplementary Table 1). They merged the paired sequences and mapped these to the human genome reference sequence using the samse command in BWA v.0.7.15-r114053 with the parameters -n 0.01, -o 2, and -l 16500, and removed duplicated molecules as described in the original publication that generated in-solution enrichment data on these libraries. The average coverage measured on the autosomal SNP targets used for the in-solution enrichment experiment was 10.8x for I3388 (the mother in the family), 25.8x for I3950 (the father in the family), 21.2x for I6714 (one of the sons in the family), and 25.3x for I3949 (the other son of the family).

The family relationships in this quartet should allow for reliable phasing of the ancient genomes. Collaborators used bcftools (v1.10.2) to calculate genotype

likelihoods of biallelic SNPs in the Afanasievo samples. Beagle 4.0 (Browning and Browning, 2007) was used for phasing by providing a genotype likelihood file (gl argument) and pedigree file (ped argument) to phase this family without any reference panel and imputation. Sites with max(GP) < 0.99 were excluded.

Evaluating phasing accuracy in the Afanasievo family

I utilised the haplotype matching algorithm in `tsinfer` to gain a qualitative understanding of phasing quality in the Afanasievo dataset. Parents were defined as "ancestors" and sons as "samples" when running the haplotype matching HMM. The resulting haplotypes are visualised in Figure 3.5, where the "parent node" of each of the son's chromosomes is colour-coded by its parental chromosome of origin.

If the quartet were perfectly phased, only a few recombination events on each chromosome would be expected in each generation (Lajugie et al., 2013). In this case, Figure 3.5 would show one of each son's chromosomes as entirely blue and the other green. Within each chromosome, switching between the shades of blue or green would demarcate recombination breakpoints within the paternal and maternal chromosomes. The observed patterns instead show that each son contains an obvious long-range phasing error: in Son1 at $\sim$ 36Mb and in Son2 at $\sim$ 32Mb. Phasing error in the parents manifests as reversed shades of blue or green when comparing between the two sons. This is particularly clear when comparing the maternal inheritance in each son. It is also evident in the paternal inherited material in "Chromo20b" of both Son1 and Son2 from $\sim$ 50Mb onwards.

Many of these switch errors may be resolved by more sophisticated phasing approaches. In a family quartet it is possible to reliably phase all sites in the children, save for those where all four individuals are heterozygous. Parental haplotype phase can also be determined within regions bounded by recombination breakpoints in the children, though these blocks cannot not be resolved relative to one another (Roach et al., 2010; Lajugie et al., 2013). Some of the patterns in Figure 3.5 may be indicative of genotype error. For instance, both sons appear to share some of same recombination breakpoints: an unusual pattern that is likely

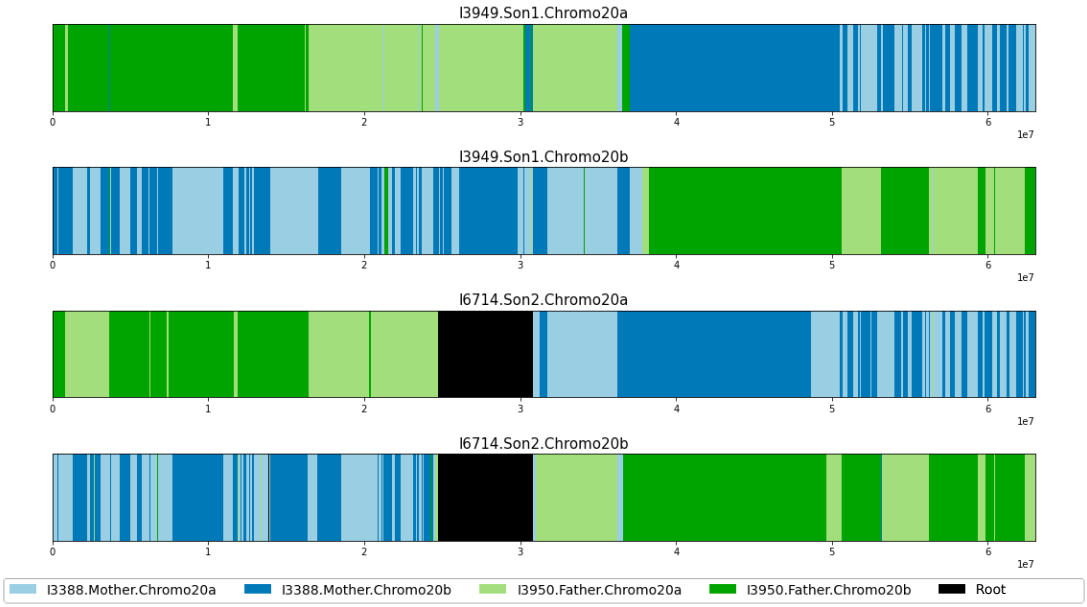


Figure 3.5: Evaluating phasing accuracy in the Afanasievo quartet. The four child chromosomes, two from each son, are matched against the parental chromosomes using the `tsinfer` haplotype matching algorithm. Each child chromosome is colour-coded corresponding to its inferred parent along the sequence. Shades of blue indicate the mothers' maternal and paternal chromosomes; shades of green indicate the fathers' maternal and paternal chromosomes. The x axis for each chromosome is in units of base pairs. "Root" indicates when a son's haplotype does not match either of the parents. This is observed at the centromere in "Son2".

explained by error. Furthermore, sporadic switches between parental chromosomes may indicate phasing errors in the parents or genotype error. Overall, this analysis suggests imperfect variant calling and phasing in these samples, as well as further opportunities to use `tsinfer` as a data quality diagnostic tool.

Descent from Afanasievo Family in UK Biobank

I used `tsinfer` and the *genomic descent* statistic to investigate genetic relationships between the Afanasievo samples and individuals in the UKB. The UKB genotype data contains only 15.8K SNPs on chromosome 20, thus requiring a modified approach to the previously described analysis of the Trumpington samples. Rather than match the Afanasievo into a tree sequence created using the sparse UKB data, I first matched them against a tree sequence inferred from the TGP dataset (see Section 3.2.1). The TGP contains 860K non-singleton, biallelic SNPs on

chromosome 20 while the Afanasievo dataset contains 308K SNPs. Note that the version of the Afanasievo dataset used in Chapter 6 was produced using a different variant calling strategy and contains a larger number of variants. The TGP tree sequence and Afanasievo dataset were reduced to $\sim$ 300K overlapping SNPs and the Afanasievo were matched against this tree sequence using the haplotype matching HMM implemented in `tsinfer`. Both the resulting tree sequence and UKB samples were then reduced to the $\sim$ 9K sites observed in all three datasets. Finally, UKB individuals were matched against the combined TGP/Afanasievo tree sequence.

The resulting tree sequence of chromosome 20 contains 974,662 UKB haploid chromosomes, 9,519 sites, 9,519 trees, 154,784,265 edges, and 3,904,144 ancestral haplotypes (consisting of TGP ancestors and samples, as well as the Afanasievo family). In this initial study, tree sequences were also inferred for chromosomes 17-22.

I calculated the *genomic descent* statistic using the Afanasievo samples as focal nodes. Participants of self-described "white British" ethnicity were assigned place of birth-defined reference sets, categorised by Nomenclature of Territorial Units for Statistics (NUTS) level 2 regions. Other participants were placed in reference sets corresponding to their self-described ethnicity.

Figure 3.6 shows the *genomic descent* statistic from the Afanasievo haplotypes for each reference set; Figure 3.7 shows results for UK-born white British individuals on a map of the UK, with Irish individuals providing results for the Republic of Ireland. Notably, the highest levels of descent are observed in individuals of self-described Irish ethnicity, as well as white British participants born in traditionally Celtic regions of the UK, such as Cornwall, Northern Ireland, Scotland and Wales. Outside of the UK and Ireland, lower levels of descent are observed in non-white British individuals of "white" ethnicity, as well as South Asian individuals and individuals of mixed ancestry. The lowest proportions of descent are observed in individuals of Chinese and African ancestry.

These findings support Cassidy et al. (2016), who showed a degree of population continuity between modern Irish populations and ancient steppe peoples. They partially support Galinsky et al. (2016), finding that steppe peoples have greater

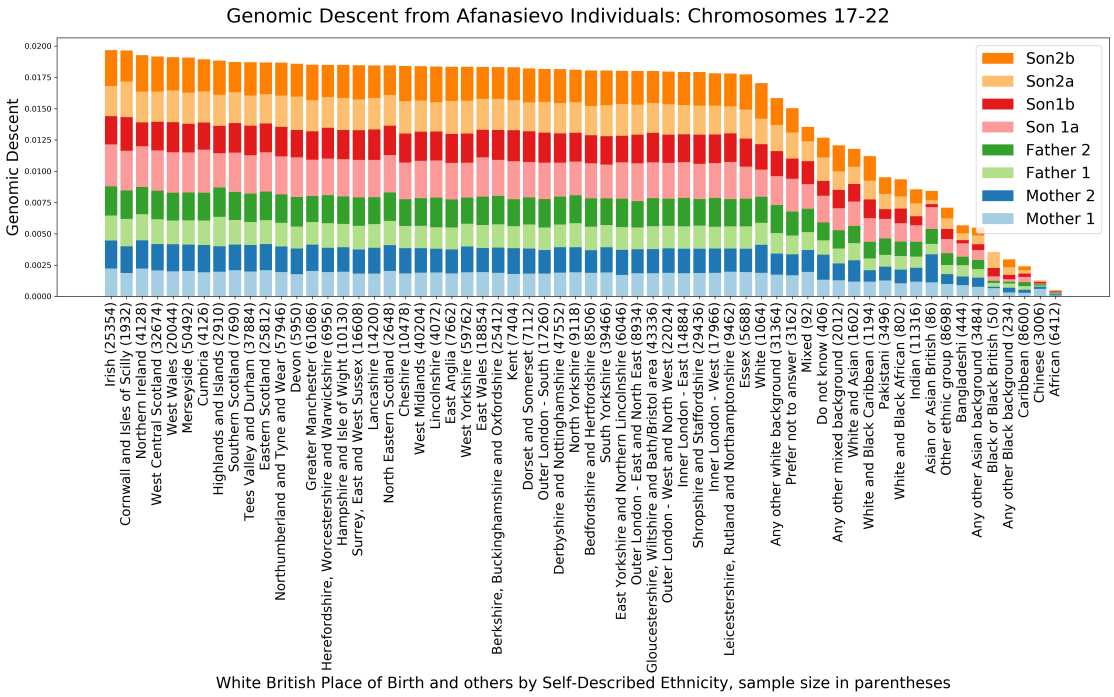


Figure 3.6: *Genomic descent* from Afanasievo family haplotypes among "white British" UK Biobank participants on Chromosomes 17-22. For UKB participants of self-described "white British" ethnicity who were born in the UK, reference sets were constructed based on the region of birth using NUTS 2 regions. All other participants were grouped by self-described ethnicity. Descent from each haplotype is denoted by colours in the stacked barchart. Results are averaged across chromosomes.

population affinity with Scottish than English individuals; however, they contradict the authors' finding that Welsh individuals do not have relatively elevated steppe ancestry.

3.4 Discussion

The statistics discussed in this chapter provide a straightforward means of understanding the population genetic relationships encoded in a tree sequence. Applying these statistics to tree sequences inferred from various genetic datasets provide results that are broadly consistent with commonly used methods such as PCA and f-statistics. The simplicity of *GNN* and *genomic descent* suggest that more sophisticated probabilistic approaches, potentially utilising information such as the distribution of time to most recent common ancestor (TMRCA) and haplotype

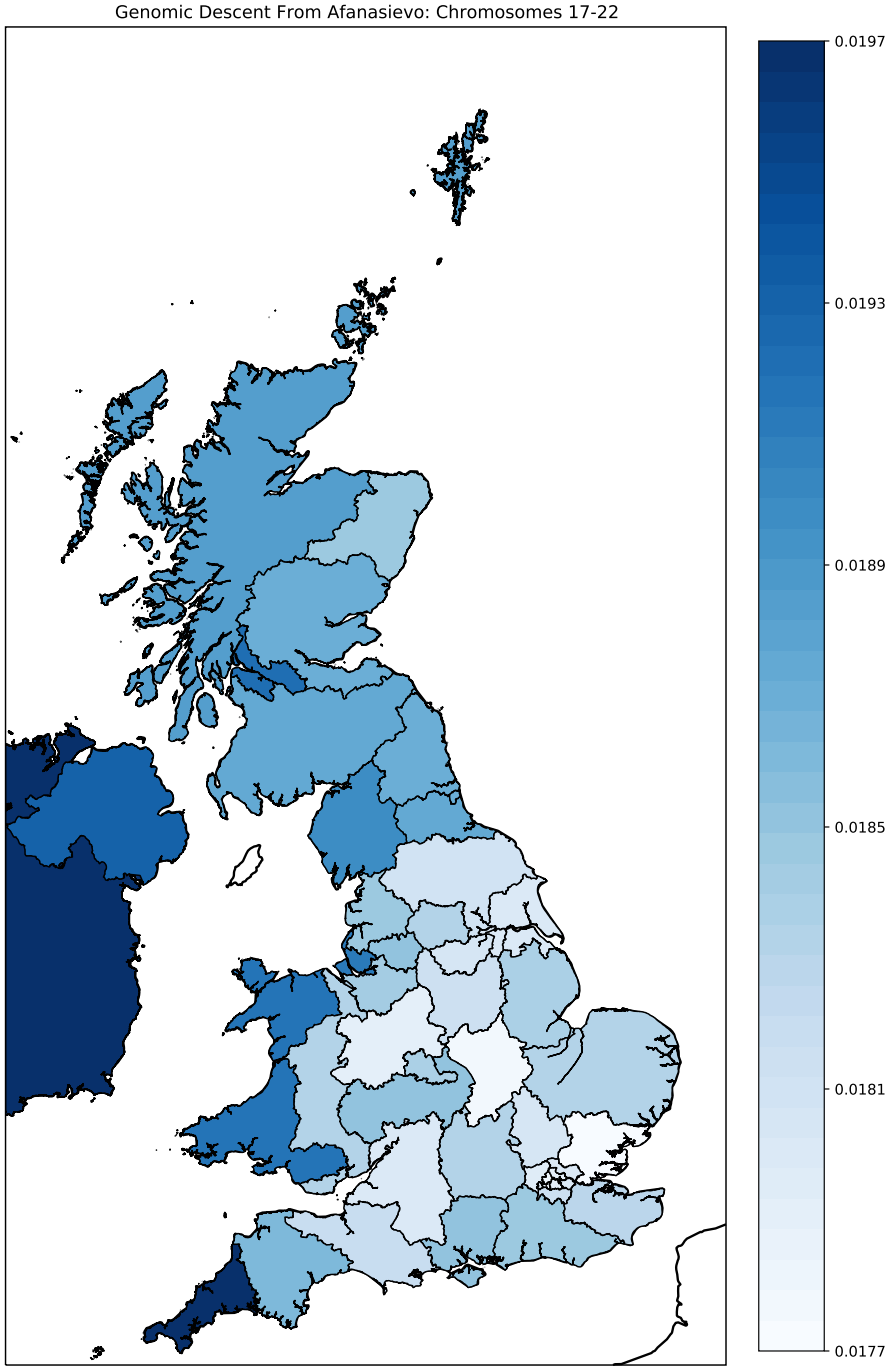


Figure 3.7: *Genomic descent* statistic from Afanasievo family haplotypes in NUTS 2 regions of the United Kingdom and in the Republic of Ireland. UK values are determined using place of birth information from participants of self-described white British ethnicity. The value for the Republic of Ireland is determined by using individuals of self-described Irish ethnicity.

lengths, will enable more powerful population genetic inference using tree sequences. Extensions incorporating these sources of information are already being explored. For instance, the *windowed-GNN* statistic offers insight into the age distribution of nearest neighbours. A simulation study of archaic introgression presented in Chapter 6 investigates shared ancestry between archaic and modern individuals using various TMRCA cutoffs, an approach which recovers more material related to archaic samples than the *genomic descent* statistic.

Multiple directions of further work would aid understanding of the signals produced by these methods. First, a comparison of commonly used population genetic statistics with the topological measures discussed here could shed light on unresolved questions in human prehistory. For instance, the relationship of the Sardinian and Basque populations to one another and to Neolithic individuals, such as the Trumpington brothers, is supported by both *genomic descent* and f-statistics, but multiple demographic scenarios could produce this signal (Chiang et al., 2018). Examining the results of these statistics in simulations could help distinguish between them. Second, simulation studies are needed to validate the performance of these novel methods in the presence of demographic processes, as well as genotype and phasing error. Finally, only point estimates of *genomic descent* are presented in this section. Incorporating a subsampling approach to estimate variance in levels of descent, as well as the length of copying tracts, should aid our understanding of the robustness of this statistic.

A more fundamental limitation of the work performed here is the inability to place ancient samples at their archaeologically-dated times in the tree sequence. In the next chapters I address this with an iterative approach for inferring dated tree sequences with modern and ancient samples.

4

Inferring dated tree sequences from modern and ancient samples

Contents

4.1 Introduction 77
4.2 `tsdate` 78
4.2.1 Conditional coalescent prior 79
4.2.2 Inside-Outside algorithm 82
4.2.3 Worked examples of `tsdate` 86
4.3 Iterative inference with ancient and modern samples . 92
4.4 Discussion 94

4.1 Introduction

I have so far considered the inference and analysis of tree sequence *topologies*. While the results of Chapter 3 show that much biological insight can be derived from these topologies, they do not encode a fundamental population genetic property: coalescence time. The distribution of coalescence times reflects key features of evolution such as historical population size (Li and Durbin, 2011) and is intimately tied to population structure (McVean, 2009) and linkage disequilibrium (McVean, 2002). If we knew the ages of internal nodes in a tree sequence, we could implement methods that use the tree sequence to infer processes such as

admixture, positive selection, and population bottlenecks, rather than search for their footprints in variation data (Harris, 2019; Ralph, Thornton, and Kelleher, 2020) (also see Section 7.1.3).

In this chapter I present `tsdate`, an approximate Bayesian method for efficiently inferring the age of coalescent events conditional on a tree sequence topology. I detail the various components of `tsdate`: the marginal prior on node age that uses the conditional coalescent introduced in Chapter 2, the inside-outside algorithm for propagating likelihoods up and down the tree sequence, and an alternative to the outside pass of the inside-outside algorithm that produces node age point estimates. I then provide simple, worked examples of `tsdate` on a tree and a tree sequence. Finally, I show how `tsinfer` is used together with `tsdate` in an iterative approach for inferring tree sequences from modern and ancient samples.

These methods were first introduced in Wohns et al. (2021). See joint authorship statement for details.

4.2 `tsdate`

We use the tree sequence topology estimated by `tsinfer` as the basis for an approximate Bayesian method to infer node age. This approach is implemented in the open-source software package `tsdate`.

Briefly, the algorithm proceeds as follows. First, a coalescent prior on the age of each node is constructed. A discretised time grid is then generated using the quantiles of these marginal prior distributions. A mutation-based likelihood function provides the likelihood of the length of edges in the tree sequence. With these pieces in place, information is propagated up and down the tree sequence using a belief propagation algorithm we call the "inside-outside algorithm". The result is an approximate posterior for the age of each node. An alternative step is also described which instead provides a point estimate for the age of each node via maximum a posteriori estimation. Note that while the equations in this section are in scaled time, in the algorithm implementation we use time expressed in generations. Note that `tsdate` does not depend on a binary tree sequence and

thus no special adjustments are required for polytomies in tree sequences inferred by `tsinfer` (see Kelleher et al. (2019)).

4.2.1 Conditional coalescent prior

The first step in the algorithm is to assign a prior distribution to the age of ancestral nodes in the tree sequence. A coalescent prior is an obvious choice (Kingman, 1982; Hudson, 1983b; Hudson, 1983a), with well-known expressions for the expectation and variance of the TMRCA of the sample (Donnelly and Tavaré, 1995; Tavaré et al., 1997; Wakeley, 2009):

$$\mathbb{E}[T_{MRCA}] = 2\left(1 - \frac{1}{n}\right); \tag{4.1}$$

$$\mathrm{Var}[T_{MRCA}] = 8\sum_{i=2}^{n}\frac{1}{i^2} - 4\left(1 - \frac{1}{n}\right)^2. \tag{4.2}$$

Additionally, both Tavaré (1984) and Takahata and Nei (1985) provide expressions for the probability distribution of the TMRCA.

For non-MRCA nodes the number of lineages remaining at any given time is unknown, so rather than use a fully tree sequence-aware prior we use an approximate approach based on assigning marginal priors to each node. Specifically, we use the number of modern samples descending from each node and the theoretical results of Wiuf and Donnelly (1999) to find a mean age and associated variance under the coalescent.

As in Chapter 2, we begin by considering a coalescent tree with n sequences, partitioned into i and $n - i$. We condition on the event E that all samples in i coalesce before sharing a common ancestor with a sample outside of i. Conditioning on E affects the order of coalescence events, but not the exponentially distributed waiting times. We seek to find the mean and variance of the time t, when all i samples have coalesced. At this time there will be only one ancestor of i remaining, but depending on how many lineages outside of i have coalesced, the number of total ancestors, α, could range from 2 to $n - i + 1$.

Equation (10) in Wiuf and Donnelly (1999) shows that the expectation of t is

$$\mathbb{E}[t] = \frac{i-1}{n}. \tag{4.3}$$

The variance can be found using Equation (4.3) and the expectation of t^2, which is calculated as

$$\mathbb{E}[t^2] = \sum_{m=2}^{n-i+1} \mathbb{P}(\alpha = m)\mathbb{E}[t^2|\alpha = m], \tag{4.4}$$

where $\mathbb{E}[t^2|\alpha]$ is given by Equation 10 in Wiuf & Donnelly (1999) as

$$\mathbb{E}[t^2|\alpha] = 8\sum_{j=\alpha}^{n} \frac{1}{j^2} + \frac{8}{n} - \frac{8}{\alpha} - \frac{8}{n\alpha}, \tag{4.5}$$

and $\mathbb{P}(\alpha = m)$ is Corollary 2 of Wiuf and Donnelly (1999) when only one ancestor in i remains:

$$\mathbb{P}(\alpha = m) = \frac{\binom{n-m-1}{i-2}\binom{m}{2}}{\binom{n}{i+1}}. \tag{4.6}$$

With a mean and variance of the age of each node under the coalescent, we next use the method of moments to fit a lognormal distribution for each marginal prior.

$$\sigma = \sqrt{log\left(\frac{\mathrm{Var}[t]}{\mathbb{E}[t]^2} + 1\right)}; \tag{4.7}$$

$$\mu = log(\mathbb{E}[t]) - \frac{1}{2}\sigma^2. \tag{4.8}$$

This lognormal distribution provides a prior, π_u, for the age of each node u in the tree sequence. For internal consistency, we use a matched lognormal distribution for the age of the TMRCA rather than the previously mentioned results from Tavaré (1984) and Takahata and Nei (1985). Although the prior is valid at any given marginal tree topology in the tree sequence, recombination introduces two complications: nodes may not span the full tree sequence and a given node may have descendant subsamples of differing sizes in the various marginal trees in which it appears. To address these points, we first iterate through each marginal tree in the tree sequence to find the total length of sequence *spanned* by node u, which we will refer to as s_u. We then determine the "sample weights" for each node by

finding the span of sequence where u has a particular number of descendent samples as a proportion of s_u. Since neighbouring marginal trees in a tree sequence are correlated (see Section 1.3), we use the difference in edges between each adjacent tree as an efficient method for calculating the sample weights.

Once we have evaluated sample weights for each node in the tree sequence, we construct a mixture distribution for the prior on node age. This is accomplished by finding the weighted average of the expectation and variance of the prior distribution using the sample weights. With the weighted means and variances, we again use the method of moments to determine the parameters of the matched lognormal distribution.

In the next chapter, we assess how well the gamma and lognormal distributions represent the age of nodes simulated under the coalescent. Figure 5.1 and the accompanying analysis shows that the lognormal approximation is better calibrated in simulations both with and without recombination. Note that another option, which I did not explore in this work, is to use discretised empirical distribution as a prior.

Time Discretisation

Our inference approach uses a time grid. While users may specify their own grid, by default we generate an "adaptive" time grid based on the marginal prior distribution of each node. This approach outperforms a naive, uniform grid, as inference is focused on times with greater probability under the prior.

The adaptive time grid is constructed by taking the union of quantiles from each ancestral node's prior distribution. The density of the time grid is determined by the user-specified number of quantiles (the default is 20). We use the quantile function corresponding to the conditional coalescent prior π_u of each node u. Many of the quantiles calculated at multiple nodes are similar or identical, so the resulting time grid is "thinned" based on another user-specified parameter, ϵ, that establishes the minimum time distance between points in the grid. This value defaults to 10^{-6}.

4.2.2 Inside-Outside algorithm

With a prior in place for ancestral nodes in the tree sequence and a time grid, we infer the age of nodes using a belief propagation approach we call the inside-outside algorithm, based on an HMM where the age of nodes are hidden states. In the case of a single tree, this equates to the standard forward-backward algorithm. In the case of a tree sequence, we must also consider the relative genomic spans associated with edges and address cycles in the undirected graph underlying the tree sequence. Cycles occur whenever a node has multiple parents and present a general problem in belief propagation (Murphy, Weiss, and Jordan, 2013).

The algorithm is efficient because it uses dynamic programming and the tree sequence traversal methods implemented in `tskit`, the tree sequence toolkit. Scaling is linear with the number of edges in the tree sequence (Figure 5.7) and quadratic with the number of time slices used.

Inside pass

We seek to compute all values in the inside matrix I for all nodes and times in the discretised time grid. $I_u(t)$ is the probability of node u at time t, which encompasses the probability of all nodes and edges in the subgraph beneath u.

We initialise the prior probability of a sample node to be one at its sampled time and zero elsewhere. We then proceed backwards in time, using the relationships between nodes encoded in the `tskit` edge table until we reach the most recent common ancestor (MRCA) nodes of the tree sequence. For each node, we visit every child as well as every time t in the time grid using

$$I_u(t) = \pi_u(t) \prod_{d \in C(u)} \sum_{t' \leq t} L_{du}(t - t' + \epsilon; D_{du}, \theta) I_d(t')^{w_{du}},$$

where $C(u)$ is the set of all child nodes of u, and as previously defined, $\pi_u(t)$ is the prior probability of node u at time t; $L_{du}(t - t' + \epsilon; D_{du}, \theta)$ is the mutation-based likelihood function of the edge from focal node u at time t to child node d at time t'; ϵ is an arbitrarily small value that is used to prevent numerical problems arising when parent and child nodes exist in the same time slice; D_{du} is the data associated

with the edge including the span of the edge and the number of mutations on it; θ is the population-scaled mutation rate; and w_{du} is the span of the edge leading from u to d divided by s_d, the total span of node d in the tree sequence. Note that the inside probability of node d is geometrically scaled by w_{du} to address overcounting if d has multiple parents.

The likelihood function gives the probability of observing k mutations on an edge of length $\delta t = t - t' + \epsilon$ with span l_{du}. It is Poisson distributed with parameter $(\theta l_{du} \delta t)/2$

$$L_{du}(\delta t; D_{du}, \theta) = \frac{\left(\frac{\theta l_{du} \delta t}{2}\right)^k}{k!} e^{\frac{-\theta l_{du} \delta t}{2}}. \tag{4.9}$$

We can factorise the inside probability as

$$I_u(t) = \pi_u(t) G_u(t),$$

where $G_u(t)$ is

$$\prod_{d \in C(u)} g_{du}(t)$$

and $g_{du}(t)$ is

$$\sum_{t' \leq t} L_{du}(t - t' + \epsilon; D_{du}, \theta) I_d(t')^{w_{du}}.$$

This factorisation will be useful in describing the outside pass in the next section.

The equation terminates at the MRCA(s) of the tree sequence. The total likelihood of the tree sequence is obtained by taking the product of the inside matrix of each MRCA.

Outside pass

Once we have iterated up the tree sequence to fill in the inside matrix, the inside probability of the MRCAs contains all of the information encoded in the tree sequence. To find the full posterior on node age, we now take account of the information in the tree sequence "outside" of the subgraph of each ancestral node. While this algorithm empirically performs well with a single inside and outside pass,

any cycles in the underlying undirected graph (which occur when recombination causes a node to have more than one parent) will result in overcounting. The alternative "outside-maximisation" pass introduced in the next section provides another approximate solution in these cases, though we find that the outside pass performs better empirically (see Figure 5.2).

Beginning with the MRCA nodes in each marginal tree (the roots), we initialise the outside value of these nodes, O_{MRCAs}, to be one at all non-zero time slices. There is no information "outside" the MRCAs because all information in the tree has already been propagated to the node and is encoded in the MRCAs' inside matrices. In a tree sequence, it is possible for a node u to be the MRCA in some of the marginal trees in which it appears but not in other trees. In these cases we find O_u by dividing the span of trees where the node is the MRCA by s_u, the total span of u in the tree sequence.

We then proceed down the tree sequence (forwards in time), again using the edge table sorted in descending order by the children's ages. At every node we calculate:

$$O_u(t) = \prod_{p \in P(u)} \sum_{t' \geq t} O_p(t')^{w_{up}} L_{pu}(t' - t + \epsilon; D_{pu}, \theta) \left(\frac{I_p(t')}{g_{up}(t')} \right)^{w_{up}}, \tag{4.10}$$

where $P(u)$ is the set of parents of node u and other terms are defined in the previous section on the inside algorithm.

Once the inside and outside passes are complete, the approximate posterior can be calculated as

$$\phi_u(t) \propto I_u(t) O_u(t). \tag{4.11}$$

Importantly, the mean value of the posterior distribution may not be consistent with the tree sequence topology. We provide the option to "constrain" node age estimates by forcing each node to be older than the estimated age of its children. The unconstrained mean and variance of each node are retained as metadata in the tree sequence. The full posterior can also be retained separately if desired.

The oldest root in tree sequences inferred by `tsinfer` is the ancestral haplotype (carrying the ancestral allele at all variable sites). We observe that mutations

mapping to edges descending from this ancestral haplotype are generally of lower quality, so in our implementation of the outside pass we include an option to avoid traversing such edges. We use this setting in all further analyses, and thus do not include age estimates for mutations appearing on these edges.

Outside-Maximisation Pass

Cycles in the undirected graph underlying a tree sequence occur when a node has more than one parent as a consequence of recombination, (an example is shown in Figure 4.5). These cycles result in the over-counting of information in the outside pass of the inside-outside algorithm. To account for this, we introduce the outside-maximisation pass as an alternative. The rationale for this approach is that the inside value of the MRCA(s) of the sample has accumulated all of the data encoded in the tree sequence. In our model, the age of a node and the age of all non-descendant nodes are conditionally independent, given the age of a node's parents. Thus, if we fix the age of the MRCA(s), we can then walk down the tree sequence fixing the age of parent nodes before considering the age of their children.

First, we set the age of each MRCA in the tree sequence to the time slice with the maximum probability in the node's inside matrix (this is equivalent to finding the maximum posterior probability for MRCA nodes). Next, we establish a traversal path using the topology such that each node is only visited once all of its direct and indirect parents have been considered. For each node we visit on this traversal path, we know the age of all of a node's ancestors as well as the relative likelihood of the subtree below each node for a given age. For any possible node age in the discretised time grid we can also compute the likelihood of the events on the branches to its direct and already-dated ancestors, as described in the inside pass. With this information we can compute the conditional posterior density for the age of the node. We set its age to the time slice with the maximum value.

Formally, for each node u we seek to calculate the time of the node, T_u as

$$T_u = \underset{t \leq U_p}{\arg\max} \left(I_u(t) \prod_{p \in P(u)} L_{pu}(T_p - t + \epsilon; D_{pu}, \theta) \right), \tag{4.12}$$

where U_P is the age of the youngest parent of node u and other variables are defined in the previous sections on the inside and outside passes.

4.2.3 Worked examples of `tsdate`

In this section I present simple, numerical examples of `tsdate` applied first to a tree and then to a tree sequence. I begin with the single tree example, illustrated in Figure 4.1a.

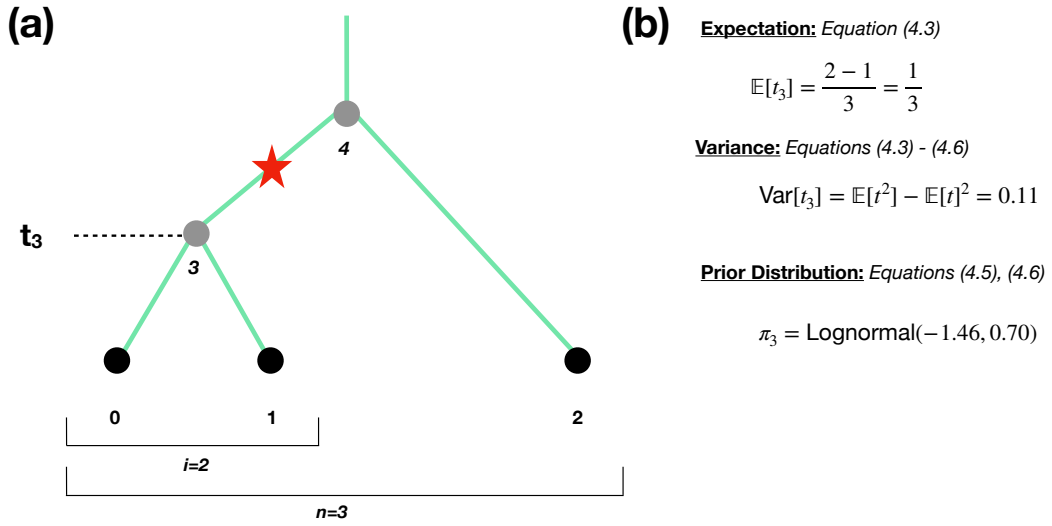


Figure 4.1: (a) Example tree with $n = 3$ samples and one mutation, indicated by the five-pointed star. t_3, the age of node 3, is indicated. This node has two descendants, $i = 2$, out of three total samples. (b) Calculating π_3, the marginal prior for node 3 that is determined by the values of n and i. The expectation, variance, and matched lognormal distribution for this node are shown alongside the equations used to find each value.

Node Age Prior

We first compute the marginal prior distribution for each node in the tree. Nodes 0-2 are fixed at time 0, so we begin by determining π_3, the prior distribution of node 3. This calculation uses equations from Section 4.2.1 and is shown in Figure 4.1b. Note that the only information we need to find π_3 is the number of descendants of node 3 and the total number of samples.

Node 4 is the MRCA of the sample, so we use Equations (4.1) and (4.2) to find $\mathbb{E}[t_4] = 1.33$ and $\text{Var}[t_4] = 1.11$. We again use moment matching to find π_4. Figure 4.2 shows the resulting distributions of π_3 and π_4.

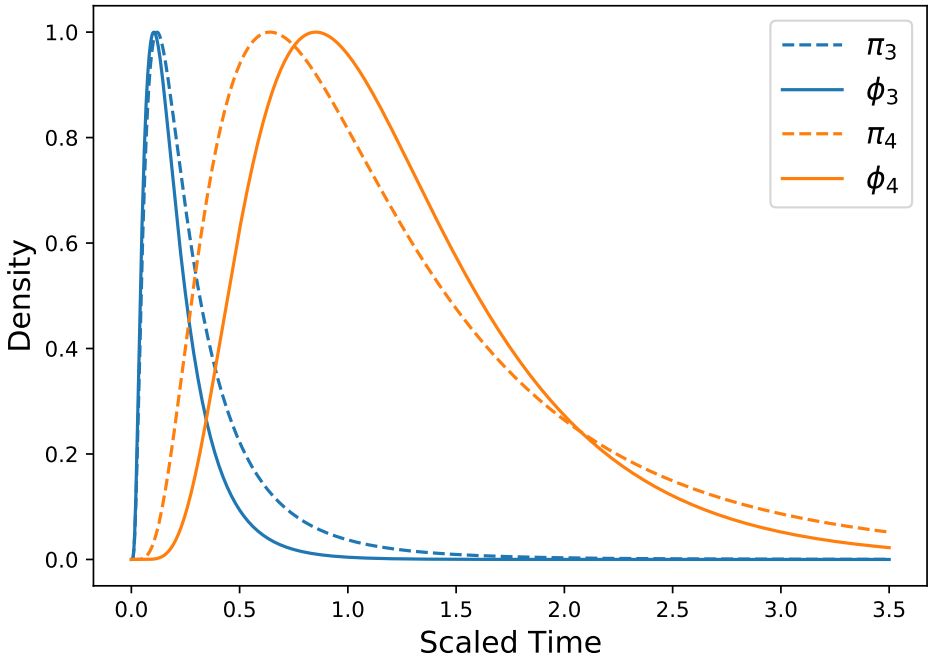


Figure 4.2: Prior and posterior distribution for nodes 3 and 4 from tree example shown in Figure 4.1a. To determine the priors, π_3 and π_4, we use moment matching to fit a lognormal distribution to the expectation and variance of node age under the conditional coalescent. The posteriors, ϕ_3 and ϕ_4, are determined by multiplying the inside and outside matrices together. I used $\theta = 1$ to compute the likelihoods. Note that these results were calculated using a dense, uniform grid, while the example uses a simple grid for ease of exposition. The probability densities are normalised by their maximum value.

Discretised Time Grid

Next, we generate an "adaptive" time grid with quantiles of the marginal prior distributions of the two internal nodes. In our example tree, we first calculate quantiles from node 3 using the quantile function for π_3. We then perform the same operation for node 4; however, quantiles are only added to the time grid if the magnitude of their difference from the closest existing points in the grid is greater than ϵ.

Here, we proceed with a time grid of $[0, 1, 2]$ to simplify the exposition of the following steps of the algorithm.

Inside Algorithm

With a prior for internal nodes and a time grid, we now use the inside pass of the inside-outside algorithm to propagate likelihoods up the tree. This is accomplished by filling out the "inside matrix" for the tree, where the number of rows is the size of the time grid and the number of columns is the number of nodes. Each entry in the matrix is calculated using Equation (4.9). The results are shown in Figure 4.3, with the calculation of $I_4(2)$ highlighted in red both graphically and symbolically.

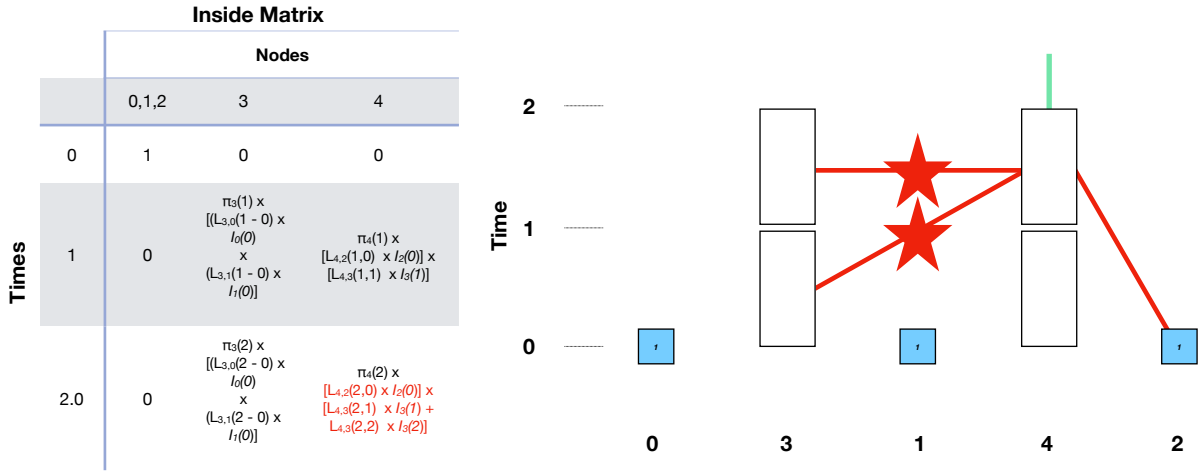


Figure 4.3: Filling in the Inside Matrix, with the calculation of $I_4(2)$ highlighted. The inside matrix is shown on the left, the schematic tree on the right. Rectangles indicate the time slices at which node 3 and 4 can exist. Likelihood terms coloured red in the inside matrix refer to red-coloured edges in the schematic tree.

Outside Algorithm

When the inside pass terminates at the MRCA of the tree, we proceed with the outside algorithm by filling in an "outside matrix" for all nodes in the tree at all time slices in the grid. Beginning at the MRCA, we know that node 4 cannot exist at time zero and there is no information "outside" of the node, so O_4 is one at time 0 and zero at all other time slices. At node 3 we seek to incorporate information from "outside" of the subtree of node 3 and its children using Equation (4.10).

The completed outside matrix is shown in Figure 4.4, with the calculation of $O_3(1)$ highlighted in red both graphically and symbolically.

After the outside pass terminates at the leaf nodes, the posterior distribution of the age of each node $\phi_u(t)$, can be calculated by multiplying the inside and outside matrices, as in Equation (4.11). The posterior distributions are shown in Figure 4.2.

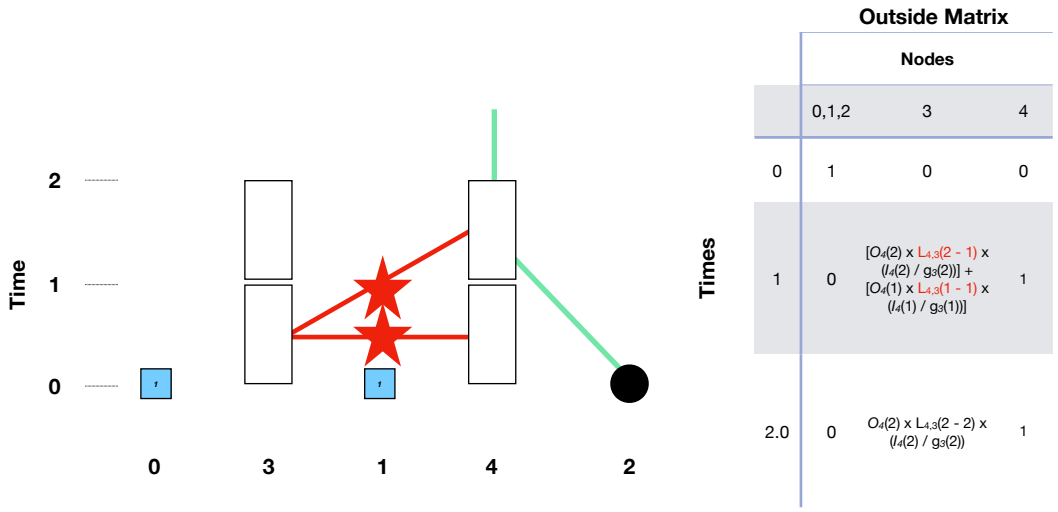
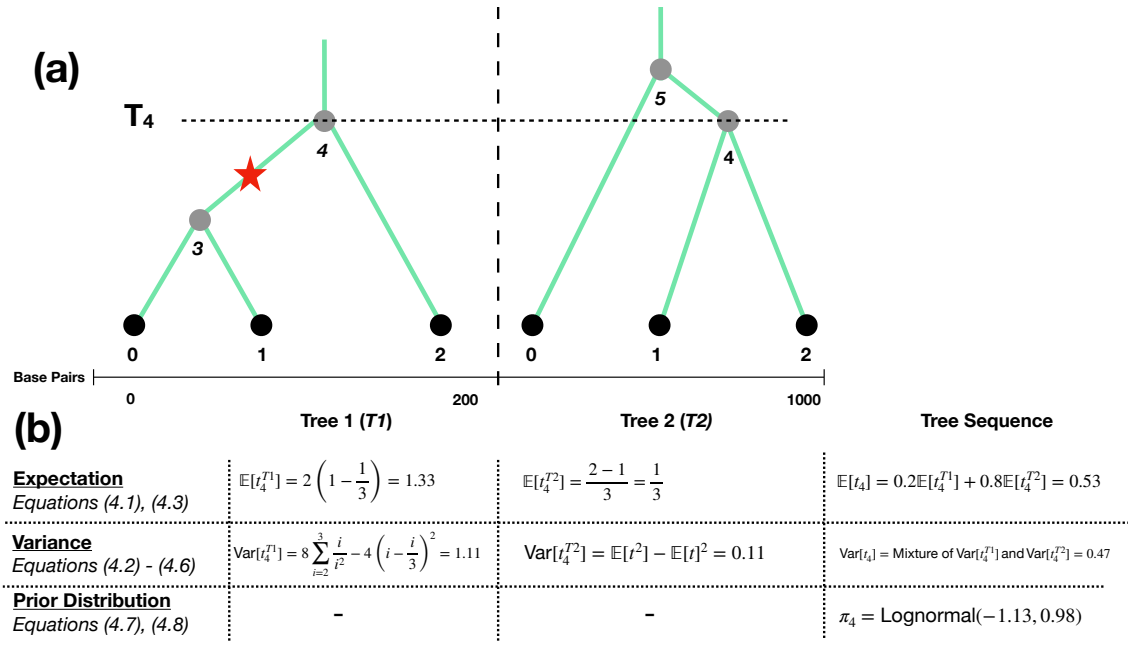


Figure 4.4: Filling in the Outside Matrix, with the calculation of $O_3(1)$ highlighted. The outside matrix is shown on the right, the schematic tree on the left. Rectangles indicate the time slices at which node 3 and 4 are evaluated. Likelihood terms coloured red in the inside matrix refer to red-coloured edges in the schematic tree.

Tree Sequence Adjustments

The `tsdate` algorithm generalises to a tree sequence by adjusting the marginal prior to account for nodes with variable numbers of leaves in different marginal trees and weighting the propagation of likelihoods in the inside and outside algorithms by edge spans. To demonstrate these features of the algorithm, I provide a worked example using the tree sequence shown in Figure 4.5a. In this tree sequence, the marginal tree accounting for the first 200 base pairs is the same as Figure 4.1. A recombination event at position 200 creates a new marginal tree describing the remaining 800 base pairs.

We begin by finding the marginal prior distributions for the internal nodes. Since nodes 3 and 5 only exists in a single tree, their prior distributions are found



(b)

	Tree 1 (*T1*)	Tree 2 (*T2*)	Tree Sequence
Expectation *Equations (4.1), (4.3)*	$\mathbb{E}[t_4^{T1}] = 2\left(1 - \frac{1}{3}\right) = 1.33$	$\mathbb{E}[t_4^{T2}] = \frac{2-1}{3} = \frac{1}{3}$	$\mathbb{E}[t_4] = 0.2\mathbb{E}[t_4^{T1}] + 0.8\mathbb{E}[t_4^{T2}] = 0.53$
Variance *Equations (4.2) - (4.6)*	$\mathrm{Var}[t_4^{T1}] = 8\sum_{i=2}^{3}\frac{i}{i^2} - 4\left(i - \frac{i}{3}\right)^2 = 1.11$	$\mathrm{Var}[t_4^{T2}] = \mathbb{E}[t^2] - \mathbb{E}[t]^2 = 0.11$	$\mathrm{Var}[t_4]$ = Mixture of $\mathrm{Var}[t_4^{T1}]$ and $\mathrm{Var}[t_4^{T2}] = 0.47$
Prior Distribution *Equations (4.7), (4.8)*	-	-	$\pi_4 = \mathrm{Lognormal}(-1.13, 0.98)$

Figure 4.5: (a) Example tree sequence with $n = 3$ samples, one mutation, and two marginal trees. Node 4 has three leaf nodes in the first marginal tree and two in the second. Nodes 0, 1 and 4 have multiple parents. A single recombination event at position 200 explains the difference between these marginal trees. This recombination event prunes the edge connecting nodes 0 and 3 and reattaches node 0 to node 5, the latter of which does not exist in the first tree. (b) Calculating π_4, the marginal prior for node 4 using the weighted average of the expectation and variance of the age of the node calculated at each marginal tree.

in the same manner as the single tree example. The calculation of π_4 is shown in Figure 4.5b. Node 4 exists in both marginal trees, so its total span, S_4, is 1000. In the first marginal tree it has three leaf nodes; in the second tree it has two. We thus compute the expectation and variance of a node with two leaf nodes and a node with three leaf nodes. We then find the weighted average of the expectation and variance with 20% weight given to three leaves and 80% weight to two leaves. These prior distributions can be used to find an adaptive time grid as in the single tree example. Figure 4.6 shows π_3, π_4, and π_5.

The traversal rules of the inside-outside algorithm apply to a tree sequence as well as a marginal tree, but the cycles induced by recombination mean that our inference approach is no longer exact. Geometric scaling is an approximate solution to the overcounting caused by cycles in the tree sequence. We geometrically scale the likelihoods of edges in the inside-outside algorithm using "edge weights". In

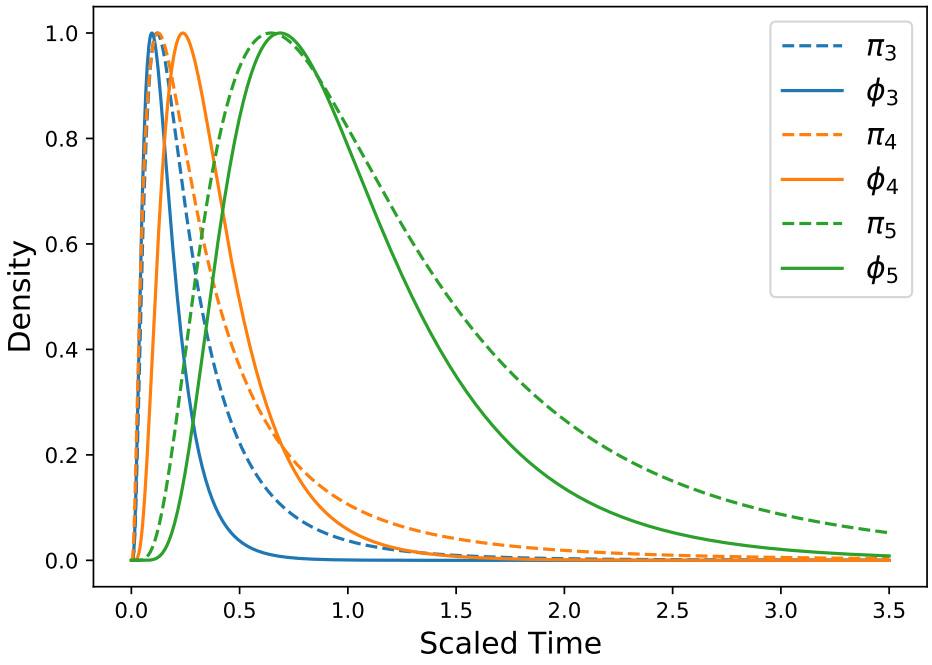


Figure 4.6: Prior and posterior distributions for nodes 3, 4, and 5 from the example tree sequence shown in Figure 4.5a. These were computed and plotted as in 4.2, see caption for details.

Equation (4.9) this is indicated by w_{du}, which is calculated as the span of the edge connecting nodes d and u divided by s_d, the total span of node d in a tree sequence. When determining I_3 in our example, both w_{03} and w_{13} are equal to 0.2. This is because s_0 and s_1 both equal 1000 and the span of the first tree is 200. Similarly, w_{05} and w_{14} are equal to 0.8. In the outside algorithm, scaling affects the edges where node 4 is the parent. Node 4 exists in both trees, so s_4 equal to 1000. w_{34} is then 0.2, and w_{14} is 0.8, since the edge between nodes 3 to 4 exists in the first tree and the edge between nodes 1 and 4 exists in the second tree.

After computing the inside and outside matrices, the approximate posterior distribution can be found as before using Equation (4.11). Note that the outside pass was used in both the single tree and tree sequence examples. If the outside maximisation pass is used, a point estimate of the age of each node is returned instead of a posterior distribution on node age.

4.3 Iterative inference with ancient and modern samples

We combined `tsinfer` version 0.2 and `tsdate` in an iterative approach that allows for the incorporation of ancient samples and improves inference accuracy in many settings.

The first step of the iterative approach is to order derived alleles appearing in the sample by their frequency. As described in Chapter 2, `tsinfer` uses allele frequency as a relative ordering of derived alleles to both build ancestral haplotypes and infer copying paths. Once alleles are ordered, it is possible to infer a tree sequence topology with `tsinfer` (Figure 4.7 Step 1).

With an inferred tree sequence topology, we next estimate the age of inferred ancestral haplotypes with `tsdate` (Figure 4.7 Step 2). If using `tsdate`'s outside pass, we do not constrain the resulting date estimates by the topology.

If ancient samples are present, we can use them to constrain the estimated age of derived alleles. The previous step (Figure 4.7 Step 2) provides date estimates for the inferred ancestors as well as for mutations. Since we estimate the age of the ancestral nodes above and below a mutation, the child node of an edge hosting a mutation is constrained by the ancient sample-informed lower bound on derived allele age. This bound is determined by gathering the haplotypes of ancient samples (either sequenced or genotyped) and examining derived alleles that can be called in these ancient samples with high confidence. If multiple ancient samples carry the same derived allele, we use the oldest sample as the lower bound on its age. Once lower bounds have been collected for all derived alleles observed in ancient samples, we compare these with our statistically inferred lower bounds on allele age, adjusting our age estimates where necessary to ensure consistency with ancient samples. Any radiocarbon-dated ancient samples with high-confidence variant calls may provide constraints in this step, including unphased and/or low-coverage samples. Although a substantial fraction of radiocarbon dates are likely inaccurate on the order of hundreds to a few thousand years, as indicated by recent re-evaluations (Piličiauskas et al., 2017; Plicht et al., 2020), we note that there is a

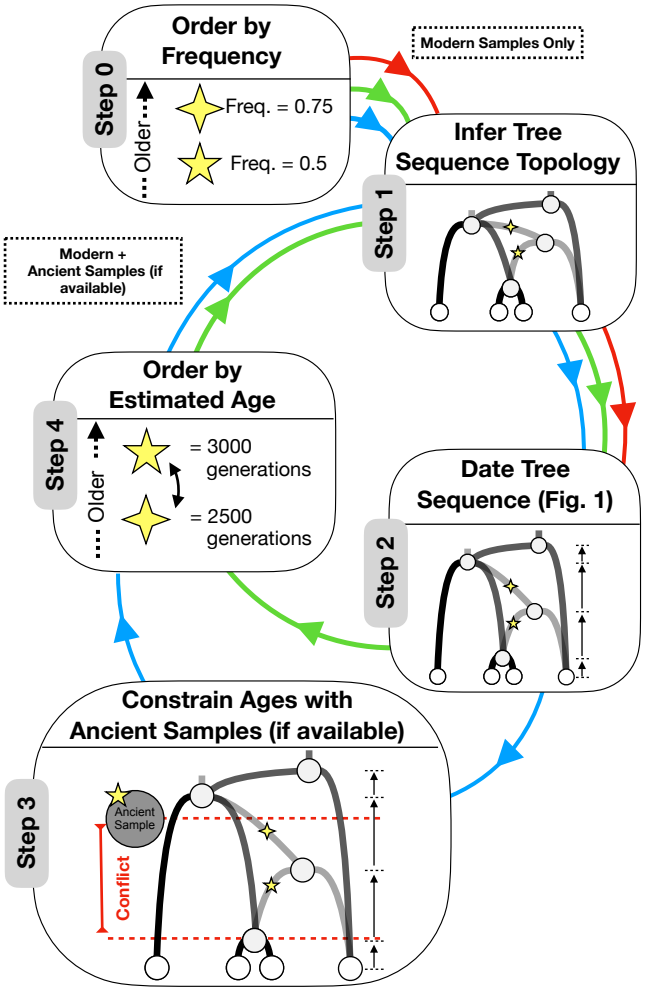


Figure 4.7: Schematic overview of iterative approach. Step 0: alleles are ordered by frequency; the mutation represented by the four-point star is thus considered to be older than that represented by the five-point star. Step 1: the tree sequence topology is inferred with `tsinfer` using modern samples. Step 2: the tree sequence is dated with `tsdate`. Step 3: node date estimates are constrained with the known age of ancient samples. Step 4: ancestral haplotypes are reordered by the estimated age of their focal mutation; the five pointed star mutation is now inferred to be older than the four-point star mutation. The algorithm returns to Step 1 to re-infer the tree sequence topology with ancient samples. Arrows refer to modes of operation: Steps 0, 1 and 2 only (red); after one round of iteration without ancient samples (green) and after one round of iteration with increasing numbers of ancient samples (blue).

low probability that errors of this magnitude will meaningfully affect tree sequences inferred using this approach. Only a subset of erroneously dated alleles will be older than the true age of the mutation, which would affect allele age estimation accuracy, and still fewer will be older than the ancestral haplotype from which they

descend, which would affect topological estimation accuracy.

With allele age estimates from step 2, possibly constrained by ancient samples in step 3, we are now able to re-infer the tree sequence topology. The revised age estimates are used to order the age of derived alleles when re-estimating ancestral haplotypes with `tsinfer`; if they are more accurate than frequency in determining a relative ordering of mutations, topological inference accuracy should be improved. Indeed, we find that the iterative approach improves accuracy when re-inferring tree sequences from variation data simulated with a uniform recombination map and without error (Figure 5.8). When reinferring tree sequences from data simulated with error or with a variable recombination map, less improvement is observed (Figure 5.5).

Ancient samples can be included in tree sequences that are (re)-inferred with estimated allele ages. This is accomplished by inserting ancient samples at their correct relative ordering among ancestors generated by `tsinfer`. Only phased ancient samples with an age estimate may be included, although we note that extending `tsinfer`'s HMM to handle diploid individuals may allow for phasing of ancient samples in this step. In `tsinfer`, samples cannot directly descend from other samples (which is highly unlikely in reality). Instead, we create "proxy" ancestral haplotypes, which at non-singleton sites are identical to each ancient sample, and insert them at a slightly older time than the ancient sample. These proxies can act as ancestors of both modern and (younger) ancient samples. Thus, the ancient samples are never themselves direct ancestors of younger ones, but their haplotypes may be. Finally, we infer copying paths between ancestors and samples to produce a tree sequence of modern and ancient samples.

4.4 Discussion

In this chapter, I present `tsdate`, an approximate Bayesian method for inferring the age of coalescent nodes conditional on a tree sequence topology. I also show how this method can be used iteratively with `tsinfer` to infer dated tree sequences from modern and ancient samples.

While the next chapter will show that `tsdate` and the iterative approach are both accurate and efficient, various aspects of these methods may be improved or enhanced. First, `tsdate` only provides an approximate solution to the cycles inherent in genealogical histories, though I show in the next chapter that the algorithm still performs well empirically in the presence of recombination. The literature on loopy belief propagation offers multiple possible solutions to this challenge (Murphy, Weiss, and Jordan, 2013). Second, the conditional coalescent-informed prior does not currently consider the archaeological age of ancient samples, which complicates the assignment of marginal prior distributions to internal nodes that are ancestors of ancient samples. If a suitable prior could be found for such internal nodes, it would be possible to both date tree sequences with ancient samples and provide genetically-informed estimates of the age of ancient samples themselves. Another potential enhancement to `tsdate` is support for variable N_e and mutation rates: the former over time and the latter both across the genome and over time. Finally, use of a recombination-based likelihood function is complicated by the difficulty of assigning recombination events to specific branches, as well as by the underestimation of the number of events. However, Deng, Song, and Nielsen (2020) show that `tsinfer` compares favourably to other inference approaches in the number of recombination breakpoints recovered, and this is further aided by a correction factor that the authors introduce. This suggests that it could be possible to use a recombination-based likelihood function if we account for the expected undercounting of recombination events in the inferred tree sequence. Additionally, it should be possible to use a recombination-based likelihood at the subset of ancestral haplotypes with more certain breakpoints produced via the path compression step of `tsinfer` (see methods of Kelleher et al. (2019)).

In the next chapter, I evaluate the accuracy of tree sequence and allele age inference using the latest implementations of `tsinfer` and `tsdate`. I also assess the effects of incorporating ancient samples, as well as our ability to recover patterns of introgression from archaic samples. In Chapter 6, I apply these methods to

infer the most comprehensive reference genealogy of modern and ancient samples constructed to date.

5

Evaluating the accuracy of dated tree sequences

Contents

5.1 Introduction 97
5.2 Evaluating `tsdate` 98
5.2.1 Accuracy of the `tsdate` prior 99
5.2.2 Accuracy of age estimation 99
5.2.3 Computational efficiency 104
5.3 Evaluating the accuracy of incorporating ancient samples 107
5.3.1 Ancient samples and allele age estimation accuracy . . . 107
5.3.2 Archaic descent simulation 109
5.4 Empirical estimates of 1000 Genomes Project variant ages on chromosome 20 115
5.4.1 Materials and Methods 115
5.4.2 Results 116
5.5 Discussion 117

5.1 Introduction

In this chapter, I assess the accuracy and scaling properties of `tsinfer` and `tsdate`. I performed simulation-based evaluation using `msprime`, incorporating realistic models of genotype and ancestral state error as well as various demographic models

implemented in stdpopsim (Kelleher, Etheridge, and McVean, 2016; Adrion et al., 2020). In addition, I used age constraints obtained from thousands of genotyped and sequenced ancient samples as the basis of a novel, empirical evaluation of the accuracy of TGP allele age estimates from tsdate and two comparable methods.

I also assess the iterative approach introduced in Section 4.3. I first used simulations to evaluate how constraints introduced by ancient samples can improve allele age estimation accuracy. I then used a model of human history with archaic admixture from Neanderthals and Denisovans to investigate how tree sequences inferred with this approach can be used to recover introgressed archaic haplotypes.

The results of this chapter show that these methods infer genome-wide genealogies and allele ages with unmatched scaling properties and accuracy comparable, or superior, to the state of the art. In the concluding section, I note potential future directions for improving these methods based on the results of the evaluations.

All text and figures in this chapter are adapted from Wohns et al. (2021) and represent my work.

5.2 Evaluating tsdate

In this section, I evaluate the accuracy and scaling properties of tsdate using msprime simulations (Kelleher, Etheridge, and McVean, 2016). I first assess how well the conditional coalescent prior captures the true age of simulated internal nodes. Next, I evaluate simulations performed both under the neutral coalescent and using a demographic model approximating the "Out of Africa Event" (Gutenkunst et al., 2009) implemented in stdpopsim. tsdate was run on the simulated tree sequence topologies, as well as on topologies inferred by tsinfer. In the latter case, topologies were inferred using the true, simulated variation data, as well as variation data injected with genotype and ancestral state error. I assess the accuracy of the resulting dated tree sequences and allele age estimates using a variety of measures. Finally, I ascertain the runtime and memory requirements of these methods with increasing sample size and sequence length. The inference accuracy and computational requirements of tsdate and tsinfer are also compared

to `Relate` (Speidel et al., 2019), a leading method for inferring dated genealogies and `GEVA` (Albers and McVean, 2020), a method for estimating allele age based on pairwise haplotype comparisons.

5.2.1 Accuracy of the `tsdate` prior

As described in Section 4.2.1, `tsdate` uses results from the coalescent conditioned on the frequency of an ancestral haplotype (Wiuf and Donnelly, 1999) to determine the mean and variance of the age of each node in the tree sequence. Moment matching is then used to fit a lognormal distribution on the age of each node. Figure 5.1 shows evaluations of the accuracy of this prior distribution, and also an alternative using a gamma distribution, in simulations with and without recombination. The results of `msprime` simulations under the neutral coalescent show that the 95% credible interval of the gamma prior includes the true node ages more frequently than the equivalent lognormal distribution. However, the credible interval of the gamma is much wider than the lognormal and skews towards younger ages. For this reason, the lognormal is the default setting in `tsdate`.

5.2.2 Accuracy of age estimation

Simulation-based evaluation of the accuracy and scaling properties of `tsinfer` and `tsdate` were performed with `msprime` version 1.0, (Kelleher, Etheridge, and McVean, 2016) and `stdpopsim` (Adrion et al., 2020). I evaluate neutral coalescent simulations with recombination as well as a more complex, three population "Out of Africa" model (Gutenkunst et al., 2009). The latter model is implemented in `stdpopsim` version 0.1.2. The effects of genotype and ancestral state errors were modelled using an error model based on an empirical comparison of the Illumina Platinum Genomes Project to matched data from the TGP (Albers and McVean, 2020).

`tsdate` was evaluated under a variety of inference settings. First, the "true" topologies produced by `msprime` were passed to `tsdate`. This is indicated in relevant figures as "`tsdate` (using simulated topology)". The second method of evaluation was to run `tsdate` on tree sequence topologies inferred from the simulated variation

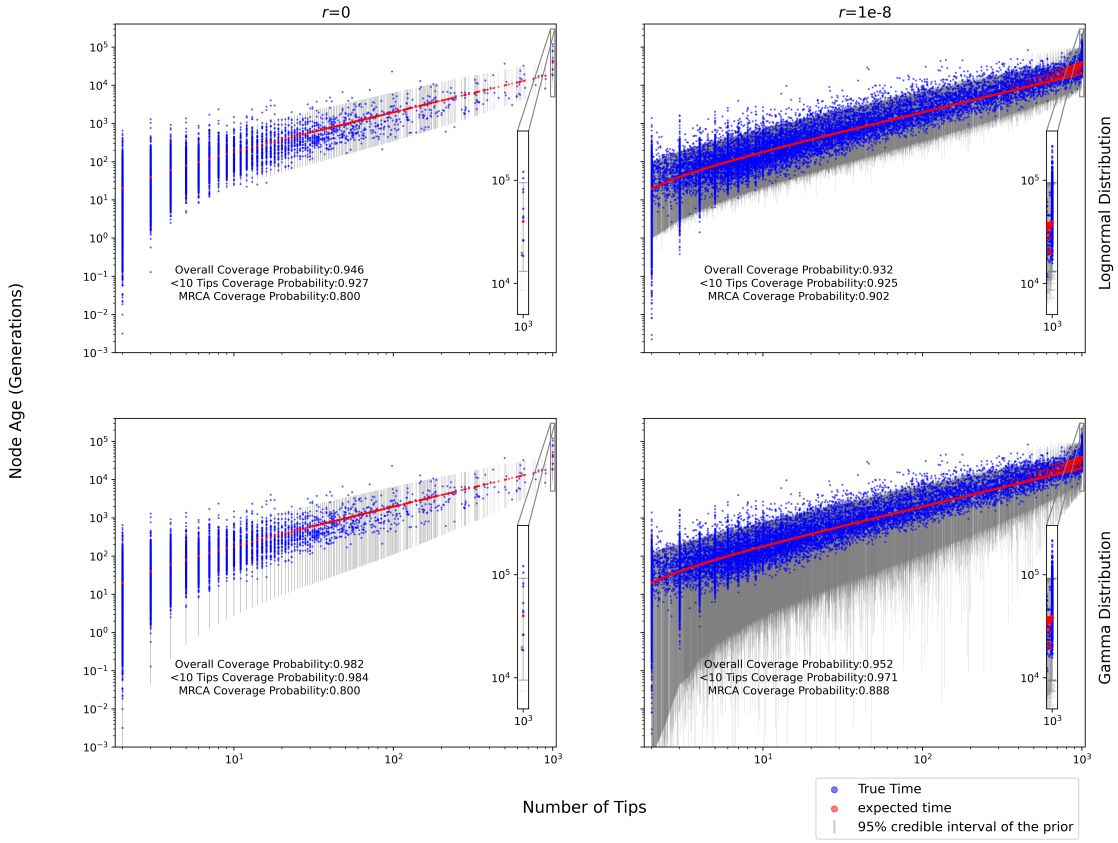


Figure 5.1: Accuracy of the `tsdate` node age prior distribution. The subplots in the left column show node ages from ten `msprime` simulations of length 500 kb with 1000 samples, N_e =10,000, and $r = 0$. The right column shows results of ten simulations using $r = 10^{-8}$ and the same parameters otherwise. The top plots show the accuracy of the lognormal approximation to the conditional coalescent and the bottom shows the gamma approximation.

data with `tsinfer`. This is referred to as "`tsdate` using `tsinfer` topologies". Both the outside and outside-maximisation passes (see Section 4.2.2) are assessed in Figure 5.2, as are various ratios of mutation to recombination rates. Since the outside pass empirically outperforms the outside-maximisation pass, only results using the former were used in subsequent simulations as well as in all results using real data. Variation data was converted to the required input formats for `GEVA` and `Relate` before running these programs using default settings. Mutation rates of 10^{-8} per base pair per generation were passed to all inference methods and recombination maps were passed to `tsinfer` and `Relate`. Ground truth allele ages were provided by `msprime`. For `tsdate` and `Relate`, allele age point estimates

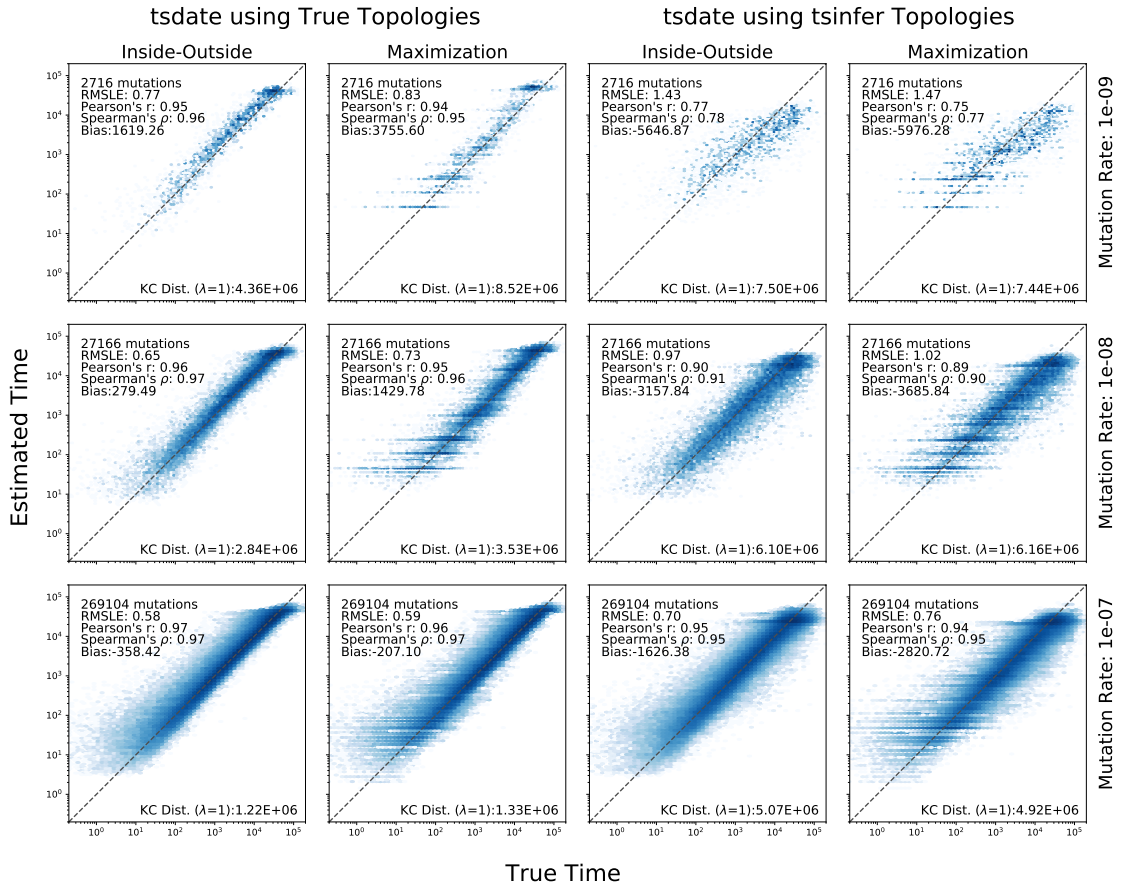


Figure 5.2: Accuracy of `tsdate` at various mutation rate settings. Each row shows the results of ten `msprime` simulations of 1 Mb with 500 samples, N_e= 10,000, and $r = 10^{-8}$. Three different values of μ were used, 10^{-9}, 10^{-8}, and 10^{-7}. In each subplot, simulated allele ages are on the x-axis and estimated allele ages are on the y-axis. The first two columns show the results of running `tsdate` on topologies simulated by `msprime`. The third and fourth columns show the results when running `tsdate` on topologies inferred by `tsinfer` from the simulated genotype data. The results of the inside pass followed by either an outside pass or outside-maximisation pass are shown.

were determined using the arithmetic mean of the ages of the upper and lower bounding ancestors. The `tsdate`-estimated age of these ancestors was given by the mean of each node's posterior distribution. The mean age of the joint clock with quality-filtered pairs was used for `GEVA`. Only sites that could be dated by all methods were evaluated. This means that singletons, $n-1$ tons, and other sites not dated by one or more methods were excluded from the evaluation.

Root mean squared log error (RMSLE), Pearson's r, Spearman's ρ, and estimator bias were used to assess the accuracy of allele age estimates. Tree inference accuracy

was measured by comparing simulated and inferred tree sequences using the KC tree distance metric (Kendall and Colijn, 2016), which includes a parameter, λ, determining the relative weight given to tree topology vs branch lengths when performing comparisons. I test both $\lambda = 0$, where all weight is given to the topology, and $\lambda = 1$, where all weight is given to branch lengths.

Figure 5.3 shows that our methods infer genealogies and allele ages with accuracy comparable or superior to the state-of-the-art in neutral coalescent simulations. When using the simulated topologies, `tsdate` has high accuracy in recovering allele age. When using inferred topologies from `tsinfer`, the accuracy of allele age estimation is comparable to `GEVA` and `Relate`.

To better compare the accuracy of each method, I plotted the inferred age of mutations from each method against one another in Figure 5.4. This plot shows that estimates from `tsdate`, using both simulated and inferred topologies, and `Relate` show the greatest similarity to one another.

Figure 5.5 shows the performance of `tsdate` on simulations of Chromosome 20 using the Out of Africa model (Gutenkunst et al., 2009; Adrion et al., 2020) and the previously described error models. Demography impacts allele age estimates since `tsdate` uses a constant value of N_e and does not currently model population structure. Using mismatch terms in `tsinfer` version 0.2 (see Section 1.5.2) improves topological inference in all simulations and allele age accuracy by most metrics. The iterative approach shows variable effects depending on the simulation model used and evaluation metric being considered. Inference using the iterative approach provides the lowest values of KC distance with $\lambda = 1$ (corresponding to the highest accuracy in branch length estimation); however, KC distance with $\lambda = 0$ and the various allele age accuracy measures show that the iterative approach generally only offers comparable or slightly improved accuracy compared to a single pass of `tsinfer` with mismatch terms. However, the simulations in Section 5.3.1 show that when inferring genealogies from data simulated with a uniform recombination map, iteration improves accuracy. Inference using error-prone variation data leads to lower accuracy by most measures.

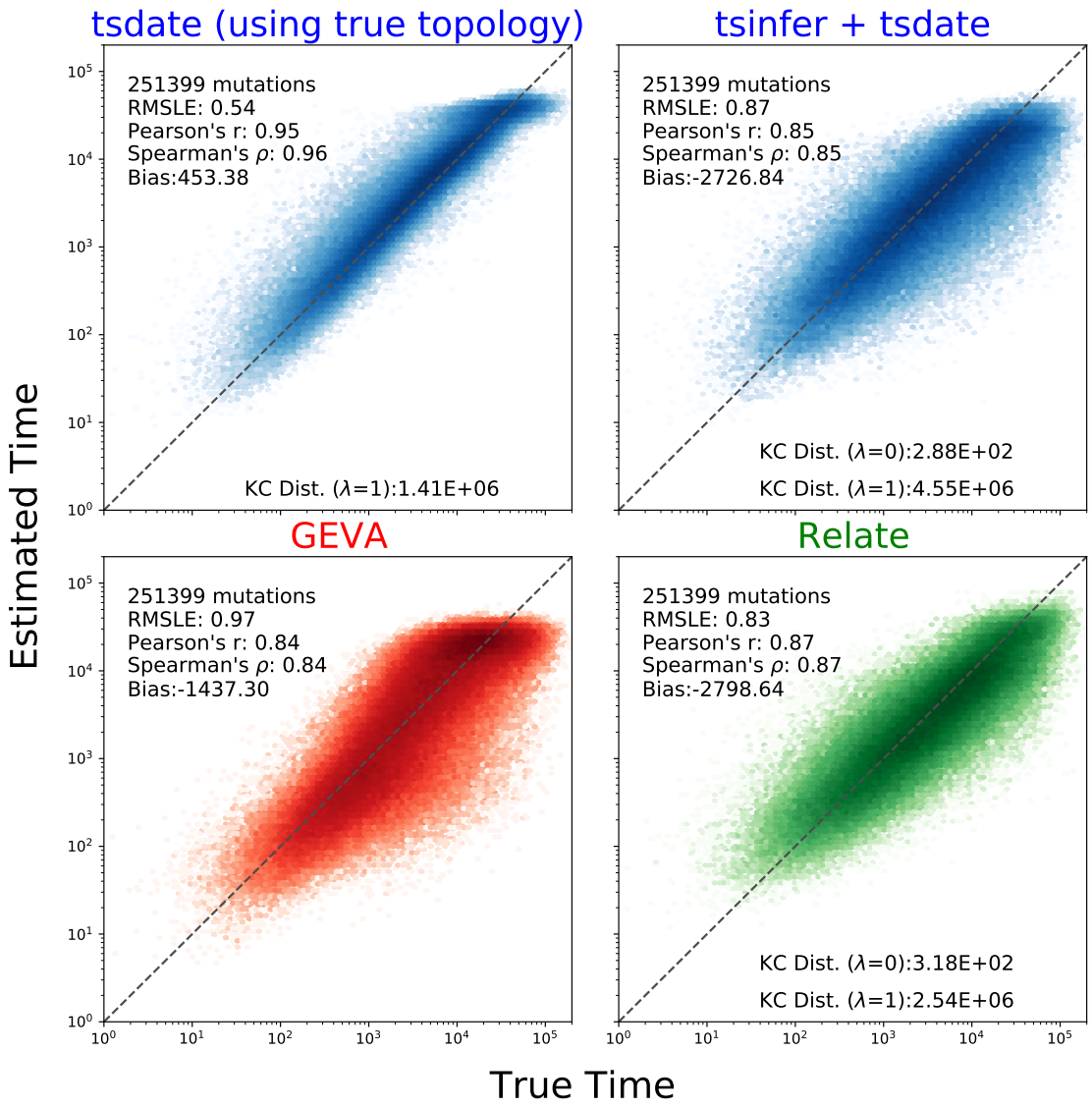


Figure 5.3: Comparison of the accuracy of `tsdate`, `GEVA`, and `Relate` on simulated tree sequence topologies. Thirty `msprime` simulations with 250 samples, 5 Mb of sequence, $N_e = 10{,}000$, and $\mu = r = 10^{-8}$ were used. The x-axis in all plots is the age of derived alleles from the simulation; estimated allele ages from the labeled method are on the y-axis. Top left subplot shows derived allele age estimates from running `tsdate` on simulated topologies. Top right subplot shows results from inferring a tree sequence topology with `tsinfer`, dating the tree sequence with `tsdate`, and using the resulting date estimates to reinfer and redate the tree sequence. Lower subplots show allele age estimates from `GEVA` and `Relate`. Only alleles dated by all methods are shown: this excludes singletons, $n-1$ tons, and alleles that were deemed to map poorly by `tsdate` or `Relate`.

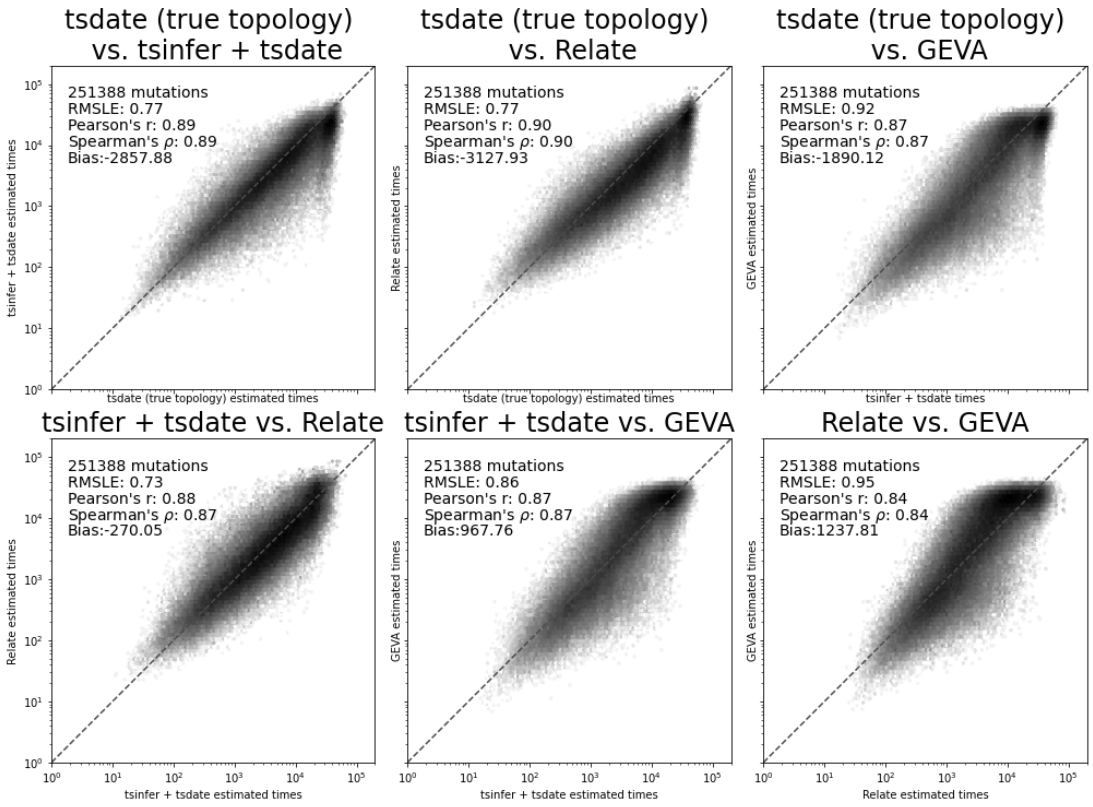


Figure 5.4: Comparing allele age estimates from `tsdate` using simulated topologies and `tsdate` using topologies estimated by `tsinfer`, as well as estimates from `Relate` and `GEVA`. The data shown in this plot is the same as that shown in Figure 5.3.

These same patterns are observed in Figure 5.6, which compares the accuracy of our methods with `GEVA` and a version of `Relate` that re-estimates effective population size and branch lengths.

5.2.3 Computational efficiency

I evaluated the computational scaling properties of `tsinfer` and `tsdate` compared to `Relate` and `GEVA` using `msprime` neutral coalescent simulations. The results are shown in Figure 5.7. The CPU run-time and memory requirements were recorded using two evaluation schemes. First, the length of the simulated chromosomes was held constant while increasing sample sizes are used. Second, sample size was fixed while sequence length increases. Default parameters were used for all methods. The maximum memory allowance for `Relate` was set to 32 Gb to allow the scaling analysis to proceed without error. The documentation of `Relate`, found

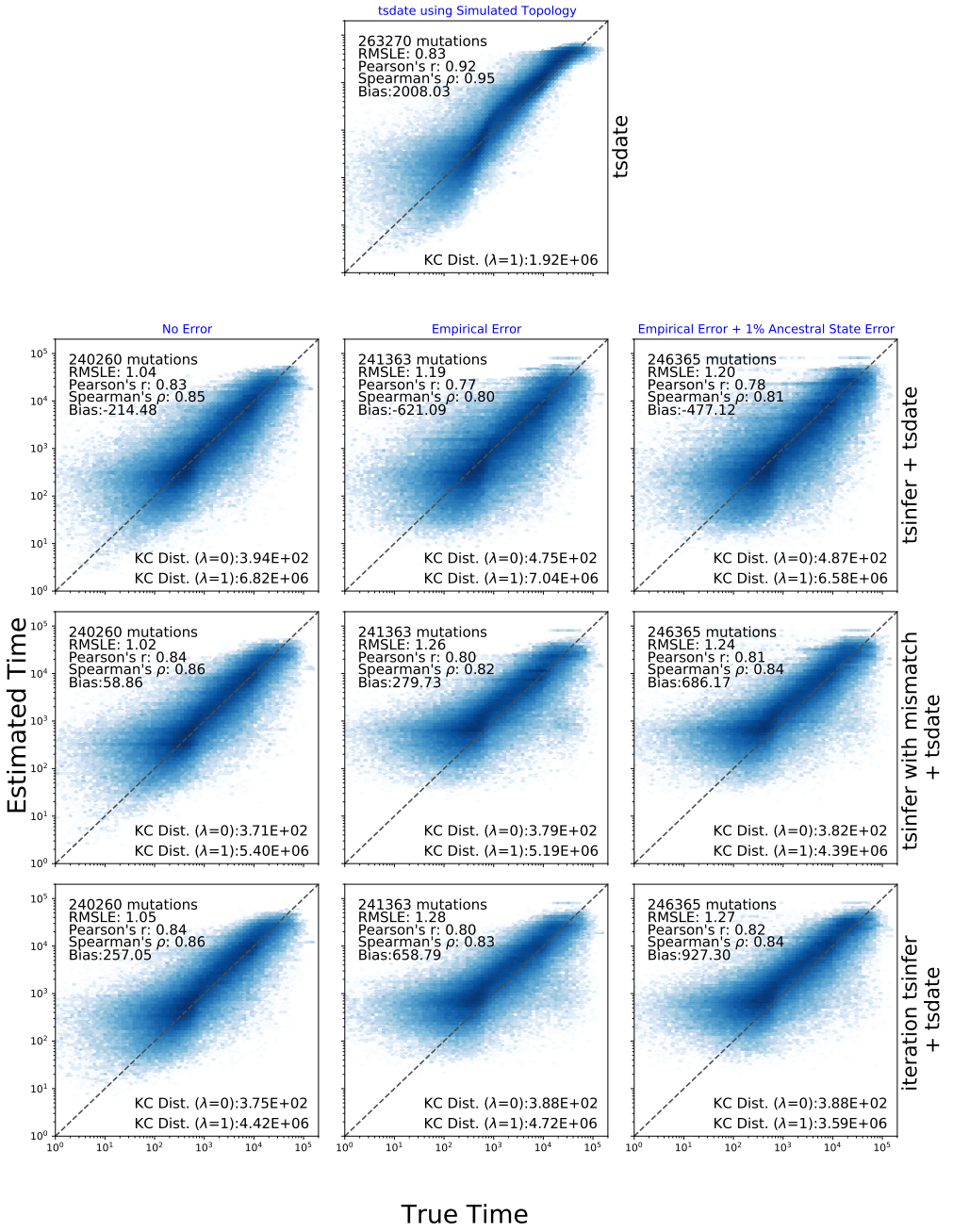


Figure 5.5: Evaluation of `tsdate` on simulated Chromosome 20 data. `stdpopsim` and `msprime` were used to simulate human Chromosome 20 using the three population Out of Africa model (Gutenkunst et al., 2009). 100 samples from YRI, CEU, and CHB were used. The top row shows the results of using `tsdate` on the simulated tree sequence topology. The second row shows the accuracy of `tsdate` on topologies from `tsinfer`. The third row shows the results of using topologies produced using `tsinfer`'s full Li and Stephens model with error. The final row shows the results of feeding the inferred dates from row two back into `tsinfer` and running `tsdate`. The first column in rows 2-4 shows simulations without genotype error, the second shows simulations with an empirical genotype error model and the third shows an empirical genotype error model and 1% ancestral state assignment error.

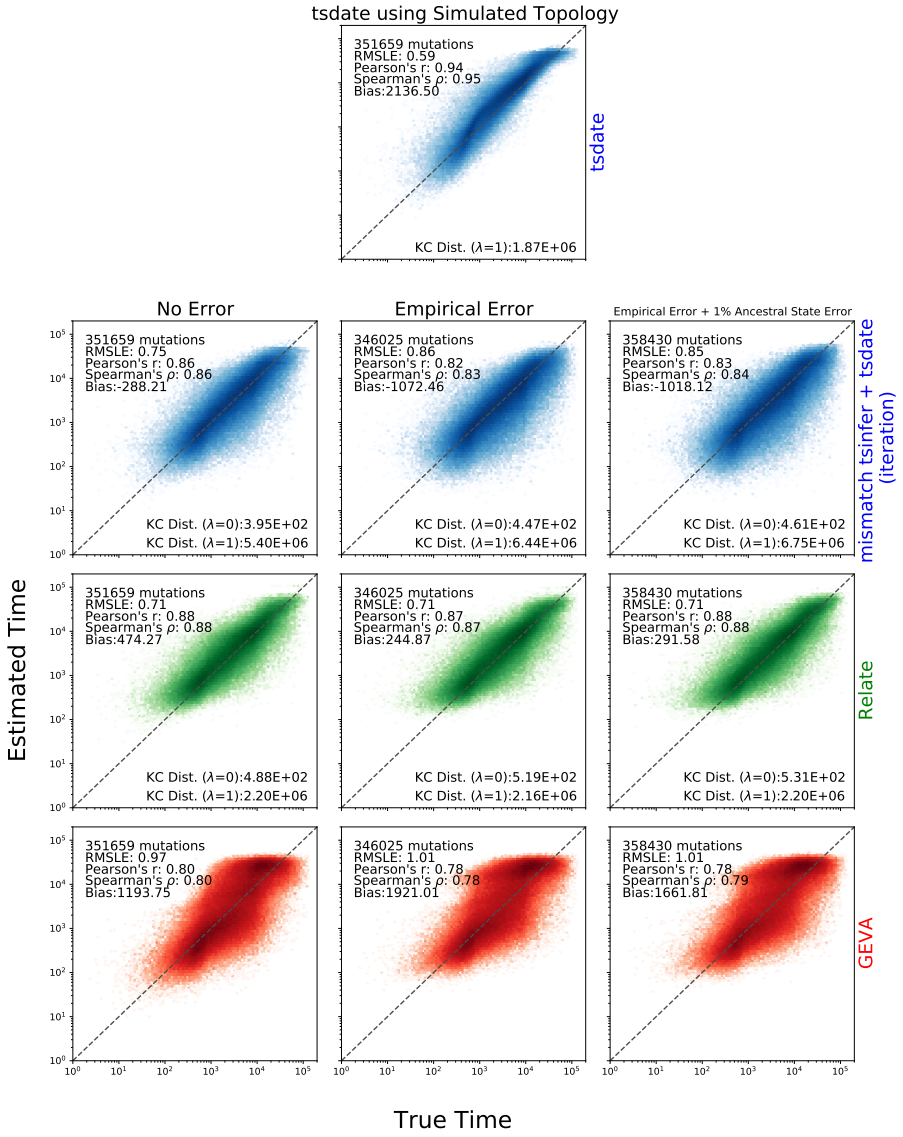


Figure 5.6: Accuracy of variant age inference. Evaluation of the accuracy of `tsdate`, `tsinfer`, `GEVA` and `Relate` on 5 Mb sections of Chromosome 20 simulated using the `OutOfAfrica_3G09` model implemented in `stdpopsim` (Gutenkunst et al., 2009; Adrion et al., 2020; Kelleher, Etheridge, and McVean, 2016) with the Chromosome 20 GRCh37 recombination map and 100 samples each from YRI, CEU, and CHB. 30 replicates were performed. The top row shows the accuracy of `tsdate` on the simulated topology. The second row shows the results of inferring tree sequences with `tsinfer`, dating the tree sequence with `tsdate`, and then re-inferring and re-dating the tree sequence (using the Chromosome 20 recombination map and a mismatch ratio of 1). The third row shows the results of `Relate` using a script provided by the authors to re-infer branch lengths and N_e. The fourth row shows the results of `GEVA` with default parameters. The first column shows simulations without genotype error, the second shows simulations with an empirical error model and the third shows an empirical error model and 1% ancestral state error. In each column, only sites dated by all three methods are shown. Summary metrics are shown in each subplot.

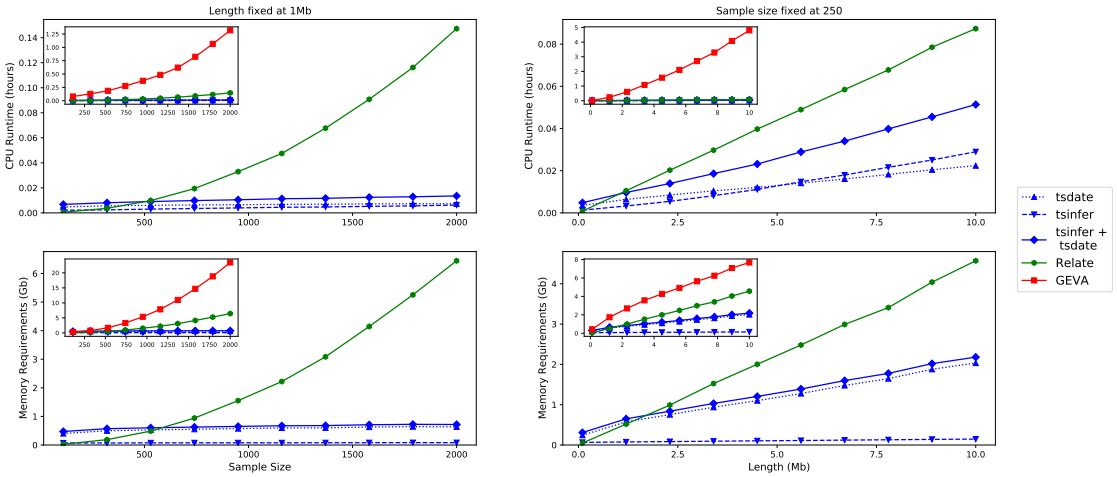


Figure 5.7: Scaling properties of `tsdate` compared to `tsinfer`, `Relate`, and `GEVA`. The left column shows the CPU and memory requirements for inference using the three methods on `msprime` simulations of 1 Mb, with $N_e = 10^4$, $\mu = r = 10^{-8}$ and sample sizes from 10 to 2000 (`Relate` continues to scale quadratically with larger sample sizes). Five replicates were performed at each sample size. The column on the right shows results of ten `msprime` simulations with sample size fixed at 250 and simulated lengths of 100 kb to 10 Mb. The main axes compare results for `tsdate`, `tsinfer`, and `Relate`. The inset plots show the same data with the addition of `GEVA`, which otherwise obscures the differences between other methods.

at `https://myersgroup.github.io/relate/getting_started.html`, notes that "Relate can become more efficient (in runtime and hard disk usage) with more memory, particularly for large sample sizes".

5.3 Evaluating the accuracy of incorporating ancient samples

5.3.1 Ancient samples and allele age estimation accuracy

Figure 5.8 shows a simulation-based evaluation of allele age inference accuracy when incorporating ancient samples into the previously described Out of Africa model (Gutenkunst et al., 2009; Adrion et al., 2020). I simulated 5 Mb of Chromosome 20 with a uniform recombination map and sampled 1008 modern samples split evenly between the three populations of the Out of Africa model: Yorubans (YRI), Han Chinese (CHB), and Utah Residents with Northern and Western European Ancestry (CEU). 40 ancient samples were also sampled. Three ancient sampling

schemes were evaluated: (1) pick samples at times randomly selected from the empirical distribution of the ages of published ancient samples, (2) pick samples from the population immediately prior to the time of the Out of Africa event in this model (5,650 generations ago), and (3) pick samples from 10,000 generations ago. 20 simulation replicates were performed with each sampling scheme.

The first sampling scheme used the estimated ages of ancient samples from the Reich Laboratory's Allen Ancient DNA resource, version 42.4, available at `https://reichdata.hms.harvard.edu/pub/datasets/amh_repo/curated_releases/index_v42.4.html`. The ancient sample times ranged from 90 years to 90,000, with a mean of 4,553 years and a standard deviation of 4,554. If the randomly sampled ages were younger than 21,200 years (the split time of CEU and CHB in the Out of Africa model using a generation time of 25 years), they were assigned to CEU or CHB with 50% probability. If older than the split time, they were assigned to the N_B population as described in Gutenkunst et al. (2009). All mutations only carried by ancient samples were removed from the simulation.

I then inferred dated tree sequences using three approaches. First, I inferred and dated a tree sequence using only the modern samples and evaluated the accuracy of the resulting allele age estimates. Second, I used these estimated allele ages to reinfer tree sequences (still only using modern samples) and evaluated the result. Third, I again used the allele age estimates to reinfer tree sequences with larger numbers of ancient samples. Allele age estimation accuracy was evaluated by comparing the true allele ages from the simulated tree sequence to the estimated allele ages using root mean squared log error (RMSLE) and Spearman's ρ. Allele ages were determined as described in the previous section. While the initial iteration step without ancient samples increases allele age estimation accuracy, adding samples drawn from the empirical distribution of ancient sample ages does not have a noticeable effect on overall accuracy. However, increasing numbers of samples from sampling schemes (2) and (3) consistently improves both MSLE and Spearman's rank correlation. This is likely because older samples will "correct" larger numbers of derived alleles that are initially estimated to be too young.

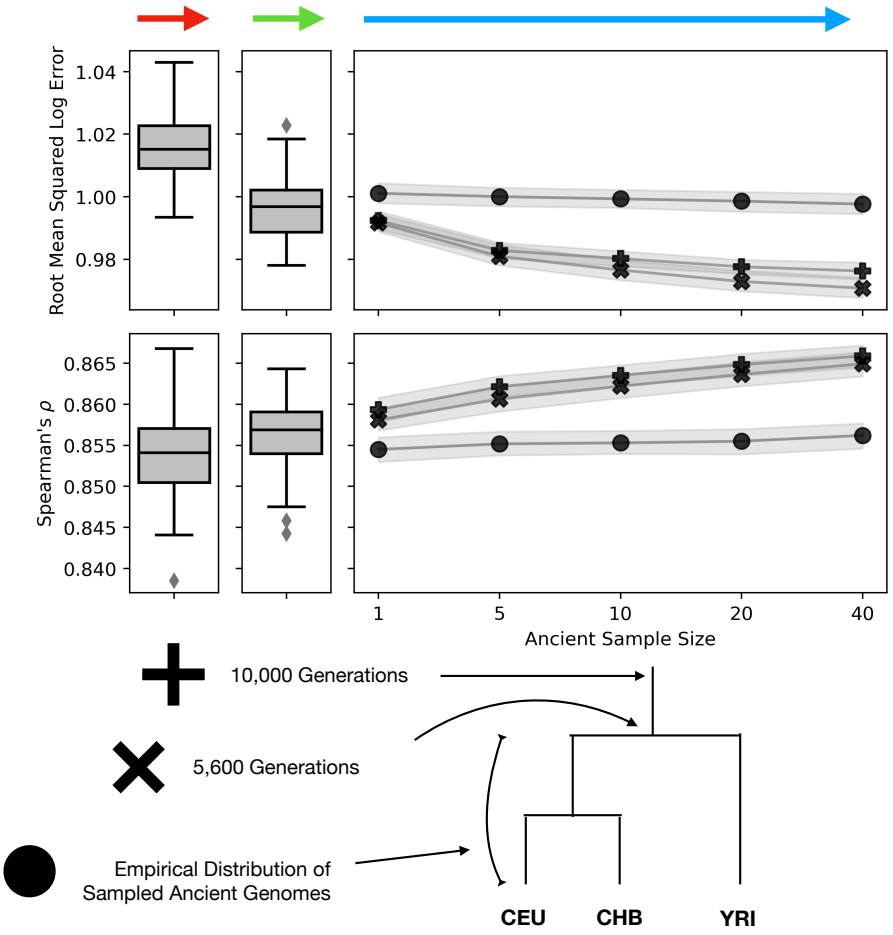


Figure 5.8: Accuracy of incorporating ancient samples into iterative inference methodology. Accuracy metrics, root-mean squared log error (top) and Spearman rank correlation coefficient (bottom), with modern samples only (first panel), after one round of iteration (second panel) and with increasing numbers of ancient samples. Coloured arrows correspond to Figure 4.7. Schematic below shows the three classes of ancient samples that were considered, reflecting where in history of humans they have been sampled from.

5.3.2 Archaic descent simulation

Tree sequences encode the genetic relationships between modern and ancient samples at every point in the genome. I used a published simulation model approximating the Out of Africa Event with archaic introgression into the ancestors of Eurasians and Papuans (Jacobs et al., 2019) to evaluate how well our inferred tree sequences reflect these relationships. Using an implementation of the simulation model from `stdpopsim` (Adrion et al., 2020), I simulated 10 replicates of 15 megabases (Mb) from Chromosome 20, sampling 200 chromosomes each from the African, West Eurasian,

East Asian, and Papuan populations. I slightly modified the ancient sampling scheme from the model by sampling three archaic individuals: the Denisovan (two chromosomes from the sampled Denisovan lineage, 2,203 generations ago), the Vindija (two chromosomes from the introgressing Neanderthal lineage, 1,725 generations ago), and the Altai Neanderthal (two chromosomes from the ancestral Neanderthal lineage, 3,793 generations ago). See Jacobs et al. (2019) Figure S5 for a schematic summarising the model. This model defines a generation to be 29 years (compared to 25 years in our other analyses).

I used migration records from the simulated tree sequences to determine patterns of archaic local ancestry. This provided a "ground truth" set of introgressed spans of sequence in modern samples. Migrations from Denisovan and Neanderthal lineages to the ancestors of modern, non-African samples in the simulated trees provide the locations of these introgressed spans of sequence in each modern sample. Importantly, in this simulation setting introgressed tracts exist regardless of whether or not they carry derived alleles and thus may be undetectable.

I examined how well patterns of common ancestry between *sampled* archaic and modern individuals reflect archaic introgression. The simulation from Jacobs et al. (2019) modelled introgression as "pulses" of admixture from archaics to moderns at discrete points in time. Consequently, the time to most recent common ancestor (TMRCA) between a modern and archaic sample is only observed to be more recent than $T_{ArchaicModern}$, the split time of modern and archaic lineages occurring 20,225 generations ago, at marginal trees where introgression occurred in the ancestors of that modern sample. However, the TMRCA between sampled archaics and moderns may not be more recent than $T_{ArchaicModern}$ at all introgressed tracts: the TMRCA is older than the split time at marginal trees where the sampled and introgressing archaics are distantly related. TMRCAs that are more recent than $T_{ArchaicModern}$, but older than T_{DenNea}, the split time of Neanderthals and Denisovans occurring 15,090 generations ago, can be attributed to introgression from *either* Neanderthals or Denisovans. Signals of introgression from these two

populations may be disentangled at marginal trees where the TMRCA of sampled archaics and moderns is more recent than T_{DenNea}.

I record tracts of common ancestry in this final category as well as the full set of introgressed tracts from the migration records using $n \times c$ matrices, where n is the sample size of modern individuals and c is the number of 1 kilobase (kb) chunks in the simulated chromosome. Two introgression matrices were constructed, one each for the Denisovans and Neanderthals. At each cell in an introgression matrix, a "1" indicates introgression in a modern individual in any marginal tree overlapping the 1 kb chunk. Three common ancestry matrices were constructed, one each for the Vindija, Altai, and Denisovan samples. In each common ancestry matrix, a "1" indicates that the TMRCA of the modern individual and *either* of the archaic individual's sampled chromosomes at any tree overlapping the chunk is less than T_{DenNea}.

Since TMRCAs of moderns and archaics more recent than T_{DenNea} reflect introgression from Denisovan or Neanderthal lineages, values of "1" in the common ancestry matrices are guaranteed to only exist where "1"'s are observed in the corresponding introgression matrix. Common ancestry between moderns and the Denisovan more recent than T_{DenNea} encompasses $\geq 99.9\%$ of introgressed Denisovan tracts (Std Dev 0.04%). This is likely attributable to the simulated population size of only $N_e = 100$ in the introgressing Denisovan D1 and D2 lineages from Jacobs et al. (2019). The Vindija Neanderthal shares common ancestry with humans more recently than T_{DenNea} for 61.3% (Std Dev 4.07%) of introgressed Neanderthal tracts. For the Altai, the value is 55.0% (Std Dev 1.99%).

Next, to investigate how well our inferred tree sequences reflect these observed patterns of introgression and common ancestry, I ran our iterative approach to infer a tree sequence of modern and archaic samples from the simulated data. I used an N_e value of 10^4 and a mutation rate of 10^{-8} per base pair per generation when running `tsdate`, the same parameters used with the real data. I then examined the inferred tree sequences, noting the spans of genome where the two "proxy ancestors"

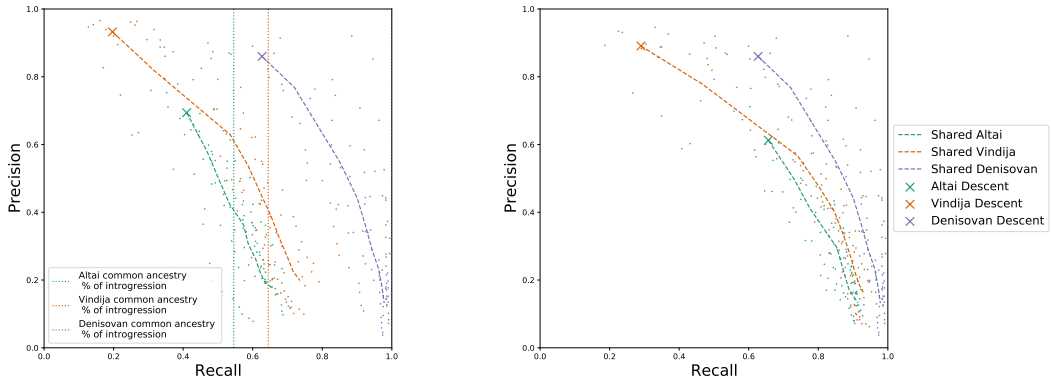
associated with each archaic individual have direct descendants among modern samples (see Section 4.3 for the definition of proxy ancestors).

Observed descent from proxy archaic ancestors should overlap with signals of introgression where sampled archaic haplotypes are a better approximation of ancestral material than `tsinfer`'s inferred ancestors at a given epoch. This approach enables an understanding of the relationships between sampled archaic and modern individuals without needing to identify a split time between moderns and archaics. Results are expected to be heavily influenced by such factors as the sampling times of ancient individuals, relationships between modern and ancient populations, error rates, and the quality of ancestors estimated by `tsinfer`. Relationships between sampled and introgressing ancient lineages may particularly affect the amount of introgressed material that can be recovered from the inferred tree sequences since, as noted previously, in this simulation model more recent common ancestry is observed between the sampled and introgressing Denisovans than between the sampled and introgressing Neanderthals.

I used the inferred tree sequences to construct three additional $n \times c$ matrices that record direct descent from proxy ancestors associated with the three sampled archaic individuals. In each cell in a matrix, a "1" records where a modern individual descends from the relevant proxy archaic sample in a marginal tree overlapping the 1 kb chunk. Comparing the binary matrices recording ground truth tracts of introgression or common ancestry from the simulations with the matrices recording inferred descent from archaic samples yielded true positives (TP), corresponding to introgression or common ancestry in the simulation and descent in inferred tree sequence, false positives (FP), where no introgression or common ancestry is noted in the simulation but descent is inferred, and false negatives (FN), where introgression or common ancestry is noted in the simulation but no descent is inferred. Table 5.1 shows the resulting rates of precision and recall, which are defined as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{5.1}$$

$$Recall = \frac{TP}{TP + FN}. \tag{5.2}$$



(a) Introgressed Tracts **(b)** Shared Ancestry Tracts

Figure 5.9: Precision-recall curves for recovering relationships between archaic and modern samples using inferred tree sequences. (a) Accuracy of recovering introgressed tracts from archaic populations. Dotted vertical lines indicate total amount of modern genetic material where modern samples and each archaic individual share common ancestry in the *simulated* tree sequence more recently than T_{DenNea} as a proportion of total introgressed material. Common ancestry in the *inferred* tree sequences occurs when the TMRCA of modern samples and an archaic sample occurs more recently than a specified time cutoff. Cutoffs ranging from the time of each archaic sample to T_{DenNea} were tested, scatter plots show the results from each archaic sample at each cutoff and each simulation replicate. Dashed lines show the average precision and recall across replicates at each cutoff. Older time cutoffs result in progressively lower precision and higher recall. Precision is maximised at points marked with an "X", where the cutoff is equal to the time of the sample, i.e. where only material directly descending from archaic proxy samples is used. (b) Precision-recall curves for recovering only the shared ancestry tracts for the three sampled archaics indicated by the dotted vertical lines in (a).

The proportion of all modern genetic material that is introgressed from each archaic population, the proportion of modern genetic material sharing common ancestry with archaic samples more recently than T_{DenNea}, and the proportion of modern genetic material that is inferred to descend from archaic samples are also given in Table 5.1.

The simulation results indicate that descent from the Vindija and Denisovan, and to a lesser extent the Altai, recovers ground truth introgression and shared ancestry with high precision. In the case of the Neanderthals, where shared ancestry tracts incompletely represent introgressed tracts, inferred descent recovers more shared ancestry tracts than introgressed tracts.

It is possible to increase recall by examining patterns of common ancestry between sampled archaics and moderns occurring more recently than a given split time in the inferred tree sequence, as was performed in the simulated tree sequence.

	Vindija		Altai		Denisovan	
	Introgressed	Common Ancestry	Introgressed	Common Ancestry	Introgressed	Common Ancestry
Precision	0.92 (0.03)	0.87 (0.05)	0.69 (0.08)	0.59 (0.08)	0.86 (0.04)	0.86 (0.04)
Recall	0.18 (0.05)	0.28 (0.07)	0.42 (0.04)	0.65 (0.06)	0.61 (0.11)	0.61 (0.11)
Simulated % of moderns	2.86% (1.12%)	1.77% (0.71%)	2.86% (1.12%)	1.57% (0.60%)	0.99% (0.24%)	0.98% (0.24%)
Inferred % of moderns	0.53% (0.15%)		1.65% (0.31%)		0.69% (0.12%)	

Table 5.1: Results of archaic descent simulations evaluating how well direct descent from archaic haplotypes in inferred tree sequences recovers ground truth introgressed tracts of sequence as well as tracts where modern samples and sampled archaics share common ancestry more recently than T_{DenNea}. Definitions of precision and recall are given in Equations (5.1) and (5.2). "Simulated % of moderns" refers to the percentage of genetic material from all modern individuals in each simulation that is introgressed from archaics or shares recent common ancestry with archaics. "Descendant % in moderns" is the percentage of modern genetic material in each inferred tree sequence which descends from each archaic individual. In each cell, the mean values from 10 simulation replicates is given with the standard deviation in brackets.

The precision and recall for recovering ground truth introgressed and shared ancestry tracts at different split times is analysed in detail in Figure 5.9, where observed precision and recall values were obtained by testing split times in intervals between the time of each archaic sample and T_{DenNea}. These results show that progressively older split times recover nearly all shared ancestry tracts, though at lower levels of precision. While it is possible to use these results to choose time cutoffs for each sample that maximise recall at a given level of precision, the optimal values derived from simulations will doubtlessly differ from the real data due to the imperfect modelling of demographic history as well as varying levels of error. In the results presented in the next chapter, I chose to report tracts that directly descend from these samples in order to avoid choosing a cutoff time while likely maximising precision.

5.4 Empirical estimates of 1000 Genomes Project variant ages on chromosome 20

To provide empirical validation of our methods, I considered the ability of our methods and alternatives to infer allele ages that are consistent with observations from ancient samples. I inferred and dated a tree sequence of TGP Chromosome 20 (without using the ancient samples) and compared the resulting point estimates and upper and lower bounds on allele age with results from `GEVA` (Albers and McVean, 2020) and `Relate` (Speidel et al., 2019).

5.4.1 Materials and Methods

The TGP Phase 3 GRCh37 VCF was accessed from TGP release 20130502 (1000 Genomes Project Consortium, 2015) A `.samples` file, a file format used with `tsinfer`, was created from this VCF file using all biallelic SNPs with high-confidence ancestral alleles from Ensembl release 100 (Hunt et al., 2018). I used `tsinfer` version 0.2 (see Section 1.5.2) to infer a tree sequence of TGP individuals with the GRCh37 recombination map, mismatch ratios of 1, and a precision setting of 15. The tree sequence was dated using `tsdate` with N_e set to 10,000 and a mutation rate of 10^{-8} mutations per base pair per generation. Default settings were used for all other parameters. Allele ages were estimated by using the arithmetic mean of the nodes above and below the oldest mutation. Allele age estimates from `GEVA` were gathered from `https://human.genome.dating/download/index`. I used the mean allele age estimate from the joint clock as a point estimate of allele age and the upper limit of the 95% confidence interval from the joint clock as an upper bound estimate (Albers and McVean, 2020). `Relate` TGP allele age estimates were downloaded from `https://zenodo.org/record/3234689#.X8EXyapKhTZ`. The publicly available `Relate` age estimates were calculated separately in each TGP population, so all available age estimates and upper bounds were included in this analysis. The point estimate of allele age was obtained by taking the mean of the upper and lower age bounds.

Empirical lower bounds were provided by radiocarbon dates of ancient samples. These dates were derived from the Reich dataset for all ancient individuals with the exception of the Afanasievo family (which we assigned to 4.6 kya), Chagyrskyaya Neanderthal (Chagyrskaya 8) (80 kya) (Mafessoni et al., 2020), Vindija Neanderthal (Vindija 33.19) (50 kya) (Prüfer et al., 2017), Altai Neanderthal (Denisova 5) (110 kya), and Denisovan (Denisova 3) (63,900 kya) (Douka et al., 2019). If multiple ancient samples carried a derived allele, the oldest sample carrying the allele was used as the lower bound on the age of that allele. This resulted in a combined dataset of 3,734 ancient individuals.

5.4.2 Results

A set of 659,804 sites for which age estimates can be found from all three methods was assembled. This excluded singletons and $n-1$ tons, which `GEVA` did not estimate, sites where the derived and alternate alleles were inconsistent (as `GEVA` estimates the age of alternate alleles), indels, and sites with low quality ancestral states. Of these, 76,889 derived alleles were observed within the combined set of 3,734 ancient samples, thus putting a lower bound on allele age. The relationship of allele age estimates from each method to allele frequency is shown in Figure 5.11, the distribution of ages is shown in Figure 5.12, and the comparisons of allele ages from the differing methods to one another is shown in Figure 5.13. Figure 5.10 shows that estimated allele ages from `tsdate` and `Relate` have the greatest compatibility with ancient lower bounds, despite the fact that the mean age estimate from `tsdate` is more recent than that of `Relate`.

Indeed, the distribution of age estimates varies between the three methods (see Figure 5.12), with `Relate` showing a higher mean estimated age compared to the other two methods. The `tsdate` mean estimated allele age at the 659,804 comparable sites was 5,919 generations and the mean upper bound on allele age was 9,012 generations. For `Relate`, the equivalent values were 6,816 and 11,732 generations. Since `Relate` provides an estimate for each population, these values

were first found by averaging the available estimates for each site, then averaging all the sites. For `GEVA` the equivalent values were 5,192 and 6,147 generations.

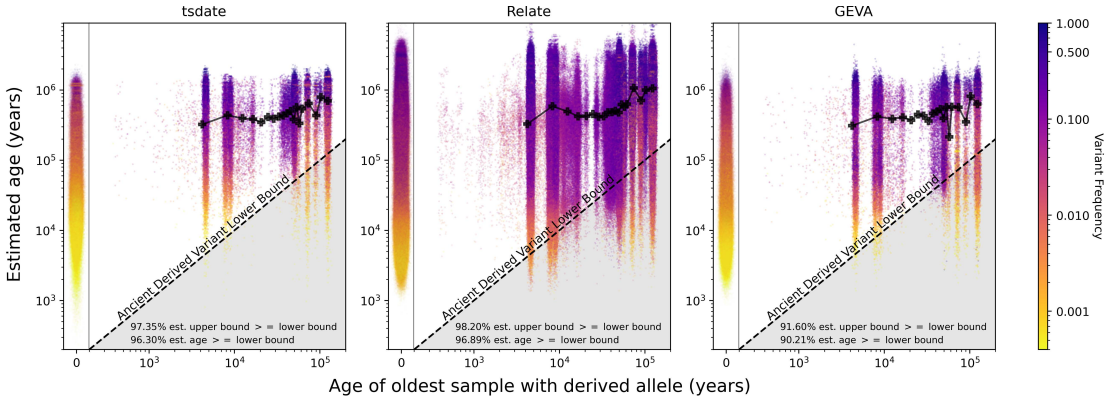


Figure 5.10: Comparison of mutation age estimates from three methods (`tsdate`, `Relate` and `GEVA`) using 3,734 ancient samples at 76,889 variants on Chromosome 20. The radiocarbon-dated age of the oldest ancient sample carrying a derived allele at each variant site in the 1000 Genomes Project is used as the lower bound on the age of the mutation (diagonal lines). Mutations below this line have an estimated age that is inconsistent with the age of the ancient sample. Black lines on each plot show the moving average of allele age estimates from each method as a function of oldest ancient sample age. Plots to the left show the distribution of allele age estimates for modern-only variants from each respective method. Additional metrics are reported in each plot. While 76,889 comparisons are shown in the `tsdate` and `GEVA` panels, `Relate` provides allele age estimates separately in each TGP population where the allele is present, resulting in 407,995 comparisons being shown in the middle panel.

5.5 Discussion

The evaluations of `tsinfer` and `tsdate` presented in this chapter show that these methods infer dated genealogies with unmatched scaling properties and accuracy which is similar, and in many cases superior, to the state of the art. Furthermore, the iterative approach improves inference accuracy in many settings considered, and the resulting tree sequences allow straightforward estimation of the introgressed haplotypes from archaic populations.

These analyses suggest multiple future directions for improving inference accuracy. Figures 5.3, 5.5, and 5.6 showed that `tsdate`'s accuracy was excellent when used on the true simulated trees, and decreased when using trees inferred by `tsinfer`. We therefore can expect dating performance to improve as topological

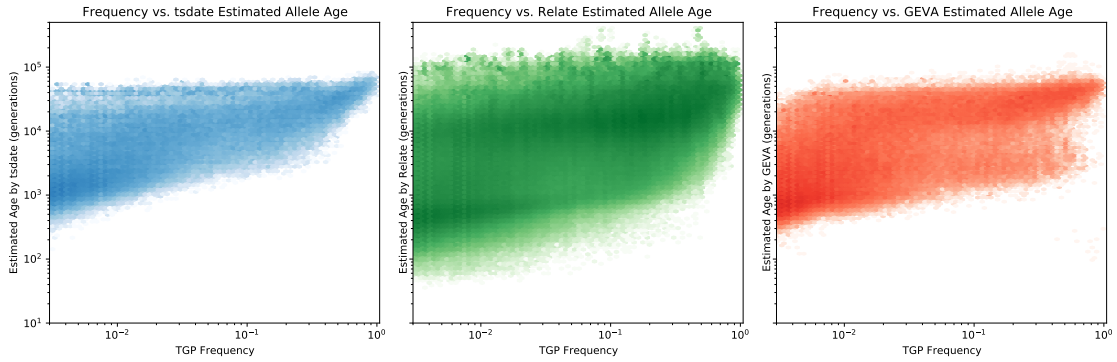


Figure 5.11: Estimated age of TGP derived alleles from `tsdate`, `GEVA` and `Relate` compared to allele frequency. Estimates of allele age were found using the arithmetic mean of the node ages above and below the mutation for `tsdate` and `Relate`. For `GEVA`, the mean age of the joint clock estimate is used. The same set of alleles was also used in Figs. 5.10a, 5.12, and 5.13. Note that `Relate` provides separate age estimates from each TGP population. As a result, the `Relate` panel contains 12,174,423 estimates, while the panels for `GEVA` and `Relate` contain 659,804 estimates.

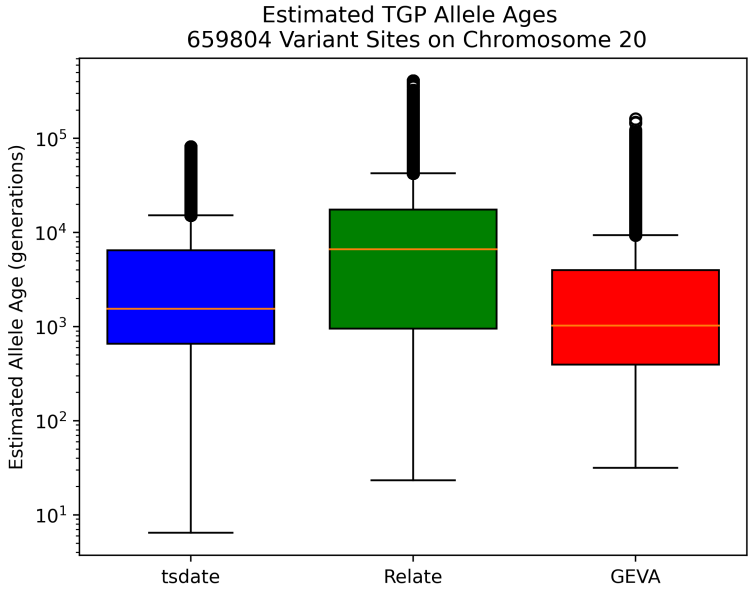


Figure 5.12: Distribution of TGP derived allele age estimates from `tsdate`, `GEVA` and `Relate`, showing that the median estimated allele age from `tsdate` is intermediate to that from `GEVA` and `Relate`. The age range of `tsdate` estimates depends on the number of time slices specified by the user (default settings were used for these results). See caption of Figure 5.11 for details of the data being plotted.

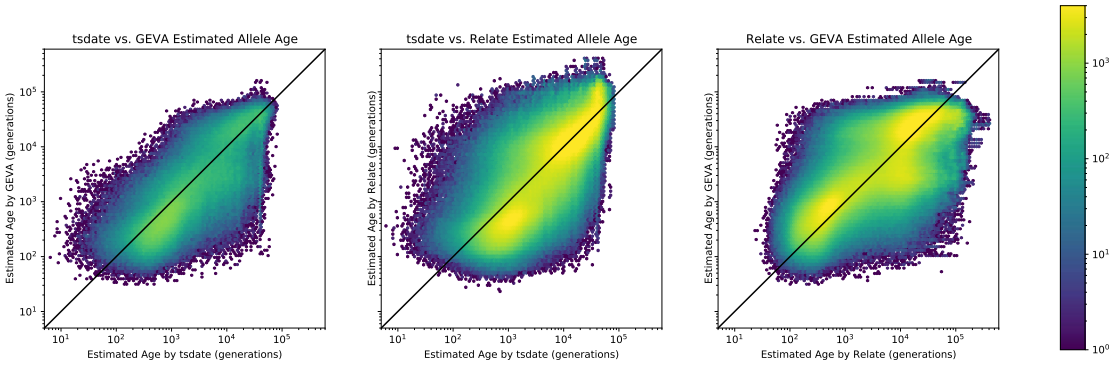


Figure 5.13: Comparison of TGP derived allele age estimates from `tsdate`, `GEVA` and `Relate`. Each hexbin plot shows estimates of variant age by two different methods. See caption of Figure 5.11 for details of the data being plotted.

inference improves. Evaluation of the iterative approach shows that improving the relative order of derived alleles can improve `tsinfer`'s accuracy, as illustrated by Figure 5.8. Multiple avenues also exist to improve the ancestral haplotype generation algorithm, such as replacing heuristic rules with a probabilistic algorithm (see Section 7.2.2). Another approach is to use matched ancestral haplotypes from an initial round of iteration in subsequent iteration steps.

Genotype and ancestral state error decrease the accuracy of inferred topologies and allele age estimates from these methods. Another iterative approach could use `tsinfer` version 0.2 to detect genotype and phasing errors in input data, before reinferring tree sequences with the corrected data. In the next chapter, I show that the number of mutations at a site is suggestive of genotype error, but more work is necessary to precisely call erroneous variants, as well as recurrent mutations (see Section 7.2.1).

The empirical evaluation of allele age estimates provided here offers a novel check on the estimated age of derived alleles, suggesting that the vast majority of estimates are consistent with ancient radiocarbon dates. Discrepancies could be caused by recurrent mutation or by inaccuracies in the variant age estimation methods, ancient variation data, or radiocarbon dating. Understanding the extent of these various phenomena could inform future use of ancient samples in this setting.

Having validated the accuracy of `tsdate` and `tsinfer` under a variety of conditions, in the next chapter I apply these methods to infer a single, unified tree sequence

describing human genetic variation from multiple modern and ancient datasets.

6

A unified genealogy of modern and ancient genomes

Contents

6.1 Introduction 122
6.2 Integrating and analysing diverse datasets with tree sequences 123
6.3 A unified genealogy of modern and ancient human genomes 124
6.3.1 Materials and methods 125
6.3.2 Results 128
6.4 Fine-scale worldwide population structure 130
6.4.1 Methods 130
6.4.2 Results 131
6.5 Modern and ancient population relationships 132
6.5.1 Methods 132
6.5.2 Results 134
6.6 Non-parametric inference of spatio-temporal dynamics in human history 137
6.6.1 Methods 138
6.6.2 Results 138
6.7 Discussion 140

6.1 Introduction

In the previous chapters I introduced, discussed, and validated methods for inferring dated genealogies from modern and ancient genomes. In this chapter, I present results from Wohns et al. (2021), where my co-authors and I use these methods to build the most comprehensive dated genealogy of sequenced human genomes constructed to date. This unified genealogy integrates the 1000 Genomes Project (TGP) (1000 Genomes Project Consortium, 2015), the Human Genome Diversity Project (HGDP) (Bergström et al., 2020), and the Simons Genome Diversity Project (SGDP) (Mallick et al., 2016), as well as sequenced archaic genomes (Meyer et al., 2012; Prüfer et al., 2014; Prüfer et al., 2017; Mafessoni et al., 2020) and a novel high-coverage family quartet from the Eurasian steppe 4,600 years ago. We also use thousands of low-coverage ancient samples to constrain ancestral relationships.

The genealogy supports the direct inference of fine-scale population genetic relationships, including the full distribution of times to common ancestor for 215 populations from the constituent datasets. We detect complex associations between Denisovans and Oceanians, identify variant sites which are likely erroneous, and investigate deviations from commonly used population genetic assumptions such as the infinite sites model. Finally, we provide a simple spatiotemporal inference method which we use to estimate the locations of 26 million ancestral haplotypes across 5 million inferred trees. This method places the centre of gravity of human ancestors in Northeast Africa by 72 thousand years ago, demonstrating that the deepest ancestral relationships are centred in Africa in autosomal data, just as they are for mitochondrial DNA and Y chromosomes.

All results and text in this chapter are adapted from Wohns et al. (2021). All text and figures were produced by myself in collaboration with my coauthors, save for Figure 6.1.

6.2 Integrating and analysing diverse datasets with tree sequences

Our ability to determine relationships among individuals, populations and species is being transformed by population-scale biobanks of medical samples (Bycroft et al., 2018; Taliun et al., 2019), collections of thousands of ancient genomes (Reich, 2018), and efforts to sequence millions of eukaryotic species (Lewin et al., 2018). Such relationships, and the resulting distributions of genetic and phenotypic variation, reflect the complex set of selective, demographic and molecular processes and events that have shaped species and are consequently a rich source of information about them (Lazaridis et al., 2014; 1000 Genomes Project Consortium, 2015; Mallick et al., 2016; Bergström et al., 2020).

However, our ability to learn about evolutionary events and processes from the totality of genomic variation, in humans or other species, is limited by multiple factors. First, combining information from multiple data sets, even within a species, is technically challenging; discrepancies between cohorts due to error (Belsare et al., 2019), differing sequencing techniques (Shi et al., 2016; Wenger et al., 2019) and variant processing (Hwang et al., 2015) lead to noise that can easily obscure genuine signal. Second, few tools can cope with the vast data sets that arise from the combination of multiple sources (Kelleher et al., 2019). Third, statistical analysis typically relies on data reduction techniques (Cavalli-Sforza and Feldman, 2003; Patterson, Price, and Reich, 2006) or the fitting of parametric models (Pritchard, Stephens, and Donnelly, 2000; Lawson et al., 2012; Patterson et al., 2012; Pickrell and Pritchard, 2012), which will inevitably provide an incomplete picture of the complexities of evolutionary history. Finally, data access and governance restrictions often limit the ability to combine data sources (Bonomi, Huang, and Ohno-Machado, 2020).

Tree sequences represent a potential solution to many of these problems (Kelleher, Etheridge, and McVean, 2016; Kelleher et al., 2019). First, building a tree sequence provides a principled approach to integrating datasets. This is due to the fact that a tree sequence inferred from data containing errors requires unparsimonious

"corrections" such as additional edges or mutations. These corrections are readily apparent in the resulting structure. Second, tree sequences are an ideal data structure when working with very large datasets. The tree sequence can not only be used to compress genetic data (Kelleher et al., 2019), but also leads to highly efficient algorithms for calculating population genetic statistics (Ralph, Thornton, and Kelleher, 2020). Finally, as argued in Chapter 1, the tree sequence, and the mapping of mutation events to it, can in principle represent the totality of what is knowable about genealogical relationships and the evolutionary history of individual variants. An inferred tree sequence may therefore be directly queried to gain non-parametric insight into the evolutionary history of sampled genomes.

In the following, we use the methods discussed in this thesis to infer a single, unified and time-resolved tree sequence from heterogeneous ancient and contemporary human genetic datasets. We validate this structure using the known age and population affinities of ancient samples and demonstrate its power through spatio-temporal inference of human ancestry. We note that while humans are the focus of this study, the methods and approaches we introduce are valid for most recombining organisms.

6.3 A unified genealogy of modern and ancient human genomes

To generate a unified genealogy of modern and ancient human genomes, we integrated data from eight sources. This included three modern datasets: the 1000 Genomes Project (TGP) which contains 2,504 sequenced individuals from 26 populations (1000 Genomes Project Consortium, 2015), the Human Genome Diversity Project (HGDP), which consists of 929 sequenced individuals from 54 populations (Bergström et al., 2020), and the Simons Genome Diversity Project (SGDP) with 278 sequenced individuals from 142 populations (Mallick et al., 2016). 154 individuals appear in more than one of these datasets (see next section). In addition, we included data from three high-coverage sequenced Neanderthal genomes (Prüfer et al., 2014; Prüfer et al., 2017; Mafessoni et al., 2020), the single Denisovan genome (Meyer

et al., 2012), and newly reported high coverage whole genome data from a nuclear family of four (a mother, a father, and their two sons with average coverage of 10.8x, 25.8x, 21.2x, and 25.3x) from the Afanasievo Culture, who lived $\sim$ 4.6 thousand years ago (kya). Finally, we used 3,589 published ancient samples from over 100 publications compiled by the Reich Laboratory to constrain allele age estimates. These ancient genomes were not included in the final tree sequence due to the lack of reliable phasing for the majority of samples, though I later discuss a potential solution to this problem in Section 7.2.3.

6.3.1 Materials and methods

The TGP GRCh38 variant calls were accessed from the European Variation Archive under accession number ERZ822766 (Lowy-Gallego et al., 2019). SGDP data was downloaded from `https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/phased_data/PS2_multisample_public` (Mallick et al., 2016). The HGDP statistically phased dataset was downloaded from `ftp://ngs.sanger.ac.uk/production/hgdp/hgdp_wgs.20190516/statphase/` (Bergström et al., 2020). The Reich laboratory's Allen Ancient DNA resource, version 42.4, provides a compilation of 3589 genotyped ancient individuals available at `https://reichdata.hms.harvard.edu/pub/datasets/amh_repo/curated_releases/index_v42.4.html` (the full list of citations for these papers is available from this link). The sequenced Denisovan, Vindija, Altai, Ust-Ishim, Loshbour, and LBK-Stuttgart samples were downloaded from `http://cdna.eva.mpg.de/neandertal/Vindija/VCF/` and the Chagyrskaya Neanderthal from `http://ftp.eva.mpg.de/neandertal/Chagyrskaya/`.

The GRCh38 reference assembly was used in all analyses in this chapter. Only a GRCh37 version of the SGDP and ancient datasets were available, so these datasets were lifted over to GRCh38 using the Picard tool (*Picard toolkit* 2019) and the hg19 to hg38 UCSC LiftOver chain (Hinrichs et al., 2006).

The Afanasievo "quartet" allows for reliable phasing of ancient genomes. The four archaic individuals are unphased, but their high levels of homozygosity, together with the distinctiveness of their haplotypes compared to modern individuals, should

mitigate the errors caused by unphased archaic genomes. We used bcftools (v1.10.2) to calculate genotype likelihood of biallelic SNPs in the Afanasievo samples. We used Beagle 4.0 by providing genotype likelihood file (gl argument) and pedigree file (ped argument) to phase this family without any reference panel and imputation. We excluded sites with max(GP) <0.99.

Once VCF files from each dataset were prepared, we created `.samples` files, a file format used with `tsinfer`, for each constituent dataset using all biallelic SNPs with high-confidence ancestral alleles from Ensembl release 100 (Hunt et al., 2018). The `.samples` files were then combined using the `merge` functionality in `tsinfer`, where the genotypes of variants missing from at least one sample were marked as "missing data" to be imputed by `tsinfer`. A tree sequence was then inferred and dated for each autosome using modern samples from the HGDP, SGDP, and TGP datasets. Empirical lower bounds on allele age were gathered from all ancient samples (including unphased and genotyped samples). Variants present in an ancient individual but missing in all modern datasets were not used. The rationale behind this approach is that most variants observed in ancient samples will also be seen in modern humans. Consequently, variants only observed in ancient samples are more likely to be erroneous. Martiniano et al. (2020) find that this simple check is appropriate for modern humans from the last 50,000 years. For instance, the authors find that 96.99% of variants in the Ust'Ishim chromosome 1 accessible regions are present in the TGP. However, it is important to note that excluding such variants could cause biases in the resulting tree sequence, for instance by underestimating the divergence between ancient samples and the ancestors of sampled moderns.

Ancient lower bounds at each variant site were compared with estimates from the inferred tree sequence of modern samples. The estimated age of a site in the inferred tree sequence is defined as the oldest most recent common ancestor which possesses a derived variant at each site. At sites where the empirical lower bound from ancient samples is older than our estimated age from `tsdate`, this bound was used as the variant age in the tree sequence inferred by `tsinfer`. With these constrained age estimates, the tree sequence was re-inferred using the four archaic individuals and

the Afanasievo quartet. Details of the incorporation of ancient samples can be found at `https://github.com/awohns/unified_genealogy_paper/all-data`.

This approach to incorporating ancient samples can introduce two types of bias that inflate mutation age estimates. First, the age estimate will be inaccurate if ancient samples carry a genuine recurrent mutation (which are expected to occur more frequently at highly mutagenic sites) which is older than the mutation carried by moderns. Second, age estimates may also be inflated if ancient data contains errors such as deamination. More work is required to understand the prevalance of these potential sources of bias, as well as to assess whether `tsinfer` can correct such events by placing recurrent and back mutations.

154 modern individuals appear in more than one dataset. This consists of 130 individuals which appear in both SGDP and HGDP, as well as 24 individuals which appear in both SGDP and the TGP GRCh38 release. Duplicated individuals are retained in all analyses for the following reasons: the scalability of our methods mean that removal of these samples would result in negligible savings in computational resources; none of the analyses we conducted would be adversely affected by duplicated samples (with the exception of the TMRCA analysis shown in Figure 6.2); and inspecting the duplicated samples revealed disagreements in genotyping and phasing. To perform this analysis, we examined haplotype data for the approximately 1.1 million sites used to infer tree sequences from the TGP, HGDP, and SGDP datasets on the long arm of Chromosome 20 (using the GRCh38 build). Across all 154 duplicated samples, the mean proportion of genotype calls that are incompatible is 0.04% (standard deviation 0.22%). This figure was found by examining the variant sites shared between each pair of duplicated individuals and computing what proportion of these variants had differing numbers of derived alleles. This discrepancy can be explained by the fact that some of the duplicated samples were sequenced using different libraries, that different variant calling pipelines were used, by the imputation of missing genotypes in HGDP individuals, and also potentially by the effects of lifting over SGDP individuals from GRCh37 to GRCh38.

We determined the number of switches between phasing configurations for each pair. The mean across all pairs is 783.37 switches (standard deviation 570.34). Since individuals have different numbers of heterozygous sites, and thus more opportunities for switches between phasing configurations, I divided the number of switches by the number of heterozygous sites in each individual. This provides the proportion of heterozygous sites where the phasing configuration switches. The average value across pairs is 3.94% (standard deviation 1.93%), with a maximum of 12.05% in individual HGDP00457 from the Biaka population in Africa and a minimum of 1.56% in individual HGDP01018 from the Karitiana population in the Americas. The greatest amount of phasing inconsistency was observed in the 30 duplicated African samples, where the phasing configuration switched at an average of 7.12% heterozygous sites, while the 11 samples from the Americas showed the greatest consistency with a switch of phasing configuration at 2.22% of heterozygous sites.

6.3.2 Results

We first merged the modern datasets and inferred a tree sequence for each autosome using `tsinfer` version 0.2. This approach uses a reference panel of inferred ancestral haplotypes to impute missing data at the 95.6% of sites that have at least one missing genotype. We then estimated the age of ancestral haplotypes with `tsdate` (see Chapter 4). We identified 6,412,717 variants present in both ancient and modern samples. For each, a lower age bound is provided by the estimated archaeological date of the oldest ancient sample in which the derived allele is found. Where this was inconsistent with the initial inferred value (559,431, or 8.7% of variants) we used the archaeological date as the variant age.

Next, we integrated the Afanasievo family and four archaic sequences with the modern samples and re-inferred the tree sequences using the iterative approach outlined in Figure 4.7b. The integrated tree sequences of each autosome together contain 26,958,720 ancestral haplotypes, 231,073,278 edges, 91,172,114 variable sites, and 245,631,834 mutations. We infer that 38.7% of variant sites require more than one change in allelic state in the tree sequence to explain the data. This may

indicate either recurrent mutations or errors in sequencing, genotype calling, or phasing, all of which are represented by additional mutations in the tree sequence. If we discount mutations affecting only a single sample (indicative of sequencing error) we find that 13,513,873 sites contain at least two mutations affecting more than one sample, implying up to 17.5% of variable sites could be the result of more than one ancestral mutation. To evaluate whether the pattern of multiple mutations at variant sites recapitulates known biological signals, we examined CpG vs non-CpG sites as well as transitions vs. transversions. We find that C to T mutations at CpG sites are $\sim 1.72x$ enriched for carrying multiple mutations, while transitions are $\sim 1.19x$ enriched over transversions. While these values are less than would be predicted by the observed mutation rates at these sites (Nachman and Crowell, 2000; Graur and Li, 2000), they indicate that the inferred tree sequence reflects the mutagenic signals of different mutation classes.

Figure 6.1 shows that a high proportion of sites with over $\sim$ 100 mutations on chromosome 20 have sequencing or alignment quality issues as defined by the TGP accessibility mask (1000 Genomes Project Consortium, 2015), or are in minimal linkage disequilibrium to their surrounding sites, suggesting they are largely erroneous. We chose to retain such sites to enable recovery of input data sources; however, future iterative approaches to the removal of such probable errors are likely to improve use cases such as imputation. Sites with over 100 mutations that are both in the TGP accessibility mask and linked to surrounding variants may still be erroneous. For instance, this subset could include clusters of "pseudovariants" in linkage with one another. It is also possible that genuine variants flanked by psuedovariants are poorly described by the marginal trees inferred by `tsinfer`. Specifically, pervasive error at sites surrounding a genuine variant could hinder the reconstruction of the focal ancestral haplotype, leading to erratic haplotype matching.

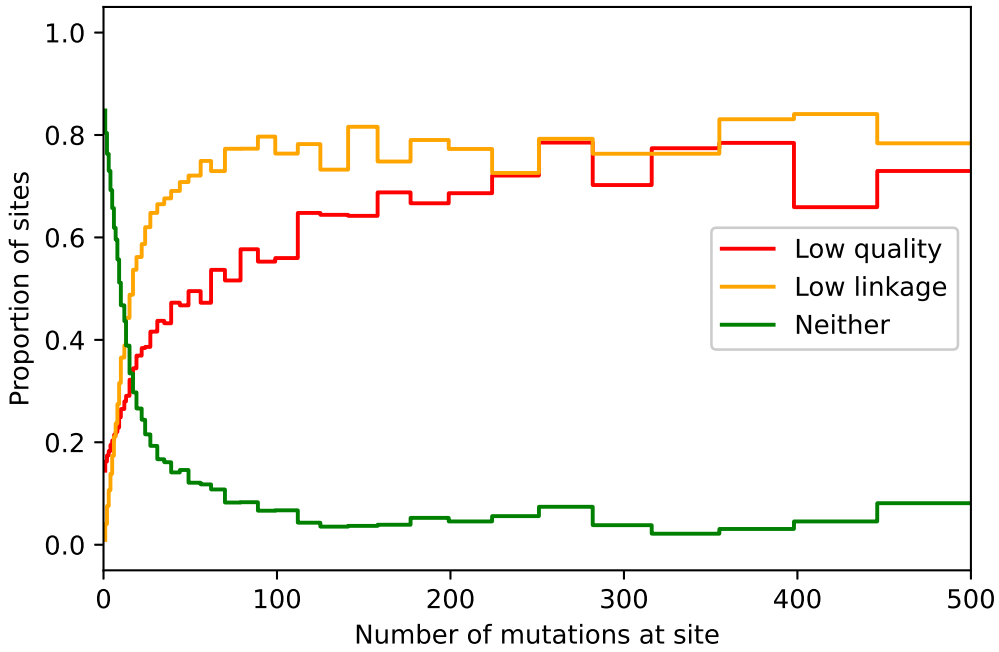


Figure 6.1: The relationship between metrics of variant-calling error and the number of mutations in the inferred tree sequence. For the combined tree sequence of HGDP, TGP, SGDP, and ancient samples, the proportion of sites in two categories - low quality and low linkage, binned by the number of mutations at the site. Low quality is defined by the TGP strict accessibility mask. Sites that fail any of the accessibility mask filters such as low coverage or low mapping quality are marked as low quality. Low linkage is defined by summing the linkage disequilibrium (r^2) for the 50 sites either side of the focal site. If this quantity is less than 10 the site is marked as low linkage.

6.4 Fine-scale worldwide population structure

To characterise fine-scale patterns of relatedness between the 215 populations of the constituent datasets, we calculated the time to the most recent common ancestor (TMRCA) between pairs of haplotypes from these populations at the 122,637 distinct trees in the tree sequence ($\sim$ 300 billion pairwise TMRCAs).

6.4.1 Methods

The full distribution of TMRCAs was computed by iterating over the trees in the tree sequence and at each tree calculating the TMRCA between pairs of chromosomes using the `mrca` function implemented in `tskit`. For efficiency, we down-sampled the chromosomes by randomly selecting up to ten samples from each population.

In populations with fewer than 10 samples, all samples are used. We then find the weighted logarithmic average of all TMRCAs associated with each population combination, weighted by the span of the tree associated with each TMRCA. The logarithmic average is used as we expect the variance of TMRCAs to increase substantially with age: all date values are log transformed before analysis for this reason. The constrained age estimates produced by `tsdate` were used in this analysis (see Section 4.2.2).

6.4.2 Results

After performing hierarchical clustering on the average pairwise TMRCA values, we find that samples do not cluster by data source (which would indicate major data artefacts), but reflect known patterns of global relatedness (Figure 6.2 and Interactive Figure 1). We conclude that our method of integrating datasets is therefore robust to biases introduced by different datasets. Numerous features of human history are immediately apparent, such as the deep divergence of archaic and modern humans, the effects of the Out of Africa event (Figure 6.2 (i)), and a subtle increase in Oceanian/Denisovan MRCA density from 2,000-5,000 generations ago (Figure 6.2 (ii)). Multiple populations show recent within-group TMRCAs, suggestive of recent bottlenecks or consanguinity. The most extreme cases are for "populations" for which only a single individual exists in our dataset, such as the four archaic individuals, and a Samaritan individual from the SGDP. We find the Samaritan has a logarithmic average within-group TMRCA of $\sim$ 1,000 generations, which is caused by multiple MRCAs at very recent times (Figure 6.2 (iii)) and is consistent with documented evidence of a severe bottleneck and extensive consanguinity in recent centuries (Bonné, 1963). Indigenous populations in the Americas, an Atayal individual from Taiwan, and Papuans also exhibit particularly recent within-group TMRCAs. It is important to note that deviations from the constant population size and mutation rate assumptions in `tsdate` may create biases in the estimates presented in this chapter. Mutation age estimates may be inflated at highly mutagenic sites such as methylated CpGs, where mutation

rates are significantly higher than 10^{-8} (Nachman and Crowell, 2000). Additionally, `tsdate` may underestimate coalescence times and mutation ages when population sizes are substantially larger than 10,000, the value of N_e passed to the method. This could differentially affect estimates of coalescence and mutation age across populations with different demographic histories. However, Figure 5.6 indicates that `tsdate` provides overall accuracy comparable to the state of the art in simulations performed under a demographic model approximating human history.

The histograms of TMRCAs for each population combination can be found in Interactive Figure 1 (see Appendix 1). Three of these histograms are also shown in Figure 6.2.

6.5 Modern and ancient population relationships

Next, to assess the ability of the unified tree sequence to recover known relationships between ancient and modern populations, we considered the patterns of descent to modern samples from archaic proxy ancestral haplotypes on Chromosome 20. Simulations, detailed in Section 5.3.2, indicate that this approach detects introgressed genetic material from Denisovans at a precision of ∼86% with a recall of ∼ 61%. A number of studies have sought to infer introgressed haplotypes from archaic individuals (Sankararaman et al., 2014; Vernot and Akey, 2014; Vernot et al., 2016; Skov et al., 2020) and one has inferred an ancestral recombination graph from modern and archaic individuals (Hubisz, Williams, and Siepel, 2020). As shown in 5.3.2, our approach identifies tracts of introgressed sequence, particularly regions where sampled and introgressing archaic individuals share more recent common ancestry, and requires no assumptions about the nature of introgression.

6.5.1 Methods

We evaluate descent from the Afanasievo family, Denisovan, and Vindija Neanderthal among younger samples in Figure 6.3, 6.4, and 6.5. The simplified tree sequence of 3,754 individuals on chromosome 20 (with unary nodes retained) is used for this analysis. Regions of over 1 Mb without variant sites were trimmed from the

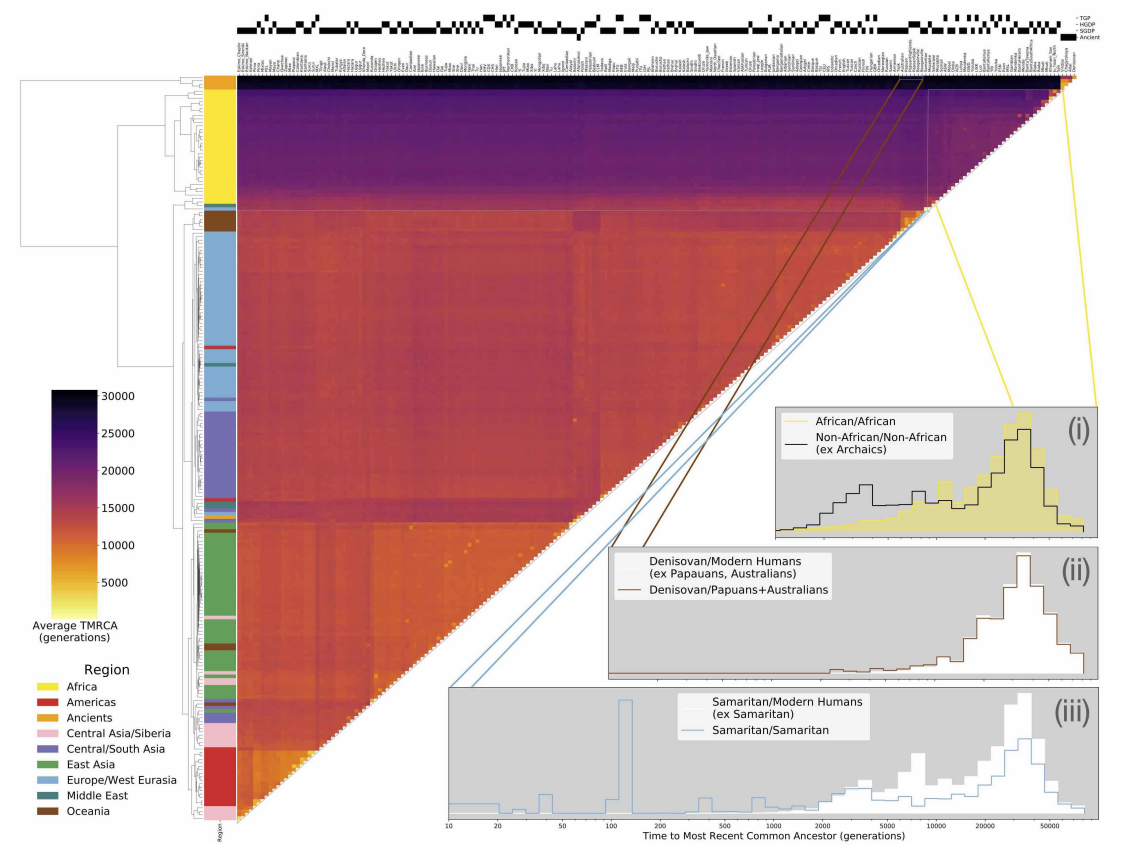


Figure 6.2: Clustered heatmap showing the average time to the most recent common ancestor (TMRCA) on Chromosome 20 for haplotypes within pairs of the 215 populations in the HGDP, TGP, SGDP, and ancient samples. Each cell in the heatmap is coloured by the logarithmic mean TMRCA of samples from the two populations. Hierarchical clustering of rows and columns has been performed using the UPGMA algorithm on the value of the pairwise average TMRCAs. Row colours are given by the region of origin for each population, as shown in the legend. The source of genomic samples for each population is indicated in the shaded boxes above the column labels. Three population relationships are highlighted using span-weighted histograms of the TMRCA distributions: (i) average distribution of TMRCAs between all non-African populations (black line) compared to African/African TMRCAs (solid yellow). (ii) Denisovans and Papuan/Australian TMRCAs (solid line), compared to Denisovans against all non-archaic populations (solid white). The subtle but unique signal is particularly evident in Interactive Figure 1 at `https://awohns.github.io/unified_genealogy/interactive_figure.html`). (iii) TMRCAs between the two Samaritan chromosomes (solid line), compared to the Samaritans/all other modern humans (solid white). Duplicate samples appearing in more than one modern dataset are included in this analysis.

inferred tree sequence, as well as regions before the first variant site and after the last variant site. We establish descent using the *proxy nodes* associated with each ancient sample in the inferred tree sequence, as detailed in Section 4.3.

Descent from ancient individuals is described using two statistics. First, the

genomic descent statistic defined in Scheib et al. (2019) is used to evaluate the overall amount of genetic material in each population which descends from the proxy ancestors associated with ancient individuals. The results of this analysis are shown for the Denisovan (Figure 6.3), Afanasievo (Figure 6.4a), and Vindija (Figure 6.5). Second, we split chromosomes into 1 kb chunks and assess descent from the ancient individuals in each chunk, as described in Section 5.3.2. Pearson product-moment correlation coefficients were calculated using `Numpy` (Harris et al., 2020) for the $n \times c$ matrices of descent, where n is the number of descendants and c is the number of 1 kb blocks. The correlation coefficients were then hierarchically clustered using the UPGMA algorithm implemented in `Scipy` (Virtanen et al., 2020). The results of this analysis are shown for the Denisovan (Figure 6.3c) and Afanasievo (Figure 6.4b).

6.5.2 Results

We find that there are descendants among non-archaic individuals, including both modern individuals and the Afanasievo, for 13% of the span of the Denisovan proxy haplotypes on Chromosome 20. The highest degree of descent among modern humans is in Oceanian populations as previously reported (Reich et al., 2010; Reich et al., 2011; Meyer et al., 2012; Jacobs et al., 2019) (Figure 6.3). However, the tree sequence also reveals how both the extent and nature of descent from the Denisovan proxy ancestors varies greatly among modern humans (Figure 6.3). In particular, we find that Papuans and Australians carry multiple fragments of Denisovan haplotypes that are largely unique to the individual. In contrast, other modern descendants of Denisovan proxy ancestors have fewer Denisovan haplotype blocks which are more widely shared, often between geographically distant individuals. While at the continental scale Oceanian samples show the closest links to the Denisovan, Figure 6.3c indicates that the non-Oceanian individuals with the most Denisovan-associated haplotypes are predominantly from South and East Asia, as well as the Americas. This finding is in agreement with other studies (Browning et al., 2018; Jacobs et al., 2019).

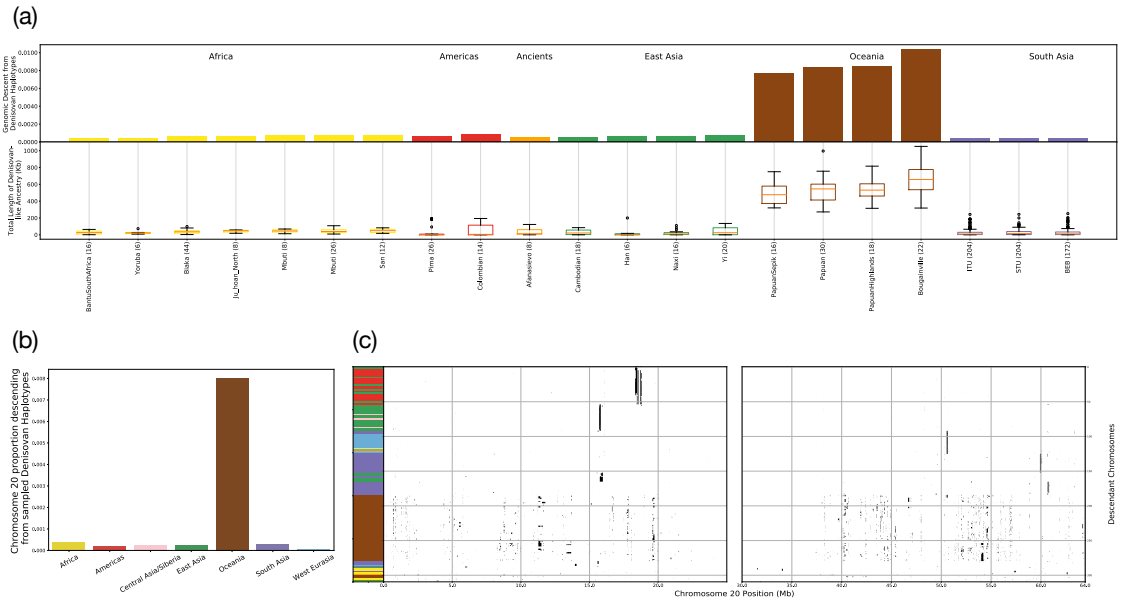
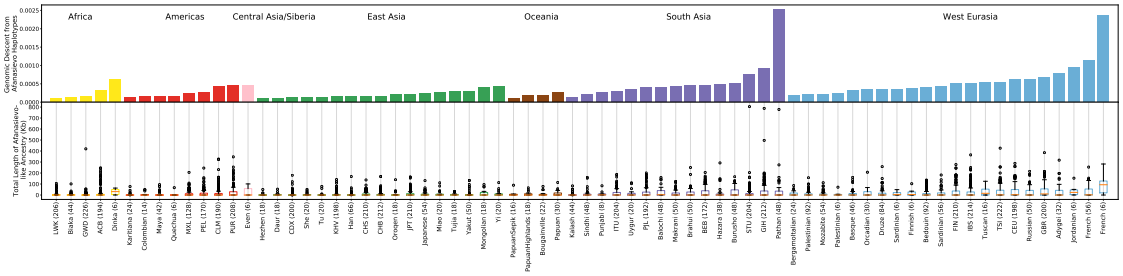


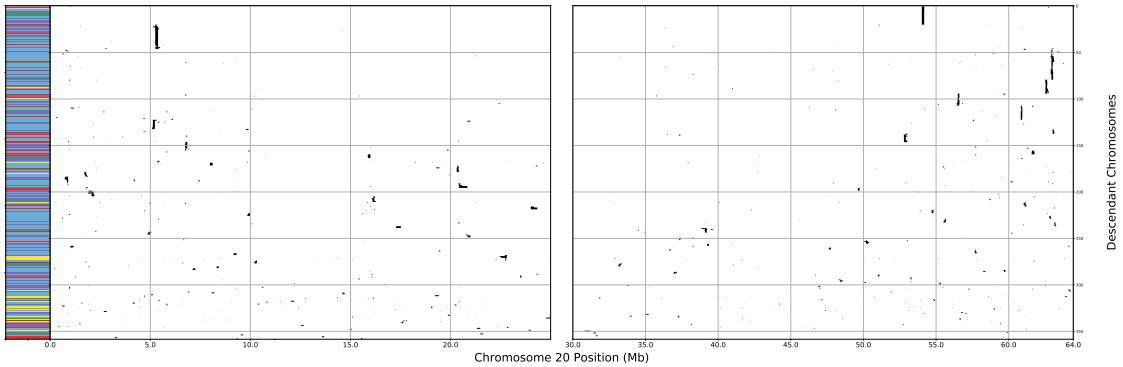
Figure 6.3: (a) Descent from Denisovan proxy ancestors on chromosome 20. The upper portion shows the proportion of genetic material in each population that descends from the proxy Denisovan ancestors (the *genomic descent* statistic(Scheib et al., 2019)). Box plots in the lower panel show the distribution of descendant material among samples from the corresponding population; box plot elements are defined in Figure 6.4a. Only populations with a mean *genomic descent* value of $\geq 0.04\%$ are shown. (b) Percentage of chromosome 20 for modern samples in each region that is inferred to descend from "proxy ancestors" associated with the sampled Denisovan haplotypes, calculated using the *genomic descent* statistic (Scheib et al., 2019). (c) Tracts of descent along chromosome 20 descending from Denisovan "proxy ancestors" in modern samples with at least 100 kilobases (kb) of total descent (colour scheme as for Figure 6.2).

For the Afanasievo family, we find the greatest amount of descent from their proxy ancestral haplotypes among individuals in Western Eurasia and South Asia (Figure 6.4a), consistent with findings from the genetically similar Yamnaya peoples (Narasimhan et al., 2019). Notably, the most frequent descendant blocks in Figure 6.4b all contain geographically disparate modern samples. These cosmopolitan patterns of descent support a contemporaneous diffusion of Afanasievo-like genetic material via multiple routes (Narasimhan et al., 2019).

For the Neanderthal samples, where there are three samples of different age, our simulations indicate that interpretation of the descent statistics is complicated by varying levels of precision and recall among lineages. Nevertheless, recall is highest at regions where introgressing and sampled archaic lineages share more recent



(a) Descent from proxy Afanasievo ancestors on chromosome 20 among HGDP, SGDP, and TGP populations. The upper panel shows the proportion of genetic material in each population that descends from the proxy Afanasievo ancestors (using the *genomic descent* statistic(Scheib et al., 2019)). Only populations with at least two individuals and a mean *genomic descent* value of $\geq$ 0.01% are shown. Box plots in the lower panel show the distribution of descendant material among samples from the corresponding population. The box indicates the inter-quartile range (IQR) of the total length of descendant material within each population. The centre line indicates the median value. The whiskers extend to the minimum and maximum values within 1.5 IQR above the third quartile and below the first quartile. Diamonds indicate outlier data points greater than 1.5 IQR above the third quartile.



(b) Patterns of descent among samples with at least 100 kb of material descending from Afanasievo proxy ancestors. Each row is a sample and each column is a 1 kb section of Chromosome 20. The region of origin of each sample is coloured on the left side of the row.

Figure 6.4: Inferred patterns of descent from Afanasievo proxy ancestors on Chromosome 20. The colour scheme for each region is the same as in Figure 6.2.

common ancestry and precision is higher for the most recent samples. Examining patterns of descent from the Vindija across autosomes indicates that modern non-African groups carry similar levels of Vindija-like material (see Figure 6.5), supporting suggestions that the proportions are very similar between East Asians and West Eurasians (Chen et al., 2020) and inconsistent with other reports (Wall et al., 2013; Prüfer et al., 2017).

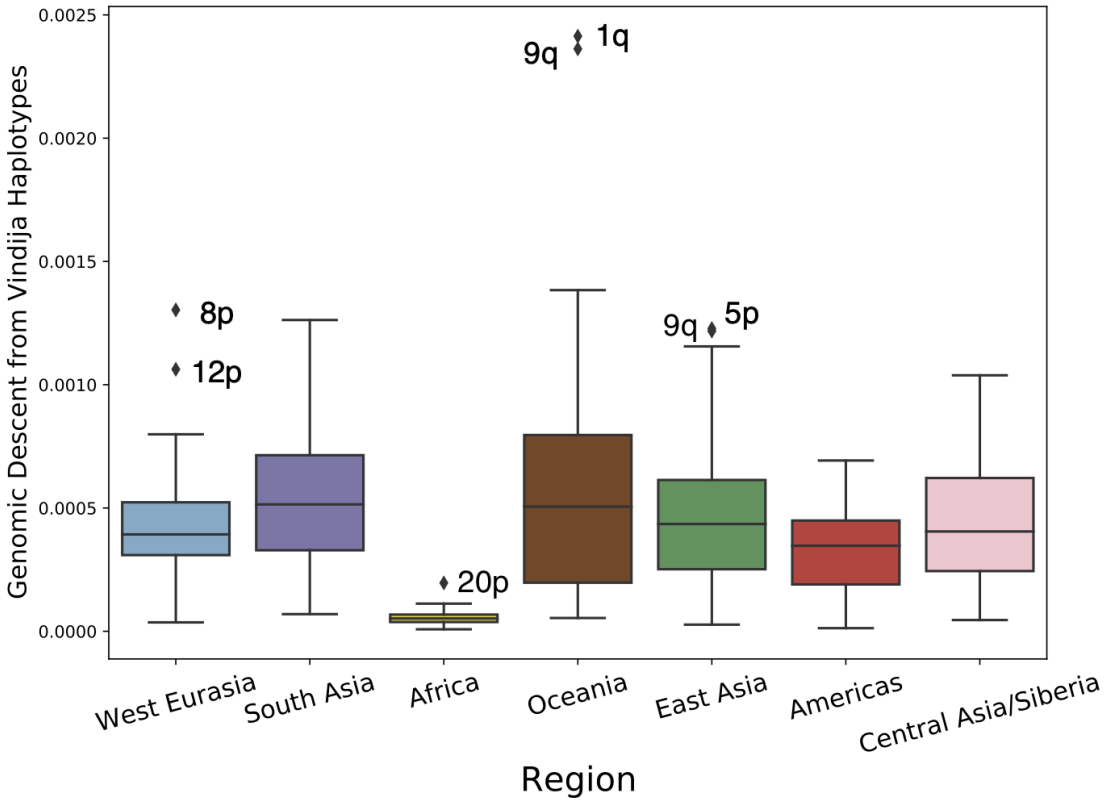


Figure 6.5: Descent from Vindija Neanderthal proxy haplotypes. Each data point is a *genomic descent* statistic calculated from the inferred tree sequence of the arms of each autosome (Scheib et al., 2019). The statistic gives the proportion of genetic material in modern samples from each region that descends from Vindija proxy haplotypes. Box plot elements are defined in Figure 6.4a. Outliers are labelled.

6.6 Non-parametric inference of spatio-temporal dynamics in human history

Tree-sequence based analysis of ancient samples demonstrates the power of the approach for characterising patterns of recent descent. To assess whether we could use the tree sequence to capture wider patterns in human history we developed a simple estimator of ancestor spatial location. We use the location of descendants of a node, combined with the structure of the tree sequence, to provide an estimate of ancestor location. The approach can use information on the location of ancient samples, though it does not attempt to capture the geographical plausibility of different locations and routes. The inferred locations are thus a model-free estimate of ancestors' location, informed by the tree sequence topology and geographic

distribution of samples. Although the relationships between genealogies and spatial structure has been an active area of research in both phylogenetics and population genetics for many years (Wright, 1943; Malécot, 1948; Felsenstein, 1975; Barton, Kelleher, and Etheridge, 2010; Battey, Ralph, and Kern, 2020; Lemey et al., 2009), our approach is the first to infer ancestral locations incorporating recombination. More sophisticated methods which also use genome-wide genealogies are currently in development, and show considerable promise (Osmond, M., and Coop, G., personal communication).

6.6.1 Methods

The latitude and longitude coordinates of individual samples are provided for SGDP individuals, while the location of sampling centres were used for the HGDP individuals. No geographic information was provided for TGP individuals, so these were not used in location inference. The coordinates of the archaeological sites associated with the Afanasievo and archaic individuals were provided.

The weighted center of gravity is determined for each ancestral node by walking up the tree sequence, visiting child nodes before their parents in the same pattern as the previously described inside algorithm (see Section 4.2.2). At each focal ancestral node, we find the geographic midpoint between each of the children of that node. The following simple approach was used to find the geographic midpoints. Latitude and longitude coordinates were converted to Cartesian coordinates (this assumes a spherical Earth). We find the average of the cartesian coordinates and then convert this back to latitude and longitude. Weights are associated the coordinates of the child nodes proportional to the total span of edges where the child descends from the focal ancestral node. This method is highly efficient, requiring less than 1 minute to compute on the combined tree sequence.

6.6.2 Results

We applied the method to the unified tree sequence, excluding TGP individuals. We find that the inferred ancestor location recovers multiple key events in human

history (Figure 6.6, Video 1). Despite the fact that the geographic centre of gravity of all sampled individuals is in Central Asia, by 72 kya the average location of ancestral haplotypes is in Northeast Africa and remains there until the oldest common ancestors are reached. Indeed, the inferred geographic centre of gravity of the 100 oldest ancestral haplotypes (which have an average age of ~ 2 million years) is located in Sudan at 19.4 N, 33.7 E. These findings reflect the depth of African lineages in the inferred tree sequence and are compatible with well-dated early modern human fossils from eastern and northern Africa (McDougall, Brown, and Fleagle, 2005; Hublin et al., 2017). We caution that our sampling of Africa is inhomogenous, and it is likely that if instead we analysed data from a grid sampling of populations in Africa the geographic centre of gravity of independent lineages at different time depths would shift. In addition, past major migrations such as the Bantu and Pastoral Neolithic expansions, both occurring within the last few thousand years, mean that present day distributions of groups in Africa and elsewhere may not represent ancestral ones, and thus the approach of using the present-day geographic distribution to provide insight is likely to give a distorted picture of ancient geographic distributions (Kalkauskas et al., 2021). Nevertheless, this analysis demonstrates that the deep tree structure is geographically centred in Africa in autosomal data, just as it is for mitochondrial DNA and Y chromosomes (Vigilant et al., 1991; Underhill et al., 2001).

Traversing towards the present, by 280 kya, the centre of gravity of ancestors is still located in Africa, but many ancestors are observed in the Middle East and Central Asia and a few are located in Papua New Guinea. At 140 kya, more ancestors are found in Papua New Guinea. This is almost 100 kya before the earliest documented human habitation of the region (O'Connell and Allen, 2015). However, our findings are potentially consistent with the proposed timescales of deeply diverged Denisovan lineages unique to Papuans (Jacobs et al., 2019). At 56 kya, some ancestral lineages are observed in the Americas, much earlier than the estimated migration times to the Americas (Llamas et al., 2016). This effect is likely attributable to the presence of ancestors who predate the migration

and did not live in the Americas, but whose descendants now exist solely in this region (Moreno-Mayar et al., 2018); the same effect may also explain the ancient ancestors within Papua New Guinea. Additional ancient samples and more sophisticated inference approaches are required to distinguish between these hypotheses. Nevertheless, these results demonstrate the ability of inference methods applied to tree sequences to capture key features of human history in a manner that does not require complex parametric modelling.

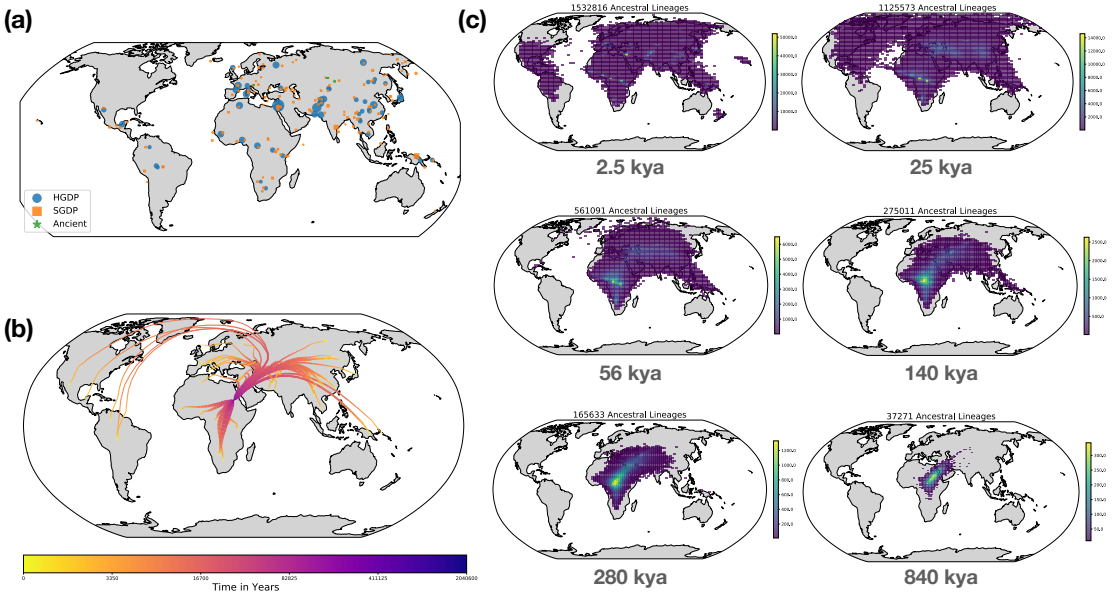


Figure 6.6: Visualisation of the non-parametric estimator of ancestor geographic location for HGDP, SGDP, Neanderthal, Denisovan, and Afanasievo samples. (a) Geographic location of samples used to infer ancestral geography. The size of each symbol is proportional to the number of samples in that population. (b) The average location of the ancestors of each HGDP population from time t=0 to $\sim$ 2 million years ago. The width of lines is proportional to the number of ancestors of each population over time. The ancestor of a population is defined as an inferred ancestral haplotype with at least one descendant in that population. (c) 2d-histograms showing the inferred geographical location of HGDP ancestral lineages at six time-points. Histogram bins with fewer than 10 ancestors are not shown. Link to ancestral location video: `https://www.youtube.com/watch?v=AvV0zBSdxsQ&feature=youtu.be`.

6.7 Discussion

A central theme in evolutionary biology is how best to represent and analyse genomic diversity in order to learn about the processes, forces and events that

have shaped history. Historically, many modelling approaches focused on the temporal behaviour of individual mutation frequencies in idealised populations (Fisher, 1930; Wright, 1931). In the last 40 years there has been a shift toward modelling techniques that focus on the genealogical history of sampled genomes and that can capture the correlation structures in variation that arise in recombining genomes (Kingman, 1982; Hudson, 1983a). Critically, while allele frequency is an idealised and unknowable quantity, there does exist a single, albeit extremely complex, set of ancestral relationships that, coupled with how mutation events have altered genetic material through descent, describes what we observe today.

However, while the empirical generation of data has transformed our ability to characterise genomic variation and relatedness structures in humans, including that of ancient individuals, developing efficient methods for inferring the underlying genealogy has proved challenging (McVean and Cardin, 2005b; Rasmussen et al., 2014). Recent progress in this area (Kelleher et al., 2019; Speidel et al., 2019), on which we build, has been driven by using approximations that capture the essence of the problem but enable scaling to population-scale and genome-wide data sets. The methods described here produce high quality dated genealogies that include thousands of modern and ancient samples. These genealogies cannot be entirely accurate, nevertheless, they enable a wealth of novel analyses that reveal features of human evolution (Harris, 2019; Ralph, Thornton, and Kelleher, 2020; Scheib et al., 2019; Stern et al., 2021). That our highly simplistic estimator of spatio-temporal dynamics of ancestors of modern samples captures key events, such as an East-African genesis of modern humans, introgression from now-extinct archaic populations in Asia and historical admixture (Kelleher et al., 2019), suggests that more sophisticated approaches, coupled with the ongoing program of sequencing ancient samples, will continue to generate new insights into our history.

Moreover, because the tree sequence approach captures the structure of human relationships and genomic diversity, it provides a principled basis for combining data from multiple different sources, enabling tasks such as imputing missing data and identifying (and correcting) sporadic and systematic errors in the underlying

data. Our results identified different types of error common in reference data sets (erroneous sites and genotyping error) as well as emphasising the importance of recurrent mutation in generating human genetic diversity (Michaelson et al., 2012; Nesta, Tafur, and Beck, 2020). Although additional work is required to correct such errors, as well as integrate other types of mutation, notably structural variation, a reference tree sequence for human variation - along with the tools to use it appropriately (Ralph, Thornton, and Kelleher, 2020; Kelleher et al., 2019) - potentially represents a basis for harmonising a much larger and wider set of genomic data sources and enabling cross data-source analyses. We note that such a reference tree-sequence could also enable data sharing and even privacy-preserving forms of genomic analysis (Bonomi, Huang, and Ohno-Machado, 2020) through compression of cohorts against such a reference structure.

There exists much room for improvement in the methods introduced here, as well as new opportunities for genomic analyses that use the dated tree-sequence structure. For example, our approach requires phased genomes, which is a particular challenge for ancient samples that typically pick random reads to create a "pseudo-haploid genome" (Green et al., 2010). However, it should be possible to use a diploid version of the matching algorithm in `tsinfer` to jointly solve phasing and imputation (see Section 7.2.3). This also has the potential to alleviate biases introduced by using modern and genetically distant reference panels for ancient samples (Hui et al., 2020). Recent work focusing on inferring genealogies for high-coverage ancient samples, and using mutations dated in such a genealogy to infer relationships of lower coverage samples through time, offers an alternative strategy for accommodating the unique challenges of ancient DNA in this context (Speidel et al., 2021). In addition, there are many possible approaches for improving the sophistication of spatio-temporal ancestor inference.

7 Discussion

Contents

7.1 *GDF5* and the unified genealogy 145
7.1.1 Marginal tree describing *rs4911178* 145
7.1.2 The age of *rs4911178* 146
7.1.3 Fine-scale genetic relationships at this locus 147
7.1.4 Possible recurrent mutations at *GDF5* 5' UTR 148
7.1.5 Conclusion 150
7.2 Future Directions 151
7.2.1 Data cleaning and annotation 152
7.2.2 Probabilistic ancestor generation algorithm 153
7.2.3 Phasing using a diploid Li and Stephens algorithm . . . 154
7.2.4 Tree sequences and fine-mapping 156
7.2.5 Tree sequences and data sharing 156
7.2.6 Conclusion 157

In this thesis, I introduce, validate, and apply novel methods for inferring dated tree sequences from modern and ancient samples, culminating in a unified genealogy of 7,524 sequenced genomes. This work represents a first draft of the full reference genealogy of humanity: a resource encoding how every human who has ever lived is related to all others at every location in the genome.

Inferring this reference genealogy can be seen as a foundational goal of population genetics. The genealogy would not only serve as a complete record of human genetic variation, but also of coalescence, mutation and recombination events. While

we will never have access to the genomes of everyone who ever lived, increasing numbers of modern and ancient genomes could be incorporated into the genealogy as they are sequenced. As shown in Chapter 6, tree sequences facilitate this type of large-scale data integration, and analyses in Kelleher et al. (2019) suggest that the resulting data structure could be surprisingly compact. The inclusion of more and more samples would also incrementally improve the fidelity of ancestral haplotypes, as well as aid the inference of past human demography. Beyond serving as a definitive evolutionary record, such a structure could provide a powerful platform for manifold downstream applications such as phasing and imputation. If phenotypic information is incorporated, the reference genealogy could aid tasks such as genetic association and fine-mapping as well.

The work presented in this thesis is a step towards this reference genealogy of humanity. In Chapter 6, I show how the unified genealogy of eight modern and ancient datasets contains information on population relationships and descent from ancient samples, and can also be used as the basis of novel analyses such as the estimation of ancestor locations. In this chapter, I return to the motivating example from Chapter 1 to demonstrate that the unified genealogy also provides increased resolution when examining the evolutionary history of loci of interest.

However, moving closer to the goal of a reference genealogy necessitates improvements to these methods as well as support for an array of downstream analyses. In Sections 4.4 and 5.5, I enumerated possible improvements to `tsdate`, `tsinfer`, and the iterative approach. In Section 7.2, I explore additional promising future directions. These include a programme to "clean" and annotate input data, infer improved ancestral haplotypes, and implement a diploid version of the `tsinfer` HMM to support phasing. I also explore approaches by which the genealogy could be used for downstream applications such as fine-mapping causal variants and consider the prospects for using tree sequences as a secure means of sharing confidential information. In Section 7.2.6, I discuss how these various future directions may be prioritised.

7.1 *GDF5* and the unified genealogy

The motivating example in Chapter 1 concerned the evolutionary history of derived alleles in a downstream enhancer of the gene *GDF5*. In Capellini et al. (2017), the authors focused on the common G/A SNP *rs4911178* in this locus. The derived allele of this variant is associated with decreased height and increased risk of osteoarthritis (Miyamoto et al., 2007; Sanna et al., 2008; Loughlin, 2015), and is highly conserved in placental mammals. Using the composite of multiple signals method (Grossman et al., 2013) as well as experiments showing that the derived allele attenuates bone growth signalling pathways, they argue that this SNP is a causal driver of observed signals of positive selection in this region. They investigated the history of this variant using TGP, Neanderthal, and Denisovan samples to define haplotypes and built a maximum likelihood phylogenetic tree over a 140 kb region that includes *rs4911178*. This analysis revealed two distinct haplotypes: the "A" haplotype with the ancestral allele of *rs4911178* and the "B" haplotype with the derived allele. Within the B haplotype, they also identified a "B*" haplotype at high frequency in non-Africans. The authors concluded that the derived allele was likely present in an ancestral African population prior to the splitting of archaic lineages and hypothesise that positive selection for the derived allele may have occurred repeatedly in northern latitudes.

7.1.1 Marginal tree describing *rs4911178*

What can the inference methods discussed in this thesis tell us about the history of *rs4911178*? I examined the marginal tree that includes this variant, as well as the trees in the region surrounding the locus. *rs4911178* is at position 20:35,364,817 (build GRCh38); the marginal tree at this position extends for 1.7 kb and contains 3,888 ancestral haplotypes, 55 variant sites, and 80 inferred mutations. As shown in Figure 7.1, there are two clearly distinct clades at this marginal tree, demarcated by the *rs4911178* mutation. Samples in the clade on the left carry the derived allele. This clade resembles characteristic examples of positive selection, with modern samples mostly descending from a long internal branch. The clade on the right

contains samples carrying the ancestral allele and exhibits a typical coalescent genealogy with an abundance of deep coalescence events.

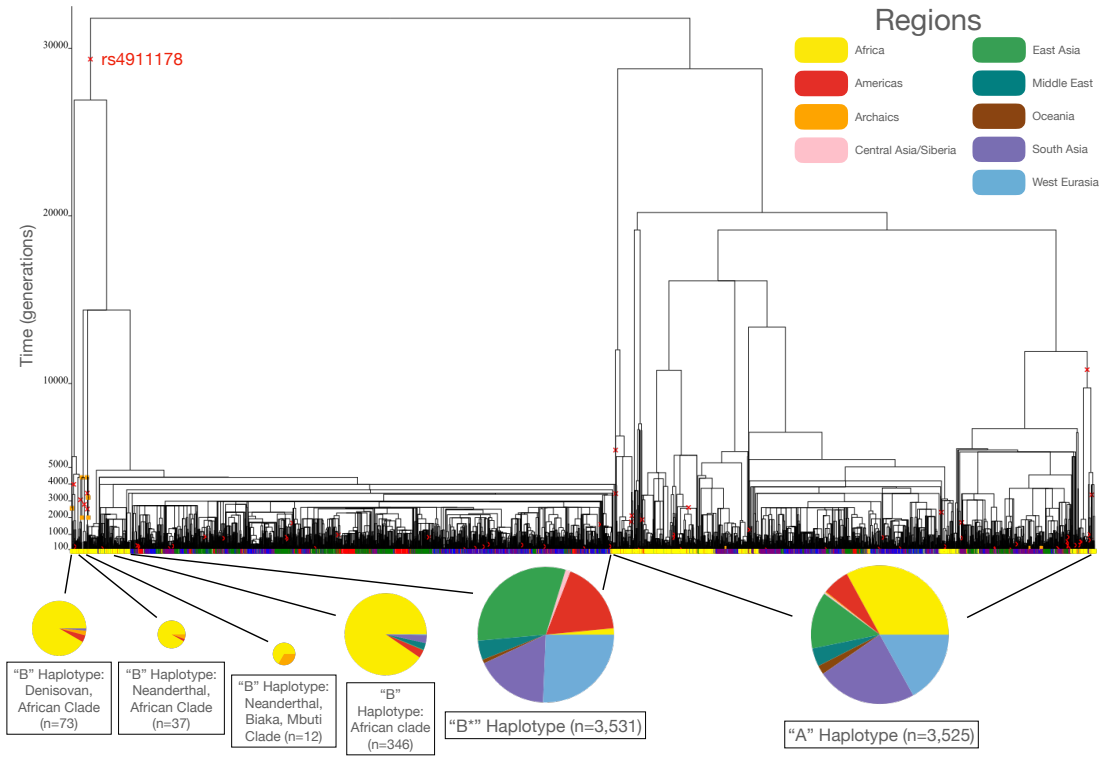


Figure 7.1: Marginal tree from the unified genealogy of Chromosome 20 (positions 35,364,386 to 35,366,101) describing *rs4911178*. The deepest coalescence separates samples carrying the derived allele from those with the ancestral allele. Mutations are marked on the genealogy with a a red "x", with the *rs4911178* mutation labelled. Note that while only variants in the interval of this tree are shown, variants from both downstream and upstream of this locus were used to build the tree. Leaf nodes are coloured by their region of origin, as indicated in the legend. Pie charts describe the membership of five clades. From left to right, the first four clades are representative of the "B" haplotype from Capellini et al. (2017), with the fifth representing the "B*" haplotype. The sixth clade corresponds to the "A" haplotype. Pie chart size is proportional, on a log scale, to the number of samples in the clade.

7.1.2 The age of *rs4911178*

The derived allele of *rs4911178* is observed in both modern and archaic samples, suggesting that it predates the split of these lineages. The unified genealogy supports this point with a single mutation explaining observed patterns of variation at this locus among the 7,508 samples included in this tree sequence. Our spatio-temporal estimator also places the origin of the allele on the African continent.

The mutation is estimated to be very old: the `tsdate` point estimate of the age of the child node of the edge on which the mutation occurred is 26,910 generations, while the estimate for the parent node is 31,790 generations. `GEVA` provides very similar estimates, with a 95% confidence interval of 25,747-33,953 generations. `Relate` provides allele age estimates separately for each TGP population. Across all populations, the mean age estimate is 27,385 generations, ranging from 16,470 generations in the Punjabi population to 42,788 generations in the Gambian population. The `Relate`-estimated minimum lower bound on the branch hosting the mutation is 1,447 generations in the Chinese Dai population; the maximum upper bound is 54,940 generations in the Yoruba population.

7.1.3 Fine-scale genetic relationships at this locus

In Figure 7.1, samples carrying the derived allele at *rs4911178* are divided into five groups. The deepest coalescence within this group is estimated by `tsdate` to occur at 26,910 generations and separates the two Denisovan samples and a group of 71 largely African samples from all other samples carrying the derived allele. In addition to the Denisovans, this group contains 67 modern samples from West African populations, Bantu-speaking populations in Kenya, and African-American and African-Caribbean populations. The remaining samples consist of one South Asian sample from the Makrani population in Pakistan, which has been shown to have recent Sub-Saharan admixture (Laso-Jadart et al., 2017), and three samples from recently admixed Americas populations that likely carry Sub-Saharan African ancestry at this locus.

The other clades of individuals with the derived allele at *rs4911178* are the descendants of a polytomy at an ancestral haplotype estimated to be 14,380 generations old. This polytomy indicates that there is not enough information to determine the branching order of the three groups. The first subtree contains one Altai Neanderthal sample, the two Chagyrskaya Neanderthal samples, and one Vindija Neanderthal sample, as well as 5 samples from the Biaka and Mbuti populations in Central Africa. The second group contains one Altai sample, one

Vindija sample, two Biaka samples, 32 samples from West Africa, Bantu-Speaking peoples in Kenya, and African-Caribbeans, and one Puerto Rican sample.

The remaining group has a TMRCA at 4,830 generations and contains 3,877 samples, 91.5% of which are from non-African populations. In Figure 7.1, I divide the group into two clades. Of the 476 African samples descending from the TMRCA at 4,830 generations, 88.7% coalesce between 3,390 and 4,830 generations. I assign these individuals to a clade labelled ""B" Haplotype African Clade" in Figure 7.1. Non-African samples have a more recent TMRCA: an inferred haplotype at 3,390 generations is ancestral to 98.7% of the 3,523 non-African samples observed to carry the derived allele of *rs4911178*. This haplotype spans $\sim$ 554 kb, including the 140 kb region studied in Capellini et al., 2017. I posit that this haplotype is consistent with the "B*" haplotype defined in Capellini et al. (2017), as it hosts derived alleles at two variants in the 5' untranslated region (UTR) of *GDF5* that the authors used to define membership in this haplotype (discussed in the next section). It is labelled as ""B*" Haplotype" in Figure 7.1.

What do these results tell us about positive selection in this region? Figure 7.2 shows the relationship between allele frequency and age for derived alleles on Chromosome 20q. While the estimated age of *rs4911178* appears in the expected range for alleles of its frequency, the B* haplotype is notably young for an allele at $\sim$ 50% frequency. This is consistent with the effects of positive selection.

7.1.4 Possible recurrent mutations at *GDF5* 5' UTR

Capellini et al. (2017) used two SNPs in the 5' UTR of *GDF5* to determine the B* haplotype: *rs143383* at 20:35,438,203 and *rs143384* at position 20:35,437,976. These SNPs are in strong LD with *rs4911178*, with r^2 values of 0.92 and 0.88 (calculated using East Asian TGP populations). These SNPs are not found in Denisovans or Neanderthals. These variants have also been shown to be strongly associated with shorter height and osteoarthritis risk (Rouault et al., 2010; Liu et al., 2013).

Notably, both are part of CpG dinucleotides, which are known to have an elevated mutation rate (Nachman and Crowell, 2000). Indeed, in the unified genealogy,

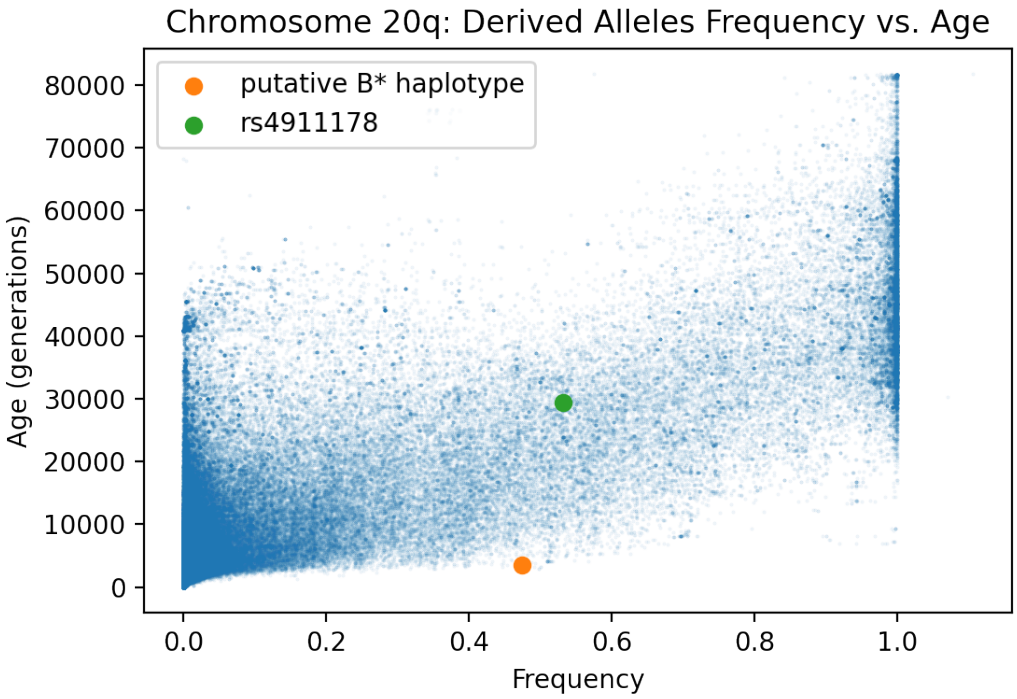


Figure 7.2: Allele frequency and estimated age from `tsdate` for 1,178,347 derived alleles on Chromosome 20q. For alleles with multiple inferred mutations, the age of the oldest mutation is used as the age of the allele. The frequency and age of *rs4911178* and the putative B* haplotype, identified in the text, are highlighted.

multiple mutations are inferred to explain the pattern of variation at these SNPs. For *rs143383*, 13 mutations are inferred, composed of six independent mutations to the derived "A" allele and seven back mutations. Two of the independent mutations and four of the back mutations occur above sample nodes and may thus be indicative of sequencing error. The oldest mutation is estimated to have occurred 2,980 generations ago. For *rs143384*, 18 mutations are inferred, including three independent mutations to the derived "A" allele and 15 back mutations. Two of the independent mutations and 11 of the back mutations occur above samples. The oldest mutation at this site is estimated to have occurred 3,450 generations ago.

While recurrent mutation at these sites is perhaps more likely given that they are part of CpG dinucleotides, the large number of back mutations at this site is unexpected. This may perhaps reflect errors in inference by `tsinfer`, pervasive sequencing error and perhaps other, more complex phenomena. Further evaluation is

clearly warranted: both of these important variants, which appear to be candidates for causal association with the phenotype of interest, and of the reliability of inferring particular recurrent mutations using `tsinfer` version 0.2.

7.1.5 Conclusion

This analysis supports many of the conclusions from Capellini et al. (2017), namely that *rs4911178* was present in the ancestral African population prior to the divergence of archaic and modern lineages, that archaic individuals and some modern Africans share similar haplotypes (though I additionally note differentiation between Denisovans and Neanderthals, as well as among Neanderthals), that essentially all non-African individuals with the derived allele share a very recent TMRCA, and that there are strong signals of positive selection at this locus. However, the fact that the derived allele does not seem to be under positive selection in Africa, despite being present at appreciable frequencies, suggests that environmental factors in Eurasia at the time of the Out of Africa event drove selection at this locus. Alternatively, a more complex explanation involving the effects of other variants in this region, likely on the B* haplotype, may explain the observed signal of positive selection. For instance, *rs143383* and *rs143384* appear to have arisen around the time of the Out of Africa event and have also been implicated as drivers of positive selection in East Asians at the *GDF5* locus (Wu et al., 2012). It thus seems possible that multiple alleles acted in concert on the B* haplotype to produce a phenotype that was a target of positive selection around the time of the Out of Africa event. Understanding the interaction of these derived alleles experimentally, as well as observing the phenotypes of individuals carrying various combinations of derived and ancestral alleles at these variants, may provide evidence for or against this alternative explanation.

Conversely, the fact that all Neanderthal and Denisovan individuals were homozygous for the derived allele at *rs4911178* is perhaps suggestive of an environmental explanation for positive selection on this allele. However, without estimates of the historical allele frequency or an understanding of whether positive selection played

a role at this locus in archaic populations, it is not definitive evidence for or against repeat selection in northern latitudes. In the absence of additional archaic genomes, an inferred tree sequence with larger numbers of African samples, for example from Choudhury et al. (2020), may also help elucidate the history of *rs491178* at deeper time depths. Additionally, more ancient human samples from outside of Africa may help our simple spatio-temporal estimator pinpoint the geographic spread of the B* haplotype, which we estimate to have originated in Central Asia. This could refine estimates of when and where the haplotype became a target of positive selection.

In conclusion, the unified genealogy of modern and ancient samples adds nuance to our understanding of the evolutionary history of *GDF5*. The tree sequence gives a more precise understanding of population relationships across this region, showing how forces such as recombination, selection, demography, migration, and recurrent mutation create the complex patterns observed at this important locus.

7.2 Future Directions

Multiple possible future directions will serve to produce a more accurate and useful reference genealogy of humanity. In this section, I discuss the prospect of adapting these methods to perform data cleaning and annotation by detecting specific genotype errors and recurrent mutations. Expanding on this point, I propose a new, probabilistic approach for generating ancestral haplotypes in the presence of recurrent mutation. I then examine how phasing errors may be identified in a tree sequence and how a diploid version of the Li and Stephens (LS) algorithm could be used to perform phasing. Tree sequences also have potential relevance to the problem of fine-mapping variants causally associated with a trait and I identify a specific use case in existing fine-mapping methods. Finally, I consider how tree sequences may be used as a platform for tasks such as imputation and phasing while retaining the confidentiality of individual-level genotype information.

7.2.1 Data cleaning and annotation

`tsinfer` version 0.2, discussed in Section 1.5.2, uses mismatch terms in the algorithm's core LS HMM to support the inference of multiple mutations at a variant site. This mismatch term can be tuned to adjust the relative probability of inferring a recombination or mutation event when resolving incongruities in haplotype matching. To generate results from simulated and real data in Chapters 5 and 6, equal probabilities were given to recombination and mutation. This choice was informed by a simulation study in the supplementary material of Wohns et al. (2021). This study shows that specifying equal probabilities to mutation and recombination produces the most accurate tree sequence topologies according to various tree distance metrics and indirect measures of inference accuracy. However, these metrics do not penalise the inference of extraneous mutations. A probable example of such mutations in the unified genealogy is shown in Section 7.1.4, though it should be noted that some of these "extraneous" mutations account for sequencing error.

Indeed, preliminary results indicate that this approach overestimates the number of mutations in tree sequences simulated using finite-sites models. This is more pronounced when inferring tree sequences from simulated variation data with genotype error added. However, it appears that mismatch probabilities may be tuned for more reliable identification of erroneous genotypes and recurrent mutation. Another approach that has shown promise is to infer a tree sequence using a lower probability of mismatch before relaying mutations on the topology using a parsimony algorithm. Sporadic mutations that only affect single samples may then be removed before reinferring a tree sequence topology.

Removing genotype errors and identifying recurrent mutations could result in more accurate inferred genealogies. This is supported by Figures 5.3 and 5.5, which suggest that genotype error has a significant impact on inference accuracy. This approach could also improve our understanding of the prevalence of genotype error and the quality of variants in commonly used population sequencing datasets, as indicated by Figure 6.1. Finally, identifying recurrent mutations gives insight into the suitability of the infinite sites model in various datasets. This latter point is

especially relevant to large collections of data such as ExAC (Lek et al., 2016) and Gnomad (Karczewski et al., 2020). The impact of recurrence was considered briefly in the former publication as a potential pitfall in the analysis of very rare variants.

In addition to simulations, various sources of real data may help validate the accuracy of this proposed data cleaning and annotation programme. In the case of genotype error, re-sequencing studies and examination of sample reads should provide empirical validation. Insight into expected patterns of recurrent mutation could be gained by examining *de novo* mutations in family trios sequenced to high coverage and the distribution of multiallelic sites. Variant annotations corresponding to elevated mutation rates, such as CpG sites, microsatellites, or minisatellites in recombination hotspots, may also be informative.

7.2.2 Probabilistic ancestor generation algorithm

The presence of recurrent mutation motivates not only a means of detecting recurrence, but also a more sophisticated ancestor generation algorithm. The heuristics used to construct ancestors in `tsinfer`, described in Section 2.3.1, do not allow for multiple ancestral haplotypes in the case of recurrent mutation. In the following, I propose a model-based, probabilistic means of ancestor reconstruction using `tsinfer` version 0.2 that draws inspiration from the HMM for shared haplotype inference in Albers and McVean (2020). Figure 7.3 provides a schematic of the proposed approach.

At a focal mutation, we can divide our dataset into the set of samples that carry the mutation (the "concordant" set) and the set that does not (the "discordant" set). To estimate the ancestral haplotype on which the mutation arose we can use the copying algorithm in `tsinfer`, where each concordant sample copies from the ancestral haplotype over some defined region with a possibility of error. The ancestral haplotype itself copies the discordant samples, which are a mosaic of even older ancestral material. The process of generating the ancestral haplotype can be considered to be the expectation step in an expectation-maximisation algorithm (EM) within the HMM framework. This model could be extended to

accommodate recurrent mutations by allowing the concordant samples to copy from a haplotype other than the ancestral one. Alternatively, we could fit a model with multiple ancestral haplotypes, where the variant site is a superposition of the two ancestral haplotypes.

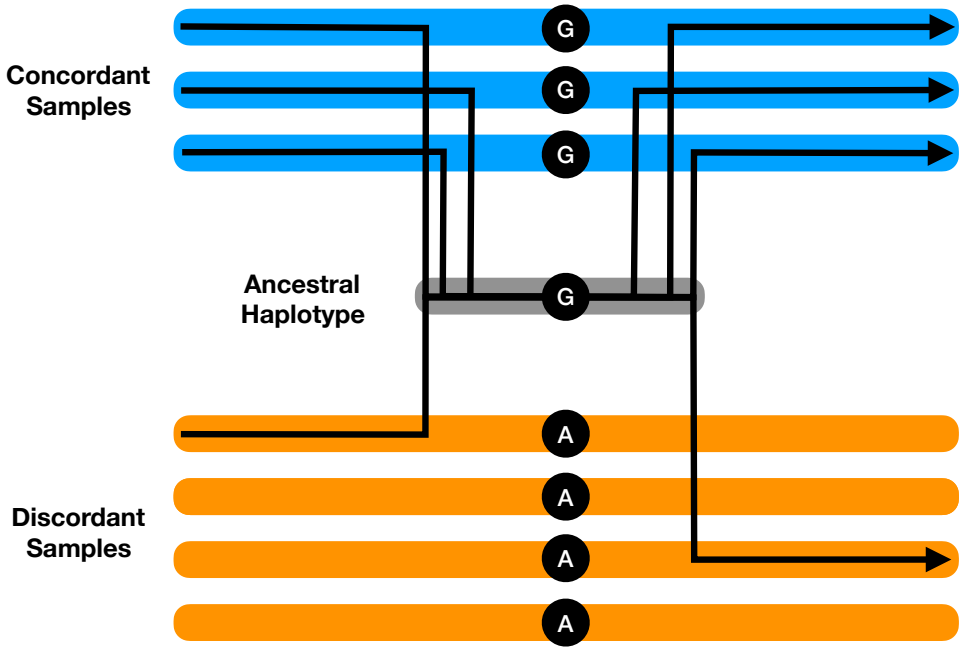


Figure 7.3: Schematic of our proposed probabilistic method of ancestral haplotype inference. The set of concordant samples copies from the ancestral haplotype over some defined region. The ancestral haplotype itself copies from ancestral material present in discordant samples.

7.2.3 Phasing using a diploid Li and Stephens algorithm

In Section 3.3.2, I showed that the LS model implemented in `tsinfer` reveals phasing accuracy in the Afanasievo quartet. In Kelleher et al. (2019), we also showed that some samples had an excess of "close breakpoints" that are potentially indicative of phasing errors (see schematic in Figure 7.4). Briefly, a phasing switch error may be represented in a tree sequence by a swap between the parental nodes

of the maternal and paternal chromosomes of an individual. Such a pattern is highly unlikely in the absence of phasing error.

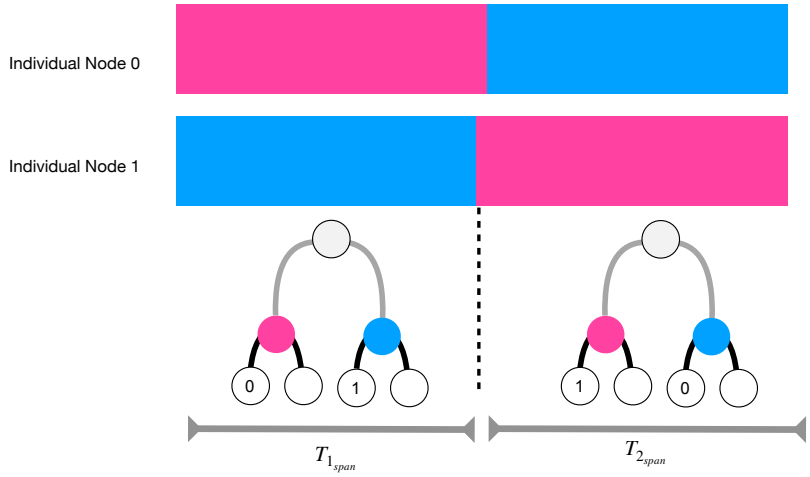


Figure 7.4: Edge breakpoints can be used to identify potential phasing errors. The two rectangles indicate the "phased" chromosomes associated with a diploid individual, coloured by the parental node of the edge where each sample node (individual nodes 0 and 1) is a child. The parental nodes are swapped at a phasing switch error.

While phasing switch errors detected with this approach may be corrected via ad-hoc methods, a more principled path forward is to utilise a diploid version of the LS algorithm with a reference panel composed of inferred ancestral haplotypes from `tsinfer`. Briefly, the currently implemented haploid version of the algorithm determines the most likely copying path for single chromosomes. In a diploid version of the algorithm, we determine the most likely copying path for a pair of chromosomes. Such an approach scales quadratically with the number of reference haplotypes (Lunter, 2018), but exploiting the tree sequence structure should allow a speedup to $\mathcal{O}(\log^2 n)$, analogous to the performance of the haploid LS model implemented in `tsinfer`.

If such an algorithm could be efficiently incorporated into `tsinfer`, it would enable an "adaptive" phasing panel to be used for each sample. This may be particularly beneficial for ancient samples, where only older ancestral haplotypes would be used as references in the phasing step.

7.2.4 Tree sequences and fine-mapping

Fine-mapping, or inferring the set of SNPs that are causally associated with a trait, is an active area of research in statistical genetics. Two recent methods that have been proposed to address this problem, *hierinf* (Renaux et al., 2020) and *KnockoffZoom* (Sesia et al., 2020), hierarchically cluster SNPs before performing inference. The clustering methods used are simplistic, in both cases it is performed using the position of SNPs on the chromosome. Measures of LD are then used to identify groups of SNPs to test. The criticism that "genetic variants do not come in a natural hierarchy" (Wang et al., 2020) has been directed towards this approach. However, genetic variants are indeed organised hierarchically in the tree sequence, which may therefore provide a superior structure for clustering. Minichiello and Durbin (2006) used a similar fine-mapping strategy with ARGs inferred by the method `Margarita` (see Section 1.5). Their results indicate that this direction shows particular promise.

7.2.5 Tree sequences and data sharing

In addition to serving as a platform for imputation, and potentially for phasing as noted in a previous section, the tree sequence offers a means of sharing confidential genotype information. Genomic data confidentiality is an increasingly pressing issue (Bonomi, Huang, and Ohno-Machado, 2020), made more urgent by the ability to identify non-sampled individuals using genetic information from their relatives (Erlich et al., 2018). In an attempt to protect the anonymity of study participants, many datasets are effectively "siloed," with access only available after an extensive application process. Public access is limited to summary statistics including allele frequency that are compiled in data collections such as Gnomad (Karczewski et al., 2020).

Tree sequences have the potential to facilitate enhanced data sharing while preserving individual-level privacy. This could be accomplished by building a tree sequence from both shareable and confidential samples, then removing the non-shareable sample information. The resulting structure would be informed by all

samples while adhering to data usage policies. This could provide a powerful platform for tasks such as imputation and phasing. However, guaranteeing confidentiality with this approach is problematic. It may be difficult to prove that individual level results could not be reconstructed from ancestral haplotypes. This may be particularly challenging when extremely rare variation is encoded in the tree sequence.

7.2.6 Conclusion

How should these future directions be prioritised? One option is to give preference to approaches that have the potential to improve a variety of downstream applications. The data cleaning and annotation programme of work appears particularly promising in this regard. Removing genotype errors and mapping recurrent mutation could both improve the accuracy of tree sequence inference and test the validity of commonly-used assumptions in population and statistical genetics. These assumptions are that input data is error-free and that variation in commonly used population sequencing datasets is well-described by the infinite sites model. A map of recurrent mutations in the human genome could also have relevance for the use of allele frequency and linkage disequilibrium in statistical genetics (Lek et al., 2016). A related direction is the inference of improved ancestral haplotypes. Supporting the inference of recurrent haplotypes will be helpful for studies of human genetic variation, and essential for the study of pathogens with a high mutation rate such as malaria and HIV. Finally, the proposed programme for fine-mapping is a promising, under-explored direction. However, a more sophisticated approach using the patterns of conditional independence encoded in the tree sequence may prove superior.

Both the phasing and data sharing future directions may require significant methodological development, but with less certain downstream utility. While utilising the inferred ancestral haplotypes as a reference panel is an appealing concept, improvements in this domain will likely be limited by the availability of suitable data to estimate the ancestral haplotypes. With regards to the data confidentiality programme, if it were possible to prove that tree sequences represent a secure means of sharing confidential genetic data, this would represent an important

step towards a practical reference genealogy of humanity. However, the practical challenges towards proving this and spurring widespread uptake are formidable.

Appendices

A

Supplementary Video and Interactive Figure from Wohns et al. (2021)

Supplementary Video 1 from Wohns et al. (2021) is available at: `https://www.youtube.com/watch?v=AvV0zBSdxsQ`. The video shows the estimated geographic locations of ancestors of HGDP, SGDP, Neanderthal, Denisovan, and Afanasievo samples over time. Each dot represents an edge in the tree sequence of chromosome 20, where the time and geographic location of the parent and child nodes of the edge have been estimated. The locations of edges at each point in time are plotted using the great circle between the parent and child nodes. Edges are coloured by the region of the descendants of the child node. If an ancestral lineage has ancestors in multiple regions, its colour is the average of the respective colours of each region.

Supplementary Interactive Figure 1 from Wohns et al. (2021) is available at: `https://awohns.github.io/unified_genealogy/interactive_figure.html`. This dynamic version of Fig. 6.2 shows the span-weighted TMRCA histogram relating each pair of the 215 populations in the integrated tree sequence of chromosome 20. Hovering over each cell displays the relevant histogram, with a vertical bar indicating the logarithmic mean TMRCA.

B

Joint Authorship Statement

The following is a statement of which parts of my thesis are not my own work.

Chapter 1: Introduction

The motivating example is reproduced from Capellini et al. (2017), including Figures 1.2, 1.3, and 1.8. I was not involved in Capellini et al. (2017), but am a co-author on another study which is currently in preparation which focuses on the evolution of this locus, the results of which are described in Chapter 7.

Figure 1.10 is reproduced from Kelleher et al. (2019). I am a co-author of this paper, but did not create the figure.

Jerome Kelleher and Yan Wong created `tsinfer` version 0.1. Jerome Kelleher implemented the support for missing data and multiple mutations in `tsinfer` version 0.2.

Chapter 2: The Relative Age of Alleles

All text, figures, and novel results (including code implementing these analyses) are my work.

Chapter 3: Analysing Tree Sequence Topologies

Section 3.2 is adapted from Kelleher et al. (2019). I was involved in the formulation and application of the *genealogical nearest neighbours* statistic, edited the descriptions and schematic caption, but did not produce Figure 3.1 or parts (a) and (b) of Figure 3.2.

Section 3.3 contains results from Scheib et al. (2019), of which I am a co-author. Collaborators provided variant data from the the Trumpington individuals and performed the PCA analysis shown in Figure 3.3a. The first two paragraphs of Section 3.3.1 include text authored by my collaborators.

In Section 3.3.2, David Reich, Ali Akbari, and Shop Mallick provided variant data from the Afanasievo family and wrote descriptions of the sequencing, variant calling, and phasing. This text appears in Wohns et al. (2021).

Chapter 4: Inferring dated tree sequences from modern and ancient samples

All text and figures in this chapter, save for the worked examples, are adapted from the relevant methods and supplemental text from that manuscript. I wrote all text that appears in this chapter, save for Section 4.3. The original text was written by a co-author, which I edited and revised.

I am the author of the software package `tsdate`, with contributions from Yan Wong. Jerome Kelleher, Yan Wong and I implemented the code in `tsinfer` version 0.2 which supports the iterative approach, while I wrote the `tsdate` code which supports the iterative approach.

Chapter 5: Evaluating the accuracy of dated tree sequences

All text, figures, and novel results are adapted from Wohns et al. (2021) and represent my work, with input from my co-authors.

Chapter 6: A unified genealogy of modern and ancient genomes

All results and text are adapted from Wohns et al. (2021). All text and figures are my work, with input from with my coauthors, save for Figure 6.1 and Supplementary Interactive Figure 1 (`https://awohns.github.io/unified_genealogy/interactive_figure.html`), which were produced by my co-authors.

Chapter 7: Conclusion

All text, figures, and novel results are my work.

C
Lessons Learned

In the following, I offer my personal reflections on a few of the lessons I have learned over the course of my DPhil.

Tree-based thinking simplifies numerous problems in biology

Everything from recombination and linkage disequilibrium to selection and migration can be clearly illustrated using tree sequences. It took a long time, perhaps the first two years of my DPhil, to switch my mental model of genetics from matrices to trees, but I have no doubt it was well worth the effort. Trees are also a powerful platform for inference. I believe `tsdate` is an example of this, and I expect that more methods utilising belief propagation on tree sequences will be introduced in the coming years.

The importance of a thorough understanding of core methods

My early efforts to understand basic features of the coalescent proved invaluable for the rest of my DPhil. With the encouragement of my supervisor Gil McVean, in the first year of my DPhil I attended Professor Simon Myers' lectures in SC1: Stochastic Models in Mathematical Genetics. A fellow DPhil student and I organised regular sessions to go over problem sheets for the course. Furthermore, early in my DPhil I found it immensely helpful to read *Gene Genealogies, Variation and Evolution* (Hein, Schierup, and Wiuf, 2005), *Coalescent Theory* (Wakeley,

2009), *Biological Sequence Analysis* (Durbin et al., 1998), and *Inferring Phylogenies* (Felsenstein, 2004), among other texts.

The value of following best practices in software development

My co-supervisors Jerome Kelleher and Yan Wong taught me the value of following best practices in software development. I learned to write more elegant, well-tested, and well-documented code by building `tsdate` and contributing to the `tskit` software ecosystem with their guidance.

Collaboration provides unexpected insights

Presentations at conferences and lab meetings have sparked novel ideas. For example, collaboration with experts in ancient DNA not only gave me access to immensely interesting datasets, but also facilitated development of our iterative approach. Honing in on the evolutionary history of the *GDF5* locus in collaboration with Terry Capellini has left me with a much better understanding of the mismatch parameters implemented in `tsinfer` version 0.2.

The value of mentors

Great mentors were the key element of my DPhil experience, both for what I learned and for my thorough enjoyment of the course.

References

1000 Genomes Project Consortium (2015). "A global reference for human genetic variation". *Nature* 526.7571, pp. 68–74.

Adrion, Jeffrey R, Christopher B Cole, Noah Dukler, Jared G Galloway, Ariella L Gladstein, Graham Gower, Christopher C Kyriazis, Aaron P Ragsdale, Georgia Tsambos, Franz Baumdicker, et al. (2020). "A community-maintained standard library of population genetic models". *Elife* 9, e54967.

Albers, Patrick K. and Gil McVean (Jan. 2020). "Dating genomic variants and shared ancestry in population-scale sequencing data". *PLOS Biology* 18.1, pp. 1–26.

Amato, Roberto, Gennaro Miele, Antonella Monticelli, and Sergio Cocozza (2011). "Signs of selective pressure on genetic variants affecting human height". *PloS One* 6.11, e27588.

Arenas, Miguel (2013). "The importance and application of the ancestral recombination graph". *Frontiers in Genetics* 4, p. 206.

Baker, SJ, S Markowitz, ER Fearon, JK Willson, and B Vogelstein (1990a). "Suppression of human colorectal carcinoma cell growth by wild-type p53". *Science* 249.4971, pp. 912–915.

Baker, Suzanne J., Antonette C. Preisinger, J. Milburn Jessup, Christos Paraskeva, Sanford Markowitz, J. K. V. Willson, Stanley Hamilton, and Bert Vogelstein (1990b). "p53 Gene Mutations Occur in Combination with 17p Allelic Deletions as Late Events in Colorectal Tumorigenesis". *Cancer Research* 50.23, pp. 7717–7722.

Barton, Nicholas H, Jerome Kelleher, and Alison M Etheridge (2010). "A new model for extinction and recolonization in two dimensions: quantifying phylogeography". *Evolution: International journal of organic evolution* 64.9, pp. 2701–2715.

Battey, C. J., Peter L. Ralph, and Andrew D. Kern (2020). "Space is the Place: Effects of Continuous Spatial Structure on Analysis of Population Genetic Data". *Genetics* 215.1, pp. 193–214.

Belsare, Saurabh, Michal Levy-Sakin, Yulia Mostovoy, Steffen Durinck, Subhra Chaudhuri, Ming Xiao, Andrew S. Peterson, Pui-Yan Kwok, Somasekar Seshagiri, and Jeffrey D. Wall (2019). "Evaluating the quality of the 1000 genomes project data". *BMC Genomics* 20.1, p. 620.

Bergström, Anders, Shane A McCarthy, Ruoyun Hui, Mohamed A Almarri, Qasim Ayub, Petr Danecek, Yuan Chen, Sabine Felkel, Pille Hallast, Jack Kamm, et al. (2020). "Insights into human genetic variation and population history from 929 diverse genomes". *Science* 367.6484.

Bonné, Batsheva (1963). "The Samaritans: a demographic study". *Human biology* 35.1, pp. 61–89.

Bonomi, Luca, Yingxiang Huang, and Lucila Ohno-Machado (2020). "Privacy challenges and research opportunities for genomic data sharing". *Nature Genetics* 52.7, pp. 646–654.

Bordewich, Magnus and Charles Semple (2005). "On the computational complexity of the rooted subtree prune and regraft distance". *Annals of Combinatorics* 8.4, pp. 409–423.

Botigué, Laura R., Shiya Song, Amelie Scheu, Shyamalika Gopalan, Amanda L. Pendleton, Matthew Oetjens, Angela M. Taravella, Timo Seregély, Andrea Zeeb-Lanz, Rose-Marie Arbogast, Dean Bobo, Kevin Daly, Martina Unterländer, Joachim Burger, Jeffrey M. Kidd, and Krishna R. Veeramah (2017). "Ancient European dog genomes reveal continuity since the Early Neolithic". *Nature Communications* 8.1, p. 16082.

Brace, Selina, Yoan Diekmann, Thomas J. Booth, Lucy van Dorp, Zuzana Faltyskova, Nadin Rohland, Swapan Mallick, Iñigo Olalde, Matthew Ferry, Megan Michel, Jonas Oppenheimer, Nasreen Broomandkhoshbacht, Kristin Stewardson, Rui Martiniano, Susan Walsh, Manfred Kayser, Sophy Charlton, Garrett Hellenthal, Ian Armit, Rick Schulting, et al. (2019). "Ancient genomes indicate population replacement in Early Neolithic Britain". *Nature Ecology & Evolution* 3.5, pp. 765–771.

Browning, Brian L, Ying Zhou, and Sharon R Browning (2018). "A One-Penny Imputed Genome from Next-Generation Reference Panels". *The American Journal of Human Genetics*.

Browning, Sharon R. and Brian L. Browning (2007). "Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering". *The American Journal of Human Genetics* 81.5, pp. 1084 –1097.

Browning, Sharon R., Brian L. Browning, Ying Zhou, Serena Tucci, and Joshua M. Akey (2018). "Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture". *Cell* 173.1, 53–61.e9.

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. (2018). "The UK Biobank resource with deep phenotyping and genomic data". *Nature* 562, 203–209.

Capellini, Terence D, Hao Chen, Jiaxue Cao, Andrew C Doxey, Ata M Kiapour, Michael Schoor, and David M Kingsley (2017). "Ancient selection for derived alleles at a *GDF5* enhancer influencing human growth and osteoarthritis risk". *Nature Genetics* 49.8, p. 1202.

Cassidy, Lara M., Rui Martiniano, Eileen M. Murphy, Matthew D. Teasdale, James Mallory, Barrie Hartwell, and Daniel G. Bradley (2016). "Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome". *Proceedings of the National Academy of Sciences* 113.2, pp. 368–373.

Cavalli-Sforza, L Luca and Marcus W Feldman (2003). "The application of molecular genetic approaches to the study of human evolution". *Nature genetics* 33.3, pp. 266–275.

Chen, Lu, Aaron B. Wolf, Wenqing Fu, Liming Li, and Joshua M. Akey (2020). "Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals". *Cell* 180.4, 677–687.e16.

Chiang, Charleston W. K., Joseph H. Marcus, Carlo Sidore, Arjun Biddanda, Hussein Al-Asadi, Magdalena Zoledziewska, Maristella Pitzalis, Fabio Busonero, Andrea Maschio, Giorgio Pistis, Maristella Steri, Andrea Angius, Kirk E. Lohmueller, Goncalo R. Abecasis, David Schlessinger, Francesco Cucca, and John Novembre (2018). "Genomic history of the Sardinian population". *Nature Genetics* 50.10, pp. 1426–1434.

Choudhury, Ananyo, Shaun Aron, Laura R. Botigué, Dhriti Sengupta, Gerrit Botha, Taoufik Bensellak, Gordon Wells, Judit Kumuthini, Daniel Shriner, Yasmina J. Fakim, Anisah W. Ghoorah, Eileen Dareng, Trust Odia, Oluwadamilare Falola, Ezekiel Adebiyi, Scott Hazelhurst, Gaston Mazandu, Oscar A. Nyangiri, Mamana Mbiyavanga, Alia Benkahla, et al. (2020). "High-depth African genomes inform human migration and health". *Nature* 586.7831, pp. 741–748.

Consortium, 1000 Genomes Project (2015). "A global reference for human genetic variation". *Nature* 526.7571, p. 68.

Daans, Melina, Frank P Luyten, and Rik J U Lories (2011). "GDF5 deficiency in mice is associated with instability-driven joint damage, gait and subchondral bone changes". *Annals of the Rheumatic Diseases* 70.1, pp. 208–213.

Darwin, Charles (1837). *Transmutation Notebook D*. Ed. by P. H. Barrett, P. J. Gautrey, S. Herbert, D. Kohn, and S. Smith.

— (1857). *Letter no. 2143*.

Deng, Yun, Yun S. Song, and Rasmus Nielsen (2020). "The distribution of waiting distances in ancestral recombination graphs and its applications". *bioRxiv*.

Donnelly, Peter and Stephen Leslie (2010). "The coalescent and its descendants". *arXiv preprint arXiv:1006.1514*.

Donnelly, Peter and Simon Tavaré (1995). "Coalescents and genealogical structure under neutrality". *Annual Reviews Genetics* 29, pp. 401–421.

Douka, Katerina, Viviane Slon, Zenobia Jacobs, Christopher Bronk Ramsey, Michael V. Shunkov, Anatoly P. Derevianko, Fabrizio Mafessoni, Maxim B. Kozlikin, Bo Li, Rainer Grün, Daniel Comeskey, Thibaut Devièse, Samantha Brown, Bence Viola, Leslie Kinsley, Michael Buckley, Matthias Meyer, Richard G. Roberts, Svante Pääbo, Janet Kelso, et al. (2019). "Age estimates for hominin fossils and the onset of the Upper Palaeolithic at Denisova Cave". *Nature* 565.7741, pp. 640–644.

Durbin, Richard, Sean R Eddy, Anders Krogh, and Graeme Mitchison (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.

Eberle, Michael A, Epameinondas Fritzilas, Peter Krusche, Morten Källberg, Benjamin L Moore, Mitchell A Bekritsky, Zamin Iqbal, Han-Yu Chuang, Sean J Humphray, Aaron L Halpern, et al. (2017). "A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree". *Genome Research* 27.1, pp. 157–164.

Erlich, Yaniv, Tal Shor, Itsik Pe'er, and Shai Carmi (2018). "Identity inference of genomic data using long-range familial searches". *Science* 362.6415, pp. 690–694.

Evans, Christopher, Sam Lucy, and Ricky Patten (2018). *Riversides: Neolithic Barrows, a Beaker Grave, Iron Age and Anglo-Saxon Burials and Settlement at Trumpington, Cambridge*. McDonald Institute for Archaeological Research, University of Cambridge.

Felsenstein, Joseph (1975). "A pain in the torus: some difficulties with models of isolation by distance". *The american naturalist* 109.967, pp. 359–368.

— (2004). *Inferring phylogenies*. Vol. 2. Sinauer associates Sunderland, MA.

Fisher, Ronald Aylmer (1930). *The Genetical Theory of Natural Selection*. Clarendon.

Forney, G David (1973). "The viterbi algorithm". *Proceedings of the IEEE* 61.3, pp. 268–278.

François, Olivier and Flora Jay (2020). "Factor analysis of ancient population genomic samples". *Nature Communications* 11.1, p. 4661.

Galinsky, Kevin J, Po-Ru Loh, Swapan Mallick, Nick J Patterson, and Alkes L Price (2016). "Population Structure of UK Biobank and Ancient Eurasians Reveals Adaptation at Genes Influencing Blood Pressure." *Am J Hum Genet* 99.5, pp. 1130–1139.

Gerlinger, Marco, Andrew J. Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, Ignacio Varela, Benjamin Phillimore, Sharmin Begum, Neil Q. McDonald, Adam Butler, David Jones, Keiran Raine, Calli Latimer, Claudio R. Santos, Mahrokh Nohadani, et al. (2012). "Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing". *New England Journal of Medicine* 366.10. PMID: 22397650, pp. 883–892.

Gilbert, M. Thomas P., Hans-Jürgen Bandelt, Michael Hofreiter, and Ian Barnes (2005). "Assessing ancient DNA studies". *Trends in Ecology & Evolution* 20.10, pp. 541 –544.

Graur, Dan and Wen-Hsiung Li (2000). *Fundamentals of molecular evolution*. Vol. 2. Sinauer Associates.

Green, Richard E., Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Markus Hsi-Yang Fritz, Nancy F. Hansen, Eric Y. Durand, Anna-Sapfo Malaspinas, Jeffrey D. Jensen, Tomas Marques-Bonet, Can Alkan, Kay Prüfer, Matthias Meyer, Hernán A. Burbano, Jeffrey M. Good, et al. (2010). "A Draft Sequence of the Neandertal Genome". *Science* 328.5979, pp. 710–722.

Griffiths, R C and Simon Tavaré (Jan. 1998). "The age of a mutation in a general coalescent tree". *Communications in Statistics. Stochastic Models* 14.1-2, pp. 273–295.

Griffiths, Robert C (1991). "The two-locus ancestral graph". *Lecture Notes-Monograph Series*, pp. 100–117.

— (1997). "An ancestral recombination graph". *Progress in Population Genetics and Human Evolution*, pp. 257–270.

Griffiths, Robert C and Paul Marjoram (1996). "Ancestral inference from samples of DNA sequences with recombination". *Journal of Computational Biology* 3.4, pp. 479–502.

Grossman, Sharon R, Kristian G Andersen, Ilya Shlyakhter, Shervin Tabrizi, Sarah Winnicki, Angela Yen, Daniel J Park, Dustin Griesemer, Elinor K Karlsson, Sunny H Wong, Moran Cabili, Richard A Adegbola, Rameshwar N K Bamezai, Adrian V S Hill, Fredrik O Vannberg, John L Rinn, Eric S Lander, Stephen F Schaffner, and Pardis C Sabeti (2013). "Identifying recent adaptations in large-scale genomic data." eng. *Cell* 152.4, pp. 703–713.

Grüneberg, H and AJ Lee (1973). "The anatomy and development of brachypodism in the mouse". *Development* 30.1, pp. 119–141.

Gusfield, Dan (2005). "Optimal, efficient reconstruction of root-unknown phylogenetic networks with constrained and structured recombination". *Journal of Computer and System Sciences* 70.3, pp. 381–398.

— (2014). *ReCombinatorics: the algorithmics of ancestral recombination graphs and explicit phylogenetic networks*. MIT Press.

Gutaker, Rafal M and Hernán A Burbano (2017). "Reinforcing plant evolutionary genomics using ancient DNA". *Current Opinion in Plant Biology* 36. 36 Genome studies and molecular genetics, pp. 38 –45.

Gutenkunst, Ryan N., Ryan D. Hernandez, Scott H. Williamson, and Carlos D. Bustamante (2009). "Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data". *PLOS Genetics* 5.10, pp. 1–11.

Haak, Wolfgang, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, Susanne Nordenfelt, Eadaoin Harney, Kristin Stewardson, Qiaomei Fu, Alissa Mittnik, Eszter Bánffy, Christos Economou, Michael Francken, Susanne Friederich, Rafael Garrido Pena, Fredrik Hallgren, Valery Khartanovich, Aleksandr Khokhlov, et al. (2015). "Massive migration from the steppe was a source for Indo-European languages in Europe". *Nature* 522.7555, pp. 207–211.

Harada, M, M Takahara, P Zhe, M Otsuji, Y Iuchi, M Takagi, and T Ogino (2007). "Developmental failure of the intra-articular ligaments in mice with absence of growth differentiation factor 5". *Osteoarthritis and Cartilage* 15.4, pp. 468–474.

Harris, Charles R., K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, et al. (2020). "Array programming with NumPy". *Nature* 585, 357–362.

Harris, Kelley (2019). "From a database of genomes to a forest of evolutionary trees". *Nature Genetics* 51.9, pp. 1306–1307.

Hein, Jotun (1990). "Reconstructing evolution of sequences subject to recombination using parsimony". *Mathematical biosciences* 98.2, pp. 185–200.

Hein, Jotun, Mikkel Schierup, and Carsten Wiuf (2005). *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press, USA.

Hinchliff, Cody E., Stephen A. Smith, James F. Allman, J. Gordon Burleigh, Ruchi Chaudhary, Lyndon M. Coghill, Keith A. Crandall, Jiabin Deng, Bryan T. Drew, Romina Gazis, Karl Gude, David S. Hibbett, Laura A. Katz, H. Dail Laughinghouse, Emily Jane McTavish, Peter E. Midford, Christopher L. Owen, Richard H. Ree, Jonathan A. Rees, Douglas E. Soltis, et al. (2015). "Synthesis of phylogeny and taxonomy into a comprehensive tree of life". *Proceedings of the National Academy of Sciences* 112.41, pp. 12764–12769.

Hinrichs, Angela S, Donna Karolchik, Robert Baertsch, Galt P Barber, Gill Bejerano, Hiram Clawson, Mark Diekhans, Terrence S Furey, Rachel A Harte, Fan Hsu, et al. (2006). "The UCSC genome browser database: update 2006". *Nucleic acids research* 34, pp. D590–D598.

Hofreiter, Michael, David Serre, Hendrik N. Poinar, Melanie Kuch, and Svante Pääbo (2001). "Ancient DNA". *Nature Reviews Genetics* 2.5, pp. 353–359.

Hubisz, Melissa and Adam Siepel (2020). "Inference of ancestral recombination graphs using ARGweaver". In: *Statistical Population Genomics*. Humana, New York, NY, pp. 231–266.

Hubisz, Melissa J, Amy L Williams, and Adam Siepel (2020). "Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph". *PLoS genetics* 16.8, e1008895.

Hublin, Jean-Jacques, Abdelouahed Ben-Ncer, Shara E Bailey, Sarah E Freidline, Simon Neubauer, Matthew M Skinner, Inga Bergmann, Adeline Le Cabec, Stefano Benazzi, Katerina Harvati, and Philipp Gunz (2017). "New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens." *Nature* 546.7657, pp. 289–292.

Hudson, Richard R. (1983a). "Properties of a neutral allele model with intragenic recombination". *Theoretical Population Biology* 23.2, pp. 183 –201.

Hudson, Richard R (1983b). "Testing the constant-rate neutral allele model with protein sequence data". *Evolution*, pp. 203–217.

— (1990). "Gene genealogies and the coalescent process". In: *Oxford surveys in evolutionary biology*. Ed. by D. Futuyma and J. Antonovics. Vol. 7. Oxford University Press, pp. 1–44.

Hudson, Richard R and Norman L Kaplan (1985). "Statistical properties of the number of recombination events in the history of a sample of DNA sequences". *Genetics* 111.1, pp. 147–164.

Huerta-Sánchez, Emilia, Xin Jin, Asan, Zhuoma Bianba, Benjamin M. Peter, Nicolas Vinckenbosch, Yu Liang, Xin Yi, Mingze He, Mehmet Somel, Peixiang Ni, Bo Wang, Xiaohua Ou, Huasang, Jiangbai Luosang, Zha Xi Ping Cuo, Kui Li, Guoyi Gao, Ye Yin, Wei Wang, et al. (2014). "Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA". *Nature* 512.7513, pp. 194–197.

Hui, Ruoyun, Eugenia D'Atanasio, Lara M. Cassidy, Christiana L. Scheib, and Toomas Kivisild (2020). "Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes". *Scientific Reports* 10.1, p. 18542.

Hunt, Sarah E, William McLaren, Laurent Gil, Anja Thormann, Helen Schuilenburg, Dan Sheppard, Andrew Parton, Irina M Armean, Stephen J Trevanion, Paul Flicek, et al. (2018). "Ensembl variation resources". *Database* 2018.

Hwang, Sohyun, Eiru Kim, Insuk Lee, and Edward M. Marcotte (2015). "Systematic comparison of variant calling pipelines using gold standard personal exome variants". *Scientific Reports* 5.1, p. 17875.

Jacobs, Guy S, Georgi Hudjashov, Lauri Saag, Pradiptajati Kusuma, Chelzie C Darusallam, Daniel J Lawson, Mayukh Mondal, Luca Pagani, François-Xavier Ricaut, Mark Stoneking, et al. (2019). "Multiple deeply divergent Denisovan ancestries in Papuans". *Cell* 177.4, pp. 1010–1021.

Kalkauskas, Antanas, Umberto Perron, Yuxuan Sun, Nick Goldman, Guy Baele, Stephane Guindon, and Nicola De Maio (2021). "Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk". *PLOS Computational Biology* 17.1, pp. 1–27.

Kapli, Paschalia, Ziheng Yang, and Maximilian J. Telford (2020). "Phylogenetic tree building in the genomic age". *Nature Reviews Genetics* 21.7, pp. 428–444.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, Kristen M. Laricchia, Andrea Ganna, Daniel P. Birnbaum, Laura D. Gauthier, Harrison Brand, Matthew Solomonson, Nicholas A. Watts, Daniel Rhodes, Moriel Singer-Berk, Eleina M. England, Eleanor G. Seaby, Jack A. Kosmicki, Raymond K. Walters, et al. (2020). "The mutational constraint spectrum quantified from variation in 141,456 humans". *Nature* 581.7809, pp. 434–443.

Kelleher, Jerome, Alison M Etheridge, and Gilean McVean (2016). "Efficient coalescent simulation and genealogical analysis for large sample sizes". *PLoS Computational Biology* 12.5, e1004842.

Kelleher, Jerome, Kevin R Thornton, Jaime Ashander, and Peter L Ralph (2018). "Efficient pedigree recording for fast population genetics simulation". *PLoS computational biology* 14.11, e1006581.

Kelleher, Jerome, Yan Wong, Anthony W Wohns, Chaimaa Fadil, Patrick K Albers, and Gil McVean (2019). "Inferring whole-genome histories in large population datasets". *Nature genetics* 51.9, pp. 1330–1338.

Kendall, Michelle and Caroline Colijn (2016). "Mapping phylogenetic trees to reveal distinct patterns of evolution". *Molecular biology and evolution* 33.10, pp. 2735–2743.

Kent, David G and Anthony R Green (2017). "Order matters: the order of somatic mutations influences cancer evolution". *Cold Spring Harbor perspectives in medicine* 7.4, a027060.

Kimura, Motoo (1969). "The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations". *Genetics* 61.4, p. 893.

Kimura, Motoo and Tomoko Ohta (1973). "The age of a neutral mutant persisting in a finite population". *Genetics* 75.1, pp. 199–212.

Kingman, John FC (1982). "The coalescent". *Stochastic Processes and their Applications* 13.3, pp. 235–248.

Krings, Matthias, Helga Geisert, Ralf W. Schmitz, Heike Krainitzki, and Svante Pääbo (1999). "DNA sequence of the mitochondrial hypervariable region II from the Neandertal type specimen". *Proceedings of the National Academy of Sciences* 96.10, pp. 5581–5585.

Krings, Matthias, Anne Stone, Ralf W Schmitz, Heike Krainitzki, Mark Stoneking, and Svante Pääbo (1997). "Neandertal DNA sequences and the origin of modern humans". *cell* 90.1, pp. 19–30.

Lajugie, Julien, Rituparna Mukhopadhyay, Michael Schizas, Nathalie Lailler, Nicolas Fourel, and Eric E. Bouhassira (May 2013). "Complete Genome Phasing of Family Quartet by Combination of Genetic, Physical and Population-Based Phasing Analysis". *PLOS ONE* 8.5, pp. 1–14.

Laso-Jadart, Romuald, Christine Harmant, Hélène Quach, Nora Zidane, Chris Tyler-Smith, Qasim Mehdi, Qasim Ayub, Lluis Quintana-Murci, and Etienne Patin (2017). "The Genetic Legacy of the Indian Ocean Slave Trade: Recent Admixture and Post-admixture Selection in the Makranis of Pakistan". *The American Journal of Human Genetics* 101.6, pp. 977–984.

Lawson, Daniel John, Garrett Hellenthal, Simon Myers, and Daniel Falush (2012). "Inference of population structure using dense haplotype data". *PLoS Genet* 8.1, e1002453.

Lazaridis, Iosif, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H. Sudmant, Joshua G. Schraiber, Sergi Castellano, Mark Lipson, Bonnie Berger, Christos Economou, Ruth Bollongino, Qiaomei Fu, Kirsten I. Bos, Susanne Nordenfelt, Heng Li, Cesare de Filippo, Kay Prüfer, Susanna Sawyer, et al. (2014). "Ancient human genomes suggest three ancestral populations for present-day Europeans". *Nature* 513.7518, pp. 409–413.

Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, James S. Ware, Andrew J. Hill, Beryl B. Cummings, Taru Tukiainen, Daniel P. Birnbaum, Jack A. Kosmicki, Laramie E. Duncan, Karol Estrada, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, Joanne Berghout, David N. Cooper, et al. (2016). "Analysis of protein-coding genetic variation in 60,706 humans". *Nature* 536.7616, pp. 285–291.

Lemey, Philippe, Andrew Rambaut, Alexei J. Drummond, and Marc A. Suchard (2009). "Bayesian Phylogeography Finds Its Roots". *PLOS Computational Biology* 5.9, pp. 1–16.

Levine, Arnold J., Nancy A. Jenkins, and Neal G. Copeland (2019). "The Roles of Initiating Truncal Mutations in Human Cancers: The Order of Mutations and Tumor Cell Type Matters". *Cancer Cell* 35.1, pp. 10–15.

Lewin, Harris A., Gene E. Robinson, W. John Kress, William J. Baker, Jonathan Coddington, Keith A. Crandall, Richard Durbin, Scott V. Edwards, Félix Forest, M. Thomas P. Gilbert, Melissa M. Goldstein, Igor V. Grigoriev, Kevin J. Hackett, David Haussler, Erich D. Jarvis, Warren E. Johnson, Aristides Patrinos, Stephen Richards, Juan Carlos Castilla-Rubio, Marie-Anne van Sluys, et al. (2018). "Earth BioGenome Project: Sequencing life for the future of life". *Proceedings of the National Academy of Sciences* 115.17, pp. 4325–4333.

Li, Heng and Richard Durbin (2011). "Inference of human population history from individual whole-genome sequences". *Nature* 475.7357, pp. 493–496.

Li, Na and Matthew Stephens (2003). "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data". *Genetics* 165.4, pp. 2213–2233.

Link, Vivian, Athanasios Kousathanas, Krishna Veeramah, Christian Sell, Amelie Scheu, and Daniel Wegmann (2017). "ATLAS: Analysis Tools for Low-depth and Ancient Samples". *bioRxiv*.

Liu, Jie, Wei Cai, Hongxin Zhang, Chuan He, and Lianfu Deng (2013). "Rs143383 in the growth differentiation factor 5 (GDF5) gene significantly associated with osteoarthritis (OA)-a comprehensive meta-analysis." *Int J Med Sci* 10.3, pp. 312–319.

Llamas, Bastien, Lars Fehren-Schmitz, Guido Valverde, Julien Soubrier, Swapan Mallick, Nadin Rohland, Susanne Nordenfelt, Cristina Valdiosera, Stephen M. Richards, Adam Rohrlach, Maria Inés Barreto Romero, Isabel Flores Espinoza, Elsa Tomasto Cagigao, Lucía Watson Jiménez, Krzysztof Makowski, Ilán Santiago Leboreiro Reyna, Josefina Mansilla Lory, Julio Alejandro Ballivián Torrez, Mario A. Rivera, Richard L. Burger, et al. (2016). "Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas". *Science Advances* 2.4.

Loog, Liisa, Mark G Thomas, Ross Barnett, Richard Allen, Naomi Sykes, Ptolemaios D Paxinos, Ophélie Lebrasseur, Keith Dobney, Joris Peters, Andrea Manica, et al. (2017). "Inferring allele frequency trajectories from ancient DNA indicates that selection on a chicken gene coincided with changes in medieval husbandry practices". *Molecular biology and evolution* 34.8, pp. 1981–1990.

Loughlin, John (2015). "Genetic contribution to osteoarthritis development: current state of evidence". *Current Opinion in Rheumatology* 27.3, p. 284.

Lowy-Gallego, Ernesto, Susan Fairley, Xiangqun Zheng-Bradley, Magali Ruffier, Laura Clarke, Paul Flicek, 1000 Genomes Project Consortium, et al. (2019). "Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project". *Wellcome Open Research* 4.

Lunter, Gerton (Aug. 2018). "Haplotype matching in large cohorts using the Li and Stephens model". *Bioinformatics* 35.5, pp. 798–806.

Mafessoni, Fabrizio, Steffi Grote, Cesare de Filippo, Viviane Slon, Kseniya A. Kolobova, Bence Viola, Sergey V. Markin, Manjusha Chintalapati, Stephane Peyrégne, Laurits Skov, Pontus Skoglund, Andrey I. Krivoshapkin, Anatoly P. Derevianko, Matthias Meyer, Janet Kelso, Benjamin Peter, Kay Prüfer, and Svante Pääbo (2020). "A high-coverage Neandertal genome from Chagyrskaya Cave". *Proceedings of the National Academy of Sciences* 117.26, pp. 15132–15136.

MalariaGEN (2021). *The Anopheles gambiae 1000 Genomes Consortium (2021): Ag1000G phase 3 SNP data release.*

Malaspinas, Anna-Sapfo, Orestis Malaspinas, Steven N Evans, and Montgomery Slatkin (2012). "Estimating allele age and selection coefficient from time serial data". *Genetics*, genetics–112.

Malécot, Gustave (1948). *Mathématiques de l'hérédité*. Paris: Masson et Cie.

Mallick, Swapan, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, Niru Chennagiri, Susanne Nordenfelt, Arti Tandon, et al. (2016). "The Simons genome diversity project: 300 genomes from 142 diverse populations". *Nature* 538.7624, p. 201.

Mani, G. S. and Bryan Campbell Clarke (1990). "Mutational order: a major stochastic process in evolution". *Proceedings of the Royal Society of London. B. Biological Sciences* 240.1297, pp. 29–37.

Marcus, Joseph H., Cosimo Posth, Harald Ringbauer, Luca Lai, Robin Skeates, Carlo Sidore, Jessica Beckett, Anja Furtwängler, Anna Olivieri, Charleston W. K. Chiang, Hussein Al-Asadi, Kushal Dey, Tyler A. Joseph, Chi-Chun Liu, Clio Der Sarkissian, Rita Radzevičiūtė, Megan Michel, Maria Giuseppina Gradoli, Patrizia Marongiu, Salvatore Rubino, et al. (2020). "Genetic history from the Middle Neolithic to present on the Mediterranean island of Sardinia". *Nature Communications* 11.1, p. 939.

Margaryan, Ashot, Daniel J. Lawson, Martin Sikora, Fernando Racimo, Simon Rasmussen, Ida Moltke, Lara M. Cassidy, Emil Jørsboe, Andrés Ingason, Mikkel W. Pedersen, Thorfinn Korneliussen, Helene Wilhelmson, Magdalena M. Buś, Peter de Barros Damgaard, Rui Martiniano, Gabriel Renaud, Claude Bhérer, J. Víctor Moreno-Mayar, Anna K. Fotakis, Marie Allen, et al. (2020). "Population genomics of the Viking world". *Nature* 585.7825, pp. 390–396.

Marjoram, Paul and Jeff D Wall (2006). "Fast "coalescent" simulation". *BMC Genetics* 7.1, p. 16.

Martiniano, Rui, Erik Garrison, Eppie R. Jones, Andrea Manica, and Richard Durbin (2020). "Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph". *Genome Biology* 21.1, p. 250.

Mathieson, Iain (2020). "Estimating time-varying selection coefficients from time series data of allele frequencies". *bioRxiv*.

Mathieson, Iain and Gil McVean (Aug. 2014). "Demography and the Age of Rare Variants". *PLOS Genetics* 10.8, pp. 1–11.

McDougall, Ian, Francis H Brown, and John G Fleagle (2005). "Stratigraphic placement and age of modern humans from Kibish, Ethiopia." *Nature* 433.7027, pp. 733–736.

McVean, Gil (2007). "The Structure of Linkage Disequilibrium Around a Selective Sweep". *Genetics* 175.3, pp. 1395–1406.

— (Oct. 2009). "A Genealogical Interpretation of Principal Components Analysis". *PLOS Genetics* 5.10, pp. 1–10.

McVean, Gilean and Niall J Cardin (2005a). "Approximating the coalescent with recombination". *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 360.1459, pp. 1387–1393.

McVean, Gilean A. T. (2002). "A Genealogical Interpretation of Linkage Disequilibrium". *Genetics* 162.2, pp. 987–991.

McVean, Gilean A.T and Niall J Cardin (2005b). "Approximating the coalescent with recombination". *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1459, pp. 1387–1393.

Melé, Marta, Asif Javed, Marc Pybus, Francesc Calafell, Laxmi Parida, Jaume Bertranpetit, Genographic Consortium, et al. (2010). "A new method to reconstruct recombination events at a genomic scale". *PLoS Computational Biology* 6.11, e1001010.

Menozzi, P, A Piazza, and L Cavalli-Sforza (1978). "Synthetic maps of human gene frequencies in Europeans". *Science* 201.4358, pp. 786–792.

Meyer, Matthias, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G. Schraiber, Flora Jay, Kay Prüfer, Cesare de Filippo, Peter H. Sudmant, Can Alkan, Qiaomei Fu, Ron Do, Nadin Rohland, Arti Tandon, Michael Siebauer, Richard E. Green, Katarzyna Bryc, Adrian W. Briggs, et al. (2012). "A High-Coverage Genome Sequence from an Archaic Denisovan Individual". *Science* 338.6104, pp. 222–226.

Michaelson, Jacob J, Yujian Shi, Madhusudan Gujral, Hancheng Zheng, Dheeraj Malhotra, Xin Jin, Minghan Jian, Guangming Liu, Douglas Greer, Abhishek Bhandari, et al. (2012). "Whole-genome sequencing in autism identifies hot spots for de novo germline mutation". *Cell* 151.7, pp. 1431–1442.

Mikic, B, RT Clark, TC Battaglia, V Gaschen, and Ernst Bruno Hunziker (2004). "Altered hypertrophic chondrocyte kinetics in GDF-5 deficient murine tibial growth plates". *Journal of Orthopaedic Research* 22.3, pp. 552–556.

Minichiello, Mark J and Richard Durbin (2006). "Mapping trait loci by use of inferred ancestral recombination graphs". *The American Journal of Human Genetics* 79.5, pp. 910–922.

Mirzaei, Sajad and Yufeng Wu (2017). "RENT+: an improved method for inferring local genealogical trees from haplotypes with recombination". *Bioinformatics* 33.7, pp. 1021–1030.

Miyamoto, Yoshinari, Akihiko Mabuchi, Dongquan Shi, Toshikazu Kubo, Yoshio Takatori, Susumu Saito, Mikihiro Fujioka, Akihiro Sudo, Atsumasa Uchida, Seizo Yamamoto, et al. (2007). "A functional polymorphism in the 5 UTR of *GDF5* is associated with susceptibility to osteoarthritis". *Nature Genetics* 39.4, p. 529.

Moreno-Mayar, J. Víctor, Lasse Vinner, Peter de Barros Damgaard, Constanza de la Fuente, Jeffrey Chan, Jeffrey P. Spence, Morten E. Allentoft, Tharsika Vimala, Fernando Racimo, Thomaz Pinotti, Simon Rasmussen, Ashot Margaryan, Miren Iraeta Orbegozo, Dorothea Mylopotamitaki, Matthew Wooller, Clement Bataille, Lorena Becerra-Valdivia, David Chivall, Daniel Comeskey, Thibaut Devièse, et al. (2018). "Early human dispersals within the Americas". *Science* 362.6419.

Morin, PA, JJ Moore, R Chakraborty, L Jin, J Goodall, and DS Woodruff (1994). "Kin selection, social structure, gene flow, and the evolution of chimpanzees". *Science* 265.5176, pp. 1193–1201.

Morrison, David A (2016). "Genealogies: Pedigrees and phylogenies are reticulating networks not just divergent trees". *Evolutionary Biology* 43.4, pp. 456–473.

Murphy, Kevin, Yair Weiss, and Michael I Jordan (2013). "Loopy belief propagation for approximate inference: An empirical study". In: pp. 467–475.

Nachman, Michael W and Susan L Crowell (Sept. 2000). "Estimate of the Mutation Rate per Nucleotide in Humans". *Genetics* 156.1, pp. 297–304.

Nakagome, Shigeki, Gorka Alkorta-Aranburu, Roberto Amato, Bryan Howie, Benjamin M Peter, Richard R Hudson, and Anna Di Rienzo (Feb. 2016). "Estimating the Ages of Selection Signals from Different Epochs in Human History". *Molecular Biology and Evolution* 33.3, pp. 657–669.

Narasimhan, Vagheesh M., Nick Patterson, Priya Moorjani, Nadin Rohland, Rebecca Bernardos, Swapan Mallick, Iosif Lazaridis, Nathan Nakatsuka, Iñigo Olalde, Mark Lipson, Alexander M. Kim, Luca M. Olivieri, Alfredo Coppa, Massimo Vidale, James Mallory, Vyacheslav Moiseyev, Egor Kitov, Janet Monge, Nicole Adamski, Neel Alex, et al. (2019). "The formation of human populations in South and Central Asia". *Science* 365.6457.

Nesta, Alex V., Denisse Tafur, and Christine R. Beck (2020). "Hotspots of Human Mutation". *Trends in Genetics*.

Nigro, Janice M., Suzanne J. Baker, Antonette C. Preisinger, J. Milburn Jessup, Richard Hosteller, Karen Cleary, Sandra H. Signer, Nancy Davidson, Stephen Baylin, Peter Devilee, Thomas Glover, Francis S. Collins, Ainsley Weslon, Rama Modali, Curtis C. Harris, and Bert Vogelstein (1989). "Mutations in the p53 gene occur in diverse human tumour types". *Nature* 342.6250, pp. 705–708.

Nosil, Patrik and Samuel M. Flaxman (2011). "Conditions for mutation-order speciation". *Proceedings of the Royal Society B: Biological Sciences* 278.1704, pp. 399–407.

O'Connell, James F and Jim Allen (2015). "The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago". *Journal of Archaeological Science* 56, pp. 73–84.

O'Fallon, Brendan D (2013). "ACG: rapid inference of population history from recombining nucleotide sequences". *BMC Bioinformatics* 14.1, p. 40.

Olalde, Iñigo, Selina Brace, Morten E. Allentoft, Ian Armit, Kristian Kristiansen, Thomas Booth, Nadin Rohland, Swapan Mallick, Anna Szécsényi-Nagy, Alissa Mittnik, Eveline Altena, Mark Lipson, Iosif Lazaridis, Thomas K. Harper, Nick Patterson, Nasreen Broomandkhoshbacht, Yoan Diekmann, Zuzana Faltyskova, Daniel Fernandes, Matthew Ferry, et al. (2018). "The Beaker phenomenon and the genomic transformation of northwest Europe". *Nature* 555.7695, pp. 190–196.

Ormond, Louise, Matthieu Foll, Gregory B Ewing, Susanne P Pfeifer, and Jeffrey D Jensen (Dec. 2015). "Inferring the age of a fixed beneficial allele". *Molecular Ecology* 25.1, pp. 157–169.

Pääbo, Svante, Hendrik Poinar, David Serre, Viviane Jaenicke-Després, Juliane Hebler, Nadin Rohland, Melanie Kuch, Johannes Krause, Linda Vigilant, and Michael Hofreiter (2004). "Genetic analyses from ancient DNA". *Annu Rev Genet* 38.1, pp. 645–679.

Palamara, Pier Francesco, Jonathan Terhorst, Yun S Song, and Alkes L Price (2018). "High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability." *Nat Genet* 50.9, pp. 1311–1317.

Parida, Laxmi, Marta Melé, Francesc Calafell, Jaume Bertranpetit, and Genographic Consortium (2008). "Estimating the ancestral recombinations graph (ARG) as compatible networks of SNP patterns". *Journal of Computational Biology* 15.9, pp. 1133–1153.

Patterson, Nick, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich (2012). "Ancient admixture in human history". *Genetics* 192.3, pp. 1065–1093.

Patterson, Nick, Alkes L Price, and David Reich (Dec. 2006). "Population Structure and Eigenanalysis". *PLOS Genetics* 2.12, pp. 1–20.

Picard toolkit (2019). `http://broadinstitute.github.io/picard/`.

Pickrell, Joseph and Jonathan Pritchard (2012). "Inference of population splits and mixtures from genome-wide allele frequency data". *Nature Precedings*, pp. 1–1.

Pickrell, Joseph K, Graham Coop, John Novembre, Sridhar Kudaravalli, Jun Z Li, Devin Absher, Balaji S Srinivasan, Gregory S Barsh, Richard M Myers, Marcus W Feldman, and Jonathan K Pritchard (2009). "Signals of recent positive selection in a worldwide sample of human populations." eng. *Genome Res* 19.5, pp. 826–837.

Piličiauskas, Gytis, Rimantas Jankauskas, Giedrė Piličiauskienė, Oliver E. Craig, Sophy Charlton, and Tosha Dupras (2017). "The transition from foraging to farming (7000–500calBC) in the SE Baltic: A re-evaluation of chronological and palaeodietary evidence from human remains". *Journal of Archaeological Science: Reports* 14, pp. 530–542.

Plicht, J van der, C Bronk Ramsey, T J Heaton, E M Scott, and S Talamo (2020). "Recent Developments in Calibration for Archaeological and Environmental Samples". *Radiocarbon* 62.4, 1095–1117.

Pritchard, Jonathan K., Matthew Stephens, and Peter Donnelly (2000). "Inference of Population Structure Using Multilocus Genotype Data". *Genetics* 155.2, pp. 945–959.

Prüfer, Kay, Cesare de Filippo, Steffi Grote, Fabrizio Mafessoni, Petra Korlević, Mateja Hajdinjak, Benjamin Vernot, Laurits Skov, Pinghsun Hsieh, Stéphane Peyrégne, David Reher, Charlotte Hopfe, Sarah Nagel, Tomislav Maricic, Qiaomei Fu, Christoph Theunert, Rebekah Rogers, Pontus Skoglund, Manjusha Chintalapati, Michael Dannemann, et al. (2017). "A high-coverage Neandertal genome from Vindija Cave in Croatia". *Science* 358.6363, pp. 655–658.

Prüfer, Kay, Kasper Munch, Ines Hellmann, Keiko Akagi, Jason R. Miller, Brian Walenz, Sergey Koren, Granger Sutton, Chinnappa Kodira, Roger Winer, James R. Knight, James C. Mullikin, Stephen J. Meader, Chris P. Ponting, Gerton Lunter, Saneyuki Higashino, Asger Hobolth, Julien Dutheil, Emre Karakoç, Can Alkan, et al. (2012). "The bonobo genome compared with the chimpanzee and human genomes". *Nature* 486.7404, pp. 527–531.

Prüfer, Kay, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, Gabriel Renaud, Peter H. Sudmant, Cesare de Filippo, Heng Li, Swapan Mallick, Michael Dannemann, Qiaomei Fu, Martin Kircher, Martin Kuhlwilm, Michael Lachmann, Matthias Meyer, Matthias Ongyerth, Michael Siebauer, et al. (2014). "The complete genome sequence of a Neanderthal from the Altai Mountains". *Nature* 505.7481, pp. 43–49.

Quach, Hélène, Maxime Rotival, Julien Pothlichet, Yong-Hwee Eddie Loh, Michael Dannemann, Nora Zidane, Guillaume Laval, Etienne Patin, Christine Harmant, Marie Lopez, et al. (2016). "Genetic adaptation and Neandertal admixture shaped the immune system of human populations". *Cell* 167.3, pp. 643–656.

Ralph, Peter, Kevin Thornton, and Jerome Kelleher (2020). "Efficiently Summarizing Relationships in Large Samples: A General Duality Between Statistics of Genealogies and Genomes". *Genetics* 215.3, pp. 779–797.

Rasmussen, Matthew D, Melissa J Hubisz, Ilan Gronau, and Adam Siepel (2014). "Genome-wide inference of ancestral recombination graphs". *PLoS Genetics* 10.5, e1004342.

Reich, David (2018). *Who We Are and How We Got Here: Ancient DNA and the New Science of the Human Past*. Oxford University Press.

Reich, David, Richard E. Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y. Durand, Bence Viola, Adrian W. Briggs, Udo Stenzel, Philip L. F. Johnson, Tomislav Maricic, Jeffrey M. Good, Tomas Marques-Bonet, Can Alkan, Qiaomei Fu, Swapan Mallick, Heng Li, Matthias Meyer, Evan E. Eichler, Mark Stoneking, et al. (2010). "Genetic history of an archaic hominin group from Denisova Cave in Siberia". *Nature* 468.7327, pp. 1053–1060.

Reich, David, Nick Patterson, Martin Kircher, Frederick Delfin, Madhusudan R Nandineni, Irina Pugach, Albert Min-Shan Ko, Ying-Chin Ko, Timothy A Jinam, Maude E Phipps, et al. (2011). "Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania". *The American Journal of Human Genetics* 89.4, pp. 516–528.

Reich, David, Kumarasamy Thangaraj, Nick Patterson, Alkes L. Price, and Lalji Singh (2009). "Reconstructing Indian population history". *Nature* 461.7263, pp. 489–494.

Renaux, Claude, Laura Buzdugan, Markus Kalisch, and Peter Bühlmann (2020). "Hierarchical inference for genome-wide association studies: a view on methodology with software". *Computational Statistics* 35.1, pp. 1–40.

Risch, Neil, Deborah de Leon, Laurie Ozelius, Patricia Kramer, Laura Almasy, Burton Singer, Stanley Fahn, Xandra Breakefield, and Susan Bressman (1995). "Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population". *Nature Genetics* 9.2, p. 152.

Roach, Jared C., Gustavo Glusman, Arian F. A. Smit, Chad D. Huff, Robert Hubley, Paul T. Shannon, Lee Rowen, Krishna P. Pant, Nathan Goodman, Michael Bamshad, Jay Shendure, Radoje Drmanac, Lynn B. Jorde, Leroy Hood, and David J. Galas (2010). "Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing". *Science* 328.5978, pp. 636–639.

Rouault, K., V. Scotet, S. Autret, F. Gaucher, F. Dubrana, D. Tanguy, C. Yaacoub El Rassi, B. Fenoll, and C. Férec (2010). "Evidence of association between GDF5 polymorphisms and congenital dislocation of the hip in a Caucasian population". *Osteoarthritis and Cartilage* 18.9, pp. 1144–1149.

Sabeti, Pardis C, Patrick Varilly, Ben Fry, Jason Lohmueller, Elizabeth Hostetter, Chris Cotsapas, Xiaohui Xie, Elizabeth H Byrne, Steven A McCarroll, Rachelle Gaudet, et al. (2007). "Genome-wide detection and characterization of positive selection in human populations". *Nature* 449.7164, p. 913.

Sánchez-Quinto, Federico, Helena Malmström, Magdalena Fraser, Linus Girdland-Flink, Emma M. Svensson, Luciana G. Simões, Robert George, Nina Hollfelder, Göran Burenhult, Gordon Noble, Kate Britton, Sahra Talamo, Neil Curtis,

Hana Brzobohata, Radka Sumberova, Anders Götherström, Jan Storå, and Mattias Jakobsson (2019). "Megalithic tombs in western and northern Neolithic Europe were linked to a kindred society". *Proceedings of the National Academy of Sciences* 116.19, pp. 9469–9474.

Sankararaman, Sriram, Swapan Mallick, Michael Dannemann, Kay Prüfer, Janet Kelso, Svante Pääbo, Nick Patterson, and David Reich (2014). "The genomic landscape of Neanderthal ancestry in present-day humans". *Nature* 507.7492, pp. 354–357.

Sanna, Serena, Anne U Jackson, Ramaiah Nagaraja, Cristen J Willer, Wei-Min Chen, Lori L Bonnycastle, Haiqing Shen, Nicholas Timpson, Guillaume Lettre, Gianluca Usala, et al. (2008). "Common variants in the *GDF5-UQCC* region are associated with variation in human height". *Nature Genetics* 40.2, p. 198.

Scheib, Christiana L, Ruoyun Hui, Eugenia D'Atanasio, Anthony Wilder Wohns, Sarah A Inskip, Alice Rose, Craig Cessford, Tamsin C O'Connell, John E Robb, Christopher Evans, et al. (2019). "East Anglian early Neolithic monument burial linked to contemporary Megaliths". *Annals of human biology* 46.2, pp. 145–149.

Schluter, Dolph (2009). "Evidence for Ecological Speciation and Its Alternative". *Science* 323.5915, pp. 737–741.

Serre, JL, B Simon-Bouy, E Mornet, B Jaume-Roig, A Balassopoulou, M Schwartz, A Taillandier, J Boue, and A Boue (1990). "Studies of RFLP closely linked to the cystic fibrosis locus throughout Europe lead to new considerations in populations genetics". *Human Genetics* 84.5, pp. 449–454.

Sesia, Matteo, Eugene Katsevich, Stephen Bates, Emmanuel Candès, and Chiara Sabatti (2020). "Multi-resolution localization of causal variants across the genome". *Nature Communications* 11.1, p. 1093.

Shi, Lingling, Yunfei Guo, Chengliang Dong, John Huddleston, Hui Yang, Xiaolu Han, Aisi Fu, Quan Li, Na Li, Siyi Gong, Katherine E. Lintner, Qiong Ding, Zou Wang, Jiang Hu, Depeng Wang, Feng Wang, Lin Wang, Gholson J. Lyon, Yongtao Guan, Yufeng Shen, et al. (2016). "Long-read sequencing and de novo assembly of a Chinese genome". *Nature Communications* 7.1, p. 12065.

Skov, Laurits, Moisès Coll Macià, Garðar Sveinbjörnsson, Fabrizio Mafessoni, Elise A. Lucotte, Margret S. Einarsdóttir, Hakon Jonsson, Bjarni Halldorsson, Daniel F. Gudbjartsson, Agnar Helgason, Mikkel Heide Schierup, and Kari Stefansson (2020). "The nature of Neanderthal introgression revealed by 27,566 Icelandic genomes". *Nature* 582.7810, pp. 78–83.

Slatkin, M and B Rannala (2000). "Estimating allele age." *Annu Rev Genomics Hum Genet* 1, pp. 225–249.

Slatkin, Montgomery (2000). "Allele age and a test for selection on rare alleles". *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 355.1403, pp. 1663–1668.

Slatkin, Montgomery and Bruce Rannala (1997). "Estimating the age of alleles by use of intraallelic variability." *American Journal of Human Genetics* 60.2, p. 447.

Smith, Joel, Graham Coop, Matthew Stephens, and John Novembre (Jan. 2018). "Estimating Time to the Common Ancestor for a Beneficial Allele". *Molecular Biology and Evolution* 35.4, pp. 1003–1017.

Song, Yun S and Jotun Hein (2005). "Constructing minimal ancestral recombination graphs". *Journal of Computational Biology* 12.2, pp. 147–169.

Song, Yun S., Yufeng Wu, and Dan Gusfield (June 2005). "Efficient computation of close lower and upper bounds on the minimum number of recombinations in biological sequence evolution". *Bioinformatics* 21.suppl_1, pp. i413–i422.

Speidel, Leo, Lara Cassidy, Robert W. Davies, Garrett Hellenthal, Pontus Skoglund, and Simon R. Myers (2021). "Inferring population histories for ancient genomes using genome-wide genealogies". *bioRxiv*.

Speidel, Leo, Marie Forest, Sinan Shi, and Simon R. Myers (2019). "A method for genome-wide genealogy estimation for thousands of samples". *Nature Genetics* 51.9, pp. 1321–1329.

Stern, Aaron J., Leo Speidel, Noah A. Zaitlen, and Rasmus Nielsen (2021). "Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies". *The American Journal of Human Genetics* 108.2, pp. 219–239.

Storm, Elaine E, Thanh V Huynh, Neal G Copeland, Nancy A Jenkins, David M Kingsley, and Se-Jin Lee (1994). "Limb alterations in brachypodism mice due to mutations in a new member of the TGFβ-superfamily". *Nature* 368.6472, p. 639.

Szöllősi, Gergely J, Eric Tannier, Vincent Daubin, and Bastien Boussau (2015). "The inference of gene trees with species trees." *Syst Biol* 64.1, e42–62.

Tajima, F (1983). "Evolutionary relationship of DNA sequences in finite populations." *Genetics* 105.2, pp. 437–460.

— (1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." *Genetics* 123.3, pp. 585–595.

Takahata, N (Jan. 1993). "Allelic genealogy and human evolution." *Molecular Biology and Evolution* 10.1, pp. 2–22.

Takahata, Naoyuki and Masatoshi Nei (1985). "Gene genealogy and variance of interpopulational nucleotide differences". *Genetics* 110.2, pp. 325–344.

Taliun, Daniel, Daniel N. Harris, Michael D. Kessler, Jedidiah Carlson, Zachary A. Szpiech, Raul Torres, Sarah A. Gagliano Taliun, André Corvelo, Stephanie M. Gogarten, Hyun Min Kang, Achilleas N. Pitsillides, Jonathon LeFaive, Seungbeen Lee, Xiaowen Tian, Brian L. Browning, Sayantan Das, Anne-Katrin Emde, Wayne E. Clarke, Douglas P. Loesch, Amol C. Shetty, et al. (2019). "Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program". *bioRxiv*.

Tavaré, Simon (1984). "Line-of-descent and genealogical processes, and their applications in population genetics models". *Theoretical Population Biology* 26.2, pp. 119–164.

— (2004). "Part I: Ancestral inference in population genetics". In: *Lectures on Probability Theory and Statistics*. Springer, pp. 1–188.

Tavaré, Simon, David J. Balding, R. C. Griffiths, and Peter Donnelly (1997). "Inferring Coalescence Times From DNA Sequence Data". *Genetics* 145.2, pp. 505–518.

Underhill, P. A., G. Passarino, A. A. Lin, P. Shen, M. Mirazón Lahr, R. A. Foley, P. J. Oefner, and L. L. Cavalli-Sforza (2001). "The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations". *Annals of Human Genetics* 65.1, pp. 43–62.

Vernot, Benjamin and Joshua M. Akey (2014). "Resurrecting Surviving Neandertal Lineages from Modern Human Genomes". *Science* 343.6174, pp. 1017–1021.

Vernot, Benjamin, Serena Tucci, Janet Kelso, Joshua G. Schraiber, Aaron B. Wolf, Rachel M. Gittelman, Michael Dannemann, Steffi Grote, Rajiv C. McCoy, Heather Norton, Laura B. Scheinfeldt, David A. Merriwether, George Koki, Jonathan S. Friedlaender, Jon Wakefield, Svante Pääbo, and Joshua M. Akey (2016). "Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals". *Science* 352.6282, pp. 235–239.

Vigilant, L, M Stoneking, H Harpending, K Hawkes, and AC Wilson (1991). "African populations and the evolution of human mitochondrial DNA". *Science* 253.5027, pp. 1503–1507.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". *Nature Methods* 17, pp. 261–272.

Voight, Benjamin F, Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K Pritchard (2006). "A map of recent positive selection in the human genome". *PLoS Biology* 4.3, e72.

Wakeley, John (2009). *Coalescent theory: an introduction*. Macmillan Learning.

Wall, Jeffrey D, Melinda A Yang, Flora Jay, Sung K Kim, Eric Y Durand, Laurie S Stevison, Christopher Gignoux, August Woerner, Michael F Hammer, and Montgomery Slatkin (2013). "Higher levels of Neanderthal ancestry in East Asians than in Europeans". *Genetics* 194.1, pp. 199–209.

Wang, Gao, Abhishek Sarkar, Peter Carbonetto, and Matthew Stephens (2020). "A simple new approach to variable selection in regression, with application to genetic fine mapping". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.5, pp. 1273–1300.

Wang, Lusheng, Kaizhong Zhang, and Louxin Zhang (2001). "Perfect phylogenetic networks with recombination". *Journal of Computational Biology* 8.1, pp. 69–78.

Watterson, GA (1976). "Reversibility and the age of an allele. I. Moran's infinitely many neutral alleles model". *Theoretical Population Biology* 10.3, pp. 239–253.

Wenger, Aaron M., Paul Peluso, William J. Rowell, Pi-Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D. Olson, Armin Töpfer, Michael Alonge, Medhat Mahmoud, Yufeng Qian, Chen-Shan Chin, Adam M. Phillippy, Michael C. Schatz, Gene Myers, Mark A. DePristo, Jue Ruan, et al. (2019). "Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome". *Nature Biotechnology* 37.10, pp. 1155–1162.

Wiuf, Carsten and Peter Donnelly (1999). "Conditional genealogies and the age of a neutral mutant". *Theoretical Population Biology* 56.2, pp. 183–201.

Wiuf, Carsten and Jotun Hein (1999). "Recombination as a point process along sequences". *Theoretical Population Biology* 55.3, pp. 248–259.

Wohns, Anthony Wilder, Yan Wong, Ben Jeffery, Ali Akbari, Swapan Mallick, Ron Pinhasi, Nick Patterson, David Reich, Jerome Kelleher, and Gil McVean (2021). "A unified genealogy of modern and ancient genomes". *bioRxiv*.

Wright, Sewall (1931). "Evolution in Mendelian populations". *Genetics* 16.2, pp. 97–159.

— (1943). "Isolation by distance". *Genetics* 28.2, pp. 114–138.

Wu, Dong-Dong, Gui-Mei Li, Wei Jin, Yan Li, and Ya-Ping Zhang (2012). "Positive selection on the osteoarthritis-risk and decreased-height associated variants at the *GDF5* gene in East Asians". *PLoS One* 7.8, e42553.

Wu, Yufeng (2008). "Association mapping of complex diseases with ancestral recombination graphs: models and efficient algorithms". *Journal of Computational Biology* 15.7, pp. 667–684.

Zeberg, Hugo and Svante Pääbo (2020). "The major genetic risk factor for severe COVID-19 is inherited from Neanderthals". *Nature* 587.7835, pp. 610–612.