UNIVERSITY OF COPENHAGEN

MSc in Bioinformatics

Modeling the prehistory of early modern humans using markers of Neanderthal introgression

Filippo Tell (ghk844)

Supervisors: Martin Petr, Fernando Racimo

August 2025

Acknowledgements

I would like to thank Fernando for giving me the opportunity of being part of his group, and all the members of the group for creating such a nice and friendly environment.

Of course, I would like to say a million thanks to Martin, both for this project and the previous ones we have worked on together. Working with you has been particularly cool, both from a personal and "professional" level. I am really happy with what I have learned.

Abstract

It is now widely accepted that the modern human lineage originated in Africa and subsequently dispersed across the entire world. During this migration, modern humans encountered Neanderthals and Denisovans, two archaic human lineages that inhabited the Eurasian continent for hundreds of thousands of years, and interbred with them. As a result of this introgression around 50 thousand years ago, a small percentage of the nuclear genomes of present-day non-African populations is formed by Neanderthal introgressed tracts.

The distribution of introgressed tracts in ancient and present-day individuals has been shaped by various selective and demographic processes over tens of thousand of years. In particular, because modern humans are known to have experienced a significant bottleneck during and after the Out-of-Africa migration, genetic drift is expected to have affected the distribution of introgressed tracts in modern humans in a significant way. Interestingly, although the bottleneck during the early prehistory of modern humans has been extensively studied in population genetics, the exact nature of the population dynamics, tribal structure, or societal organization of prehistoric hunter-gatherer populations remains relatively poorly understood. For instance, it appears quite unlikely, from a geographical perspective, that a small group of modern humans could occupy large geographical ranges in Eurasia for tens of thousands of years, as implied by the single-population, panmictic models of modern human demographic history that have been successfully used in population genetics for decades. Given that the spatio-temporal and genomic distribution of introgressed tracts in modern humans must have been shaped by complex prehistoric demographic processes, could the information stored in them be used to model the migratory patterns and social organization of prehistoric people?

In this study, we introduced a new introgression-based summary statistics called the tract frequency spectrum (TFS), which describes the distribution of introgressed tracts between individuals. Using extensive simulations across a series of alternative models of

a prehistoric population structure, we found that the TFS does have a potential to be a useful summary statistic informative about the demographic history of modern humans. We also developed a proof-of-concept algorithm to infer a spatiotemporal graph of relationship between individuals based on the extent of their sharing of introgressed tracts, which we propose will provide a useful data structure for future work, similarly to the TFS. Additionally, we developed a simulation-based implementation of a previously proposed metric of the correlation between shared drift between individuals and the extent of their tract sharing, and show that it could be used as an informative statistic for modeling the process of Neanderthal introgression.

Overall, this study provides a new modeling and statistical framework for a follow up work which will attempt to move from traditionally used idealized single-population scenarios of modern human prehistory to more detailed models which include ancient population structure and dynamics.

Contents

Introduction

The story of human evolution features frequent episodes of gene flow and migration. Paleoanthropology has found numerous examples of ancient groups of humans migrating out of Africa, where the entire genus *Homo* originated, at different time points in our history, with the earliest documented dispersal as old as about 2 million years ago (Mya) [1, 2].

At a later point, about 600 thousand years ago (kya), a group of humans which diverged from what eventually became the ancestors of anatomically modern humans (AMH) left Africa and eventually populated a considerable part of Eurasia. Outside of Africa, this human lineage later split and differentiated into at least two known lineages of so-called archaic humans, Neanderthals and Denisovans [3, 4]. Neanderthals, named after the "Neander's Valley" in Germany where their remains were first discovered (*Neanderthal* in German), inhabited a wide range of Eurasia [5], while Denisovans, named after the Denisova cave in Siberia, ranged from Siberia to Southeast Asia [6].

Meanwhile, the modern human lineage, the ancestor of all present-day people, is thought to have emerged within Africa around 300 kya [7], although the nature of this process appears to be complex [8]. Around 60 kya, a population of anatomically modern humans migrated out of Africa and, during the process of expanding through Eurasia, met and interbred with the two aforementioned archaic human groups, Neanderthals and Denisovans. Finally, around 40 kya, paleontological evidence suggests that anatomically modern humans remained the only human evolutionary lineage on the planet, as archaic humans disappeared from the archaeological record [9]. Subsequently, modern humans continued to colonize the world to the geographical extent they occupy today.

Of course, the above-mentioned examples of ancient migrations and gene flow still paint a significantly simplified picture of human evolution, which must have been even more intricate and dynamic, characterized by more migratory waves, population structure, and admixture events between related groups which we currently known

very little about. For example, genomic data suggests that Denisovans and Neanderthals have themselves interbred with each other [10]. Additionally, further genomic evidence suggests gene flow in the opposite direction, from the ancestors of AMH into archaic humans [11, 12].

Although many aspects of human evolutionary history have been historically characterized by investigating fossil evidence or analyses of present-day human genomes, many of the breakthroughs have come through the sequencing of human ancient DNA (aDNA), that is the DNA isolated from human remains [13].

The role of ancient DNA in unveiling human history

Ancient DNA (aDNA) is DNA isolated and sequenced from ancient biological samples, such as bones and tissues. It is often characterized by short, degraded, and chemically modified fragments [14]. The origins of the field can be traced to the 1980s, when fragments of mitochondrial DNA from an extinct quagga museum specimen were successfully sequenced and analysed [15]. Soon after, the invention of the polymerase chain reaction (PCR) [16] significantly advanced the field, allowing the sequencing of longer stretches of aDNA, reaching approximately 27 thousand base pairs for certain samples [17].

However, aDNA "revolution" on a massive scale began in the early 2000s with the advent of next-generation sequencing (NGS)[18]. The application of NGS to aDNA made it soon possible to sequence million of base pairs per sample, significantly surpassing the capabilities of the standard PCR [19]. However, numerous challenges had to be overcome in order to obtain DNA sequences of sufficient quality. In particular, the high contamination of aDNA samples by exogenous DNA sources (e.g. microbes and humans) and the post-mortem degradation of the DNA [20, 21] necessitated technological advancements in both molecular and computational biology to minimize these major problems in aDNA analysis. As a result, it is now possible to obtain almost full-length DNA sequences from organisms that lived tens of thousands (and, recently, even millions [22]) of years ago. In certain cases, the depth of coverage is comparable to that of present-day samples [23, 24, 25].

As a result of these technical developments, aDNA has become a powerful tool in population genetics, helping to uncover the evolutionary histories of extant and extinct species in an increasing level of detail. Its ability to reveal ancient variations over

time and space clarifies the relationships between groups and their migratory patterns, which shaped present-day populations, at a scale which would not be reachable with present-day DNA alone.

In the context of the study of human history, aDNA has been crucial in revealing critical events about our past and origin. Not only it has uncovered evidence of more recent dramatic, continental-scale migrations [26, 27, 28], but it has also opened a window into our distant past by shedding light on the genomes of two of our extinct archaic relatives: Neanderthals and their sister group Denisovans [3, 4, 29, 30].

Sequencing the nuclear genomes of Neanderthals and Denisovans has allowed researchers to discover that these two groups of archaic humans interbred with the ancestors of non-Africans which resulted in the gene flow of genetic material from Neanderthals and Denisovans into [3, 4, 29, 31, 32]. Today, approximately 2% of the nuclear genomes of people of primarily non-African ancestry is derived from Neanderthals [3], while Southeast Asian populations can trace an even larger component of their genomes (about 5% for certain groups) to Denisovans [33].

After finding evidence of gene flow from archaic into modern humans, scientists began asking more detailed questions. For instance, when and where did modern and archaic humans meet and interbreed? What were the dynamics of this population contact? How often did they meet? What impact, if any, has the archaic introgressed DNA had on the biology of modern humans?

Although significant progress has been made over the last 15 years in answering these questions, additional research is needed to develop a better understanding of archaic human introgression. The increasing number of archaic and ancient modern human DNA samples, as well as improvements in both the experimental and computational methods, will hopefully help us to obtain a clearer picture of the encounters between archaic and modern humans, both from a historical and biological perspective.

Neanderthal introgression

The first draft of the Neanderthal nuclear genome was obtained in 2010 through the sequencing of three specimens unearthed in the Vindija Cave in Croatia [3]. Comparative analyses of this composite genome to genomes of present-day humans revealed that all non-Africans today were approximately 4% closer to Neanderthals than present-day Africans. The study concluded that the most plausible explanation for this observation is that gene flow occurred from Neanderthals into the ancestors of non-African populations [34]. Subsequent studies, which included more Neanderthal and ancient

modern human genomes, further supported this hypothesis [29, 31, 35, 36] and also allowed researchers to estimate the time of this introgression [34]. Additionally, the identification of long Neanderthal introgressed tracts in the genomes of two early modern humans (40 kya) from Romania and western Siberia suggested that gene flow from Neanderthal to their ancestors occurred only several generations prior in these individuals' genealogy, suggesting that introgression likely happened multiple times [31, 35, 36].

Today, the scientific community widely accepts the introgression hypothesis. However, beyond the broad strokes, relatively little is known about the extent to which early modern human populations interacted with archaic humans, the extent to which they contributed to later populations, and even less about the geographical distribution of introgression or social dynamics. With more and more data available, it is perhaps time to ask more ambitious questions about these more detailed aspects of this admixture event, such as its geographical location, precise timing, and social dynamics (e.g., what were the group sizes of archaic and modern humans involved in the population contact, and how frequently did this contact happen), as well as the effects of the introgressed material on the biology of early and present-day modern humans.

The Middle East is presumed to be one of the most likely locations where Neanderthals and modern humans met and interbred. The archaeological record suggests that these two groups likely inhabited the area contemporaneously [34, 37]. However, Neanderthals occupied an extensive habitat spanning the majority of the Eurasian continent [5] and there is evidence of recent admixture (a few generations prior) [31, 35, 36] in certain European early modern humans individuals older than 40 kya. Although these lineages did not seem to contribute significantly to present-day Eurasian populations, it is difficult to imagine that the Neanderthal DNA carried by present-day populations is originated from a single pulse admixture event between an ancestral Eurasian population and Neanderthals in the Middle East, as the necessarily simplified population genetic models generally assume[38].

Initially, the timing of the Neanderthal admixture with modern humans was estimated somewhere between 50 and 65 kya, based on the aforementioned old (>40 kya) early modern human samples [29, 31, 36]. However, a recent study [39] based on new early modern human (45 kya) genomes from individuals in Germany and the Czech Republic, which represent the deepest known split from the Out-of-Africa lineage that contributed to the genomes of all present-day Eurasians, constrained the timing of the Neanderthal admixture event to between 45 and 49 kya. Their analysis also

supports the scenario of a single pulse of admixture based on the length distribution of the Neanderthal tract. However, as noted in the study, the authors could not exclude other potential scenarios, such as extended or multiple pulses of admixture, due to the limitations of the method applied [40]. A similar conclusion was reached by another study [41], where once again the hypothesis of multiple pulses of introgression could not be eliminated.

Significant progress has been made in understanding the influence of this admixture on human traits, as well as the evolutionary forces that shaped the genomic distribution of Neanderthal tracts in present-day genomes. The pattern of introgressed alleles in modern and early modern human genomes suggests that these alleles experienced different fates when they entered our genome and were subject to strong selective pressure. Some genomic regions appear to have been significantly depleted soon after entering the modern human genomic background (also called Neanderthal deserts) [39, 42, 43, 44]. Others have been identified as potential candidates of adaptive introgression [45, 46, 47, 48].

Given that Neanderthals inhabited Europe for about 400 thousand years [9, 49], it is plausible that some of their alleles were well adapted to the local environment and might have been beneficial to early modern human populations that arrived in those regions. Candidate regions for adaptive introgression are regions involved in brain development, neuronal function, adaptive and innate immunity, lipid metabolism, skin and hair pigmentation and the musculoskeletal system [50].

The growing interest in archaic admixture has driven the development of computational and probabilistic methods to detect archaic introgressed segments in modern human genomes [51, 52, 53]. Advances in aDNA analysis, including methods for obtaining complete high-fidelty genomes from fossil remains and imputing low-coverage genomes [54], have made it possible to obtain large panels of ancient and present-day modern human genomes to study how the Neanderthal ancestry has changed over time and space.

However, to fully understand how admixture occurred and what contributed shaping the Neanderthal ancestry in present-day genomes, it is essential to develop models which integrate the demography and dynamics of the early modern humans who left Africa. It is plausible that different demographic scenarios of the Out-of-Africa event and subsequent population and social dynamics of hunter-gatherer populations would have resulted in different patterns of Neanderthal ancestry in present-day populations

and over time, different amounts of genetic drift and, consequently, influencing the selection forces acting on introgressed DNA.

Models of early modern humans Out-of-Africa dispersal

A single origin of modern humans somewhere in East Africa followed by a subsequent Out-of-Africa dispersal about 60 kya is supported by both archaeological and genetic evidence [55, 56, 57]. Several models for an Out-of-Africa dispersal have been proposed [58, 59, 60], however, the details of modern human expansion history following this dispersal are still largely unresolved [61].

Most working population genetic models of human demography used in the literature are largely based on the assumption that the population migrating Out-of-Africa experienced a severe bottleneck, that significantly reduced its population size [58]. Similar conclusions have been obtained in other studies using different statistical methods [59, 62]. This reduction is described in terms of effective population size (N_e) and it has been estimated of being less than 2000 mating individuals. N_e is a core concept in population genetics, serving as a measure of the rate of genetic drift in an idealized, panmictic population underlying most theoretical and practical work in population genetics. However, in the context of studying prehistory, particularly in a geographical setting, it is not straightforward to visualize it and apply it to a real population involving a given census number of individuals (almost always unknown) [63, 64]. For instance, the idea that a single, homogeneous human population of a few thousand people left Africa and remained of the same size with the constant panmictic population dynamics is, of course, quite unrealistic.

The population genetic inference methods used in the aforementioned studies to detect this severe bottleneck are built on the assumption of panmixia. While this assumption formed the basis for the first population genetics models [65, 66, 67], and it continues to be extremely useful both in theory and practice, it is rarely observed in natural populations, which generally feature a significant degree of population structure. Ignoring the structure and applying methods that assume panmixia can lead to misleading if not meaningless results about the population history [68, 69, 70]. For example, a structured population that did not experience changes in population size might be wrongly interpreted as experiencing bottleneck by statistical methods that assume a panmictic model [68].

While it is very likely that some form of population bottleneck indeed did occur, simply by the fact that a *subset* of the African genetic variation separated during the Out-of-Africa migration, it seems intuitively unlikely that a single-population bottleneck describes the full picture. Alternative and more complicated scenarios, with multiple dispersion Out-of-Africa as proposed by other studies, might be an example of models whose aspects might apply as well [60, 71].

Archaeological and genetic evidence suggests that certain archaic and ancient groups were socially organized in small and interconnected groups [24, 72, 73]. A similar structure of multiple, interconnected, relatively small groups may have also been common among modern humans who left Africa. If this is the case, it would imply that the Out-of-Africa modern humans were not a single, homogenous and panmictic population and that other alternative models featuring a degree of spatial population structure might better reflect the reality.

In this study, we applied a simulation-based approach to develop a set of alternative, proof-of-concept scenarios of the Out-of-Africa event and subsequent demography of prehistory hunter-gatherer populations in Eurasia. We investigated the plausibility of these scenarios against the traditional panmictic population genetic models using standard population genetics summary statistics as well as several novel statistics and data structures based on the spatio-temporal and genomic distribution of sharing of Neanderthal tracts. Our motivation to focus on statistics based on introgressed Neanderthal tracts was motivated by our hypothesis that they might potentially carry a unique advantages and power to disentangle the complexities of social and population dynamics of prehistoric societies. Finally, we present an initial comparison of our results with statistics obtained from empirical ancient and present-day data, leveraging a large panel of recently published imputed human genomes [26].

Project aims

This thesis project represents the first step to move from traditionally used population genetic models of early modern human (EMH) prehistory toward more detailed models of population structure, social dynamics, and census size. Our goal was to develop a set of alternative models of EMH prehistory which would capture the real historical demographic processes in a more realistic way, and evaluate their feasibility using established population genetic statistics, as well as newly developed metrics based on the spatio-temporal and genomic distribution of the introgressed tracts. Specifically, we were interested in examining how would different demographic scenarios and modes of hybridization and introgression (e.g. single pulse vs. multiple pulses) reflect in these summary statistics.

Eventually, our goal is to apply these summary statistics in simulation-based model inference such as Approximate Bayesian Computation (ABC). Because the reliability of simulation-based inferences critically depends on the informativeness of summary statistics used [74, 75, 76], we decided to approach the problem first by simulating different demographic scenarios for an Out-of-Africa migration of modern humans, each with its own distinguishing features, followed by Neanderthal introgression. We then evaluated how are different features of these scenarios reflected in these summary statistics, with a particular focus on investigating which statistics pick up important signals about the models and which do not. Finally, we applied these summary statistics to an empirical dataset of Neanderthal tracts inferred in present-day modern human genomes to investigate, to what degree are our summary statistics of choice consistent with the assumed models of modern human demographic history.

For organizational clarity, the goals of the overall project have been divided into three chapters.

Chapter 1: Demographic scenarios for an Out-of-Africa dispersal of modern humans

The first aim of this thesis was to define a set of demographic scenarios, with a particular focus on the population dynamics and structure of the earliest modern humans that have inhabited Europe after the Out-of-Africa migration. Our goal was to compare the performance of these models with a simplified, single-population, panmictic demographic model, as a representative of the baseline model that underlies many of the results in population genetic literature [38]. Because our next aim was to develop novel population genetic summary statistics based on introgressed tracts, we needed to consider the introgression event itself. Although it is generally assumed that a vast majority of introgressed tracts entered the modern human genetic background through a single "pulse" of introgression, there is a considerable uncertainty into the exact nature of introgression dynamics or, in fact, into whether or not multiple introgression events occurred as modern humans expanded across Eurasia. To investigate the potential of our summary statistics to shed light on this question, each of our scenarios was implemented in two versions: one with a single extended pulse of Neanderthal introgression in the ancestral Out-of-Africa (OOA) population, and one with two pulses: the first pulse into the OOA ancestors, and the second pulse private to a respective subsequent Eurasian lineage. We then evaluated the qualitative feasibility of each model by comparing standard population genetics summary statistics, such as heterozygosity, against present-day population data. This comparison will allow us to establish whether the proposed models are reasonable in light of observed data, particularly in terms of the amount of genetic drift manifested in each model.

Chapter 2: Developing tract-based summary statistics

The second aim of this thesis was to develop novel metrics of introgressed tracts, including a graph data structure which captures the relationship between individuals based on the sharing of introgressed tracts across space and time. Our primary motivation for this was the hypothesis that the exact nature of spatio-temporal population and social dynamics within and between groups of EMH during (and after) the Out-of-Africa migration would be the primary factor determining the amount of genetic drift acting on introgressed tracts. As a result, fitting parameters of population models against

metrics obtained from introgressed tracts in EMH individuals across space and time could serve as an important source of information for fitting population dynamics of the earliest EMH groups in Europe. To this end, we evaluated these metrics on demographic scenarios proposed in section , and study how are aspects of these scenarios reflected in tract-based summary statistics and the tract graph data structure.

Chapter 3: Exploratory applications to empirical data

The third aim of this thesis was to apply the summary statistics explored in section to empirical data of modern human genomes, using coordinates of introgressed Neanderthal tracts previously inferred in our group (Refoyo-Martínez, *et al.*, in prep.) using the IBDmix software[52]. Although determining a set of models which most closely fit to the real, unknown, historical and demographic process will require the development of a dedicated model inference procedure, such as Approximate Bayesian Computation, we nevertheless aimed to assess whether any of the results obtained for the proposed scenarios of an Out-of-Africa migration and Neanderthal introgression are reasonable given the observed data, and which of them should be clearly disqualified or modified accordingly.

Chapter 1: Demographic scenarios of an early modern human prehistory

1

Since our primary goal was to develop a set of alternative models of early modern human (EMH) prehistory which would capture the real historical demographic processes in a more realistic way, and evaluate their feasibility using established and new population genetic statistics, we dedicated the first chapter of this study to the implementation and validation of different demographic model scenarios.

We designed three demographic models of an Out-of-Africa dispersal of modern humans and subsequent admixture with Neanderthals, primarily focusing on alternative demographic histories of a West Eurasian population in terms of their ancient social dynamics and population structures. In addition, we implemented two variants of Neanderthal introgression involved in each of these models: a single-pulse event and a two-pulses event.

To get a sense of how well these models align with empirical data, we compared the distribution of individual heterozygosity between simulated present-day individuals and empirical data.

For simplicity and clarity, we will refer to these three different models as **Model 1**, **Model 2**, and **Model 3** (see section 1.1.1).

1.1 Materials and Methods

1.1.1 Demographic models and simulations

We designed three primary scenarios of Out-of-Africa migration of modern humans followed by Neanderthal introgression, focusing exclusively on the demography of the West Eurasian population. All models were implemented using the R package *slendr* (version 1.1.0) [77] and the final outcome of each model was a tree-sequence object simulated with *slendr*'s built-in *msprime* back-end script operated by the function *msprime()* [78], with mutations laid over the genealogies using the function *ts_mutate()*. In all simulations, we sampled 50 West Eurasian and 5 African present-day individuals. We modeled each individual with a 100 Megabases (Mb) diploid genome, a mutation rate of 1.0×10^{-8} per base pair per generation, and a uniform recombination rate of 1.0×10^{-8} per base pair per generation.

Each model shares a common baseline topological structure: an ancestral population splits into an African population ("AFR") and a Neanderthal population ("NEA") at 550 kya [79]. The Out-of-Africa population ("OOA") then branches from the "AFR" population at 70 kya [80] and at 42 kya the West Eurasian population ("EUR") emerges from the "OOA" population. Building upon this common topology, each individual scenario then differs in the structure and history of the "EUR" population:

- **Model 1**: This scenario presents an unstructured and constant size (N_e=15000) "EUR" population, and is intended to capture a typical idealized model of human demographic history, such as the one exemplified by the Gravel *et al.* model [58].

- **Model 2**: The second scenario presents a structured "EUR" population characterized by 9 demes of equal size (fixed N_e=2000). At 30 kya, these demes merge three by three into three ancestral populations: Western Hunter Gatherer ("WHG"), Anatolian farmers ("ANA"), and the Yamnaya steppe herders ("YAM"). At 8 kya, we merged the "ANA" and the "WHG" populations (Neolithic farmer migration), followed by a merge with the "YAM" population at 5 kya (steppe migration), resulting in a single present-day West Eurasian population [81].

- **Model 3**: The third scenario is similar to **Model 2**, presenting 9 equal size "EUR" demes. However, unlike in **Model 2**, these demes are not completely independent. At 30 kya, we introduced gene flow between demes with consecutive deme numbers, to resemble a spatial scenario in which neighboring demes exchange

migrants. This gene flow continues until 3 kya, when all the demes merged together into a single "EUR" population.

Each model presents only two present-day populations, "AFR" and "EUR", from which we sampled the individuals.

To introduce Neanderthal introgression, we implemented two variants of each of the three models above, named **A** and **B**:

- **Variant A**: These scenarios feature a single extended gene flow from the "NEA" population into the "OOA" at a 3% rate from 49 kya to 45 kya [39].
- **Variant B**: These scenarios are characterized by a first extended gene flow from the "NEA" population into the "OOA" at a 2% rate from 49 kya to 45 kya, and by a second episode at a 1% rate between 42 kya and 40 kya from "NEA" into each independent population that descended from the "OOA" population.

In the context of developing summary statistics informative about important aspects of early human prehistory, contrasting variants **A** and **B** were intending to investigate the power of different metrics summarizing the sharing (or, alternatively, uniqueness) Neanderthal tracts between individuals, and their potential to contribute to resolving the question of the additional pulses of introgression in the evolutionary history of ancient and present-day Eurasians.

We kept all the model parameters fixed, except for the N_e of the "OOA" population that has been parametrized. Section S4.1 summarize the models parameters.

1.1.2 Heterozygosity

For each model described in section 1.1.1, we computed the heterozygosity from a given simulated tree sequence for each sampled individual using the *slendr* function *ts_diversity()*. To reduce the effect of stochasticity, we simulated each model across 10 replicates, and then averaged the heterozygosity values for all sampled individuals across all replicates to estimate the density distribution. The distributions were estimated and plotted using the default parameters of the *geom_density()* function in the *ggplot2* R package [82]. For each model, 500 individuals were used to estimate the heterozygosity distribution for the EUR population, and 50 individuals were used for the AFR population.

For comparison with empirical data, we used autosomal heterozygosity estimates for West Eurasians and African samples obtained from the Simons Genome Diversity Project [83] (Supplementary Table 1).

1.2 Results

1.2.1 Visualization of model scenarios

To verify that the demographic history models implemented in *slendr* corresponded to our specification (see section 1.1.1), we visualized a representation of each model through the *plot_model* function in the *slendr* R package [77]. However, we realized that as soon as a given model reaches certain degree of complexity (particularly a higher number of population and gene-flow events), the default plotting behavior of the *plot_model* function produced cluttered representations, resulting in figures which made it challenging to verify the models as intended (see Figure S7). We solved this problem by modifying the *plot_model* source code, utilizing the *ggbreak* R package [84] to introduce an optional break in the y-axis of the plotted model through its function *scale_y_break*. This makes it possible for a user of the *slendr* package to narrow down model visualization to highlight the most relevant phases of demographic history. This fix has been proposed in an issue on the GitHub repository (https://github.com/bodkan/slendr) of *slendr* for the consideration of the developers.

Figure 1.1 shows a visual representation of the demographic history of **Model 1** as produced by the *slendr* function *plot_model*. **Model 1** is the simplest demographic scenario among the three, and is intended to capture the simplified nature to traditionally used population genetic models of human demography, as exemplified by the model proposed by Gravel *et al.* [58]. Most importantly, this model involves a single panmictic European population whose ancestors experienced a strong bottleneck during the Out-of-Africa ("OOA") event, and which later evolves into a West Eurasian ("EUR") population.

Figure 1.2 shows a visual representation of the demographic history captured by **Model 2**. **Model 2** introduces a certain degree of structure in the demographic history of the "EUR" population, making it slightly more complex compared to the idealized panmictic population of **Model 1**. In particular, in this scenario the West Eurasian population is structured into 9 independent demes that eventually merge to form the three main ancestral populations of present-day Europeans, inspired by the study by Lazaridis *et al.* [81]: Western Hunter-Gatherers ("WHG"), Anatolian farmers ("ANA"), and the Yamnaya steppe herders ("YAM"). Subsequently, these three populations also merge into a single panmictic "EUR" population, but do so at a later stage at 3 kya, intended to capture, in a simplified manner, the large-scale migrations in the past ten thousand years [81].

Finally, Figure 1.3 presents the demographic scenario we call **Model 3**. Similarly to **Model 2**, this scenario exhibits a structured history of the "EUR" population composed by nine demes. However, this scenario features a continuous gene flow events between the "neighboring" pairs of demes before they, again, merge into a single panmictic "EUR" population, similarly to **Model 2**. In other words, this model is again intended to capture a potential geographic and population substructure in the early-modern human population, but in a different, still idealized, manner.

A detailed description and specifics of the implementation of the models, model parameters, and their variants can be found in section 1.1.1.

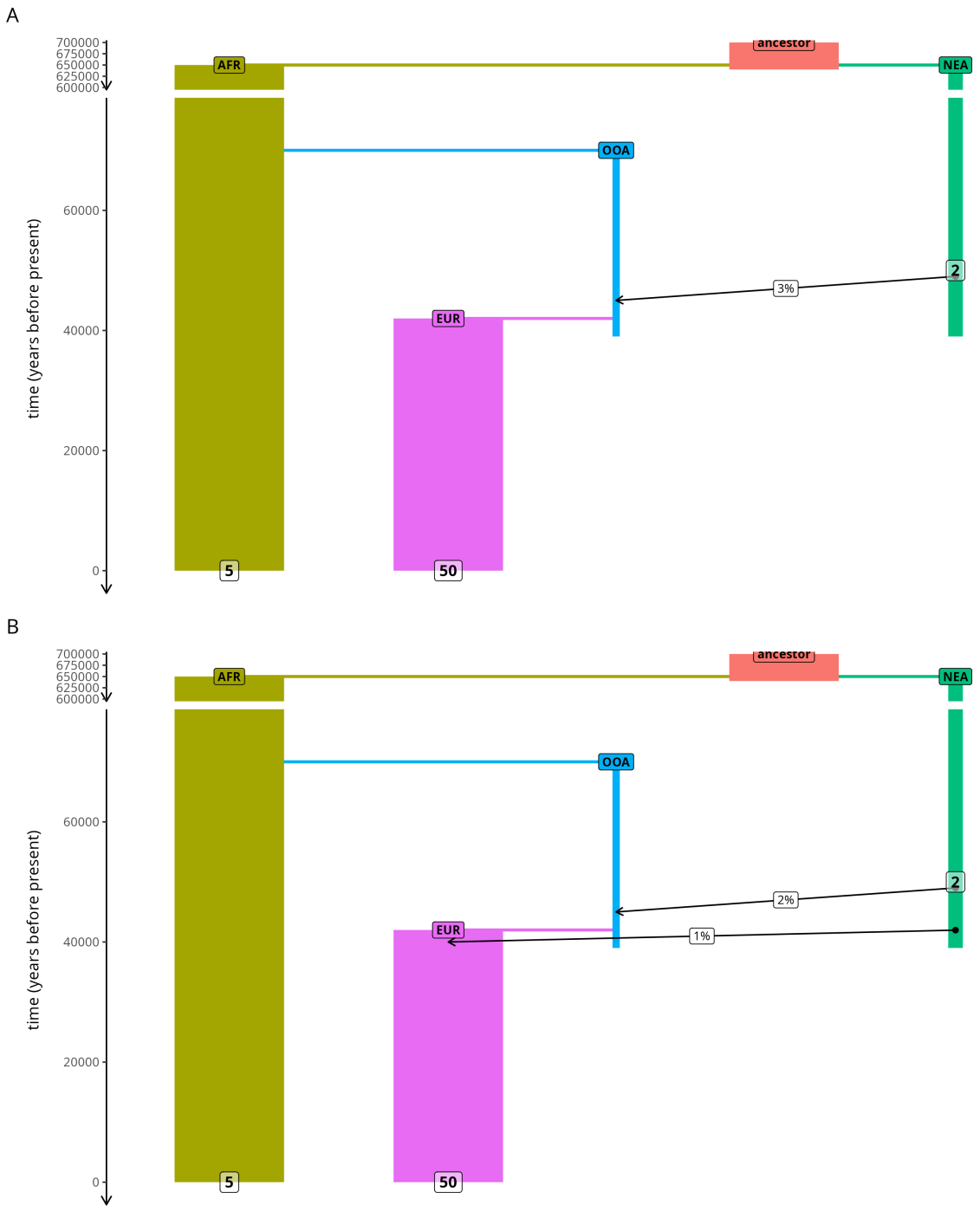

Figure 1.1: Schematic representation of the demographic history of Model 1. (**A**) shows the variant **A** of Neanderthal ("NEA") introgression for this model, with one single extended pulse of introgression into the Out-of-Africa ("OOA") population. (**B**) shows the variant **B** of "NEA" introgression for this model, with two independent extended pulses of introgression: one into the "OOA" and one into the "EUR". The arrows indicate the gene-flow events, and the percentages correspond to the proportion of migration over the entire time-period of each respective gene-flow event. The numbers within boxes correspond to the number of individuals sampled from that population at a given time point. The y-axis represents the time in years before present, with gaps representing our modification to the default plotting procedure of the *slendr* package to facilitate clearer plotting and model verification.

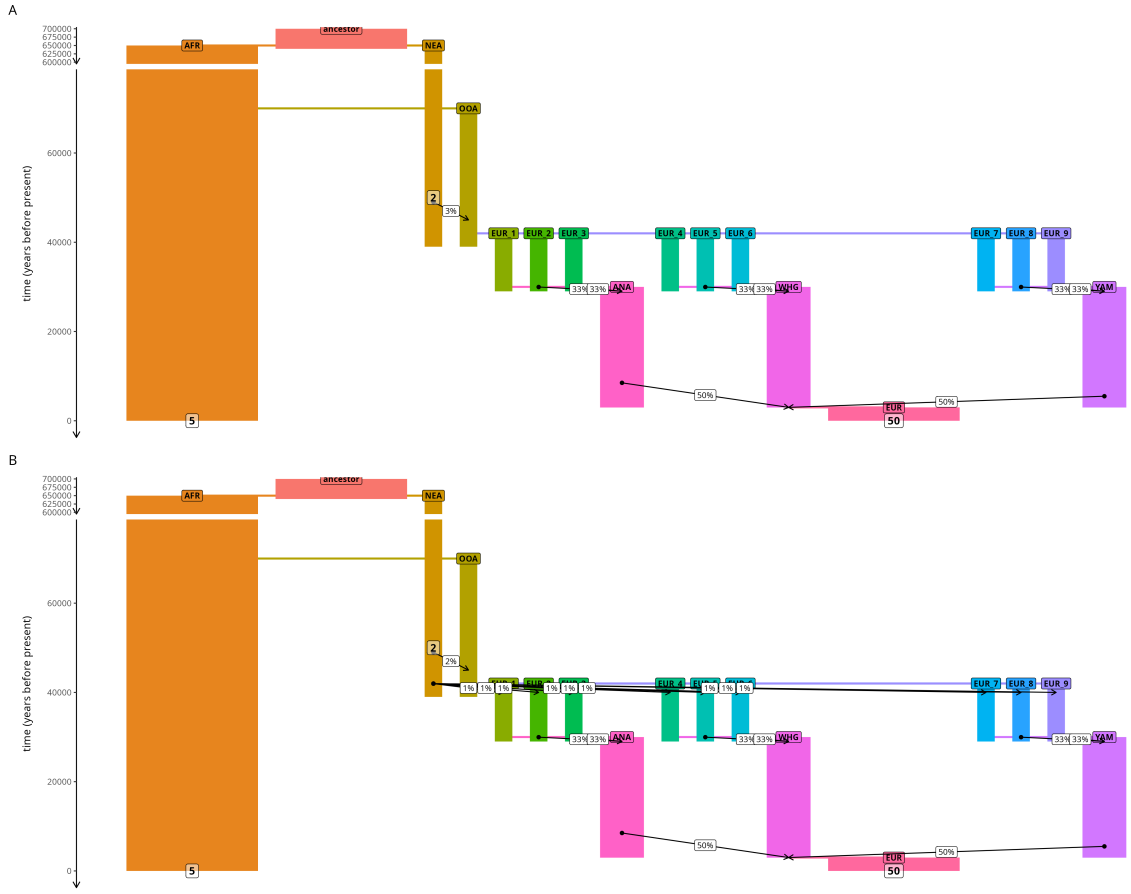

Figure 1.2: Schematic representation of the demographic history of Model 2. (**A**) shows the variant **A** of Neanderthal ("NEA") introgression for this model, with one single extended pulse of introgression into the Out-of-Africa ("OOA") population. (**B**) shows the variant **B** of "NEA" introgression for this model, with two independent extended pulses of introgression: one into the "OOA" and one into each "EUR" deme. The arrows indicate the gene flow events, and the percentages correspond to the proportion of migration. The numbers within boxes correspond to the number of individuals sampled from that population at a given time point. The y-axis represents the time in years before present, with gaps representing our modification to the default plotting procedure of the *slendr* package to facilitate clearer plotting and model verification.

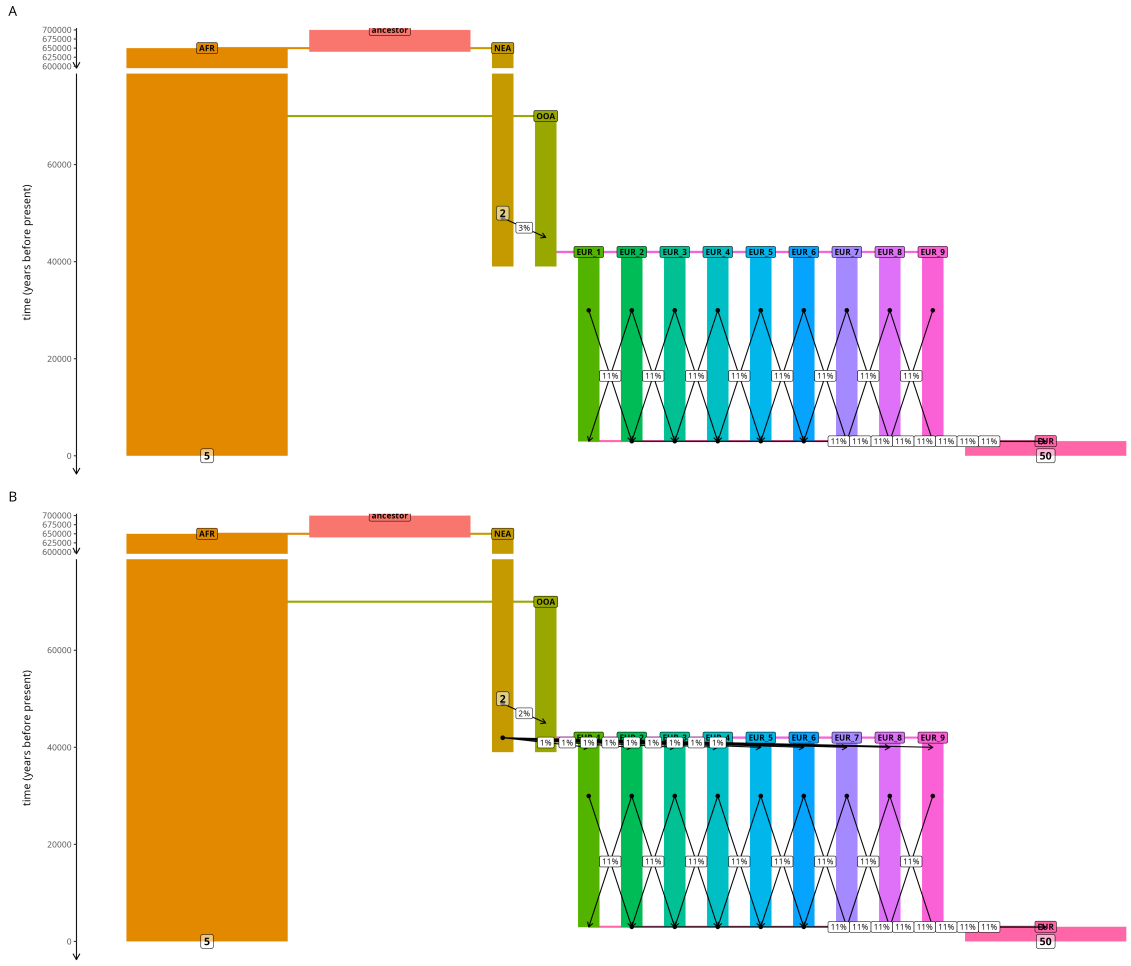

Figure 1.3: Schematic representation of the demographic history of Model 3. (**A**) shows the variant **A** of Neanderthal ("NEA") introgression for this model, with one single extended pulse of introgression into the Out-of-Africa ("OOA") population. (**B**) shows the variant **B** of "NEA" introgression for this model, with two independent extended pulses of introgression: one into the "OOA" and one into each "EUR" deme. The arrows indicate the gene flow events, and the percentages correspond to the proportion of migration. The numbers within boxes correspond to the number of individuals sampled from that population at a given time point. The y-axis represents the time in years before present, with gaps representing our modification to the default plotting procedure of the *slendr* package to facilitate clearer plotting and model verification.

1.2.2 Comparison of heterozygosity values

Although intended to capture interesting of real-world history compared to the idealized, single-population panmictic models, the alternative scenarios implemented in this chapter (i.e., **Models 1, 2** and **3**) are still quite idealized, and, at this stage of the project, have not been backed by rigorous modeling. Nevertheless, we were interested in assessing how well would these models align with empirical data. To this end, we computed individual heterozygosities in present-day African ("AFR") and West Eurasians ("EUR") individuals recorded in tree-sequences simulated from each of the three models, and compared their distributions against empirical heterozygosity values computed from the Simons Genome Diversity Project (SGDP) data set [83](Supplementary Table 1).

Figure 1.4 shows the results for both populations across all of our simulations. As expected given that the models have been implemented as initial frameworks for more detailed inference, it is clear that none of the scenarios align with the observed data, with the heterozygosity distributions in both the simulated populations being significantly lower relative to the empirical data (which would, at a later stage, become one of the target statistics in future inference).

As expected, heterozygosities obtained for the "AFR" population do not change regardless of the model of reference or the N_e of the "OOA" population. This is expected, as the history of the "AFR" population in our models remains unaltered in all of our models and it is not influenced by the N_e of the "OOA" population.

On the other hand, the distribution of the individual-based heterozygosities of the "EUR" population changes significantly as function of the N_e of the ancestral "OOA" population. Specifically, lower N_e values, representing a stronger bottleneck of the ancestral "OOA" population, lead consistently to an increased loss of heterozygosity in present-day "EUR" samples. Moreover, the different model scenarios, regardless they exhibit more complex demic structure or not, do not show significant changes in their heterozygosity values.

It is also worth mentioning that, at least in this qualitative comparison, the two variants of the models involving population structure within the earliest prehistoric Eurasian populations, **A** and **B**, do not seem to differ in terms of expected heterozygosities.

A detailed description of how the heterozygosity was computed from the simulated tree-sequence data structure using the *slendr* R package, as well as how we obtained and summarized the empirical heterozygosities from the SGDP panel can be found in section 1.1.2.

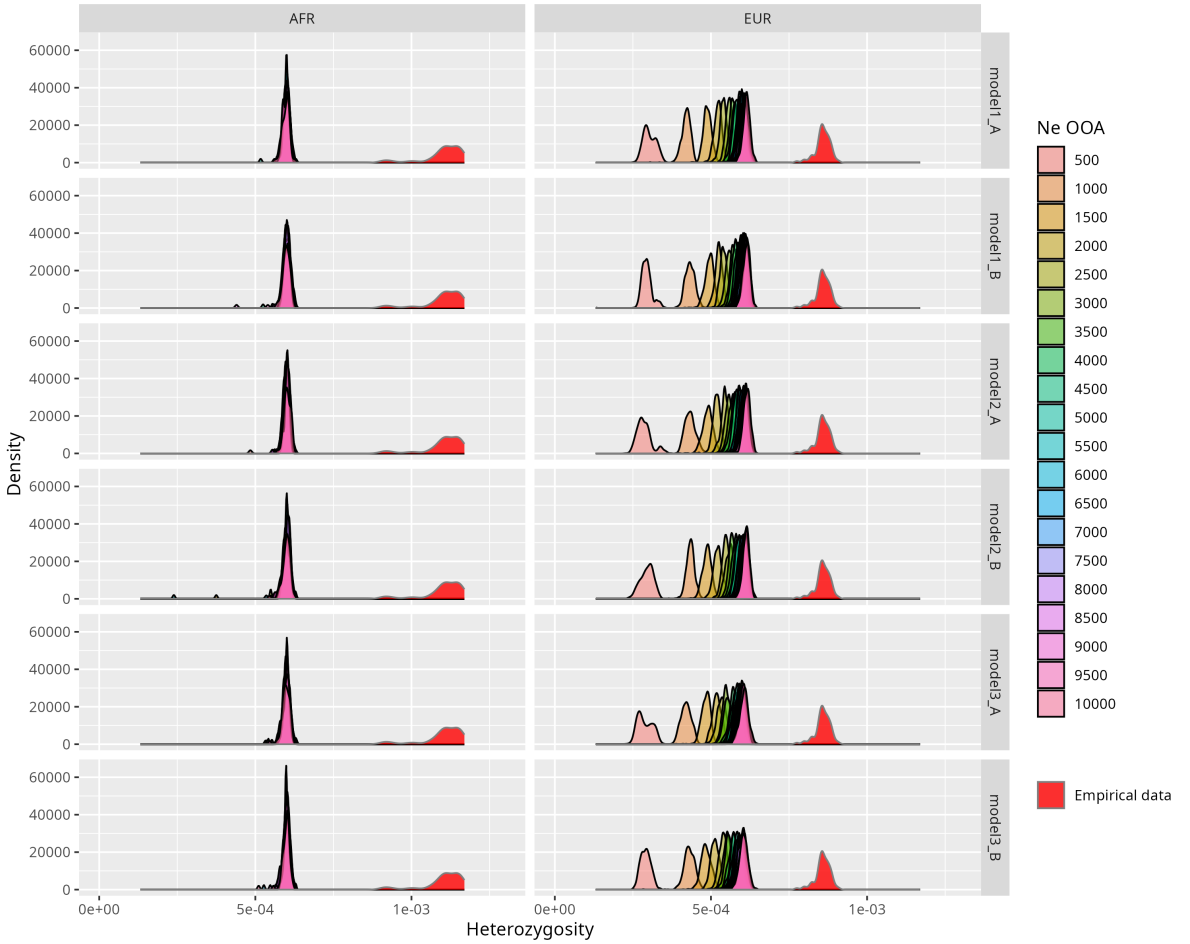

Figure 1.4: Comparison of individual-based heterozygosity distributions between simulated and empirical data. The figure shows the plots of the distribution of the heterozygosity for all of our simulated models against empirical data (vertical order of facets). The left facet column ("AFR") corresponds to the distribution for the African population, while the right facet column ("EUR") corresponds to the distribution for the West Eurasian population. The empirical heterozygosity distributions are shown in red. The legend to the right shows the color for different N_e of the "OOA" population.

1.3 Discussion

In the first chapter of this study, we developed a set of proof-of-concept demographic models as potential alternatives to the traditionally used single-population panmictic models of modern human history. In particular, we focused on the population structure during after the Out-of-Africa migration, subsequent Neanderthal introgression, and the history of the earlier modern humans in Eurasia . Our overarching goal was to provide an initial framework for future simulation-based inference of detailed population and social dynamics of prehistoric populations.

Visualization of the topology of our models verified that the model scenarios that we implemented in the *slendr* simulation package corresponded to our specifications (see section). In the process, we also provided a fix to *slendr* developers which improves the readability of model representations generated with *slendr*, particularly for complex models.

However, when we evaluated how well our models aligned with empirical data using genome-wide heterozygosity as the comparison metric, we assessed that our models were still relatively far from real data (see Figure 1.4). This is unsurprising, because our proposed models are not yet based on rigorous statistical inference, and highlights the need for careful model-selection procedure across a range of informative summary statistics. Moreover, as we hypothesized, simple population genetic metrics such as genome-wide heterozygosity do not necessarily have enough discriminatory power to distinguish our different demographic scenarios, which is why we sought to develop other, more powerful, summary statistics in the following chapter.

Although genome-wide heterozygosity metrics are important summary statistic that should certainly be used in future simulation-based inference, such as Approximate Bayesian Computation, the discrepancy in values obtained between simulated and empirical data is not of particular concern at this stage of our project. Our primary goal, which we addressed in the next two chapters, was to evaluate how the different scenarios are reflected in the new summary statistics that we implemented.

Chapter 2: Developing tract-based summary statistics

2

Having successfully established several candidate models of varying levels of population structure in prehistoric Eurasia in Chapter 1, we dedicated the second chapter of this study to exploring the potential of Neanderthal introgressed tracts as a source of information to fit models of early human demographic history.

In this regard, we implemented four different approaches to encode information about introgressed tracts in form of binary matrices which we then studied in this and the following chapter. Each of these approaches extracts different informative patterns from the data and our goal was to evaluate, which of these patterns might be potential useful for modeling human prehistory, population dynamics, and social structure.

Building on these four tract-encoding approaches, we developed an introgression tract-based summary statistic that we call the tract frequency spectrum (TFS), which is conceptually equivalent to the site frequency spectrum (SFS), but consider the distribution of introgressed tracts sharing among individuals. We explored the TFS using the several candidate models of varying levels of population structure in prehistoric West Eurasia that we established in Chapter 1.

Additionally, as an alternative approach of extracting useful information from the tracts, we explored the possibility of reconstructing the spatio-temporal relationships between individuals by building a graph structure based on the sharing of introgressed tracts.

2.1 Materials and Methods

2.1.1 Extracting true coordinates of Neanderthal from simulations

In all the simulations with admixed sampled individuals, we extracted the exact genomic coordinates of Neanderthal introgressed tracts into a tabular format using the *slendr* function *ts_tracts()*, which exploits a tree-sequence algorithm to uniquely label each ancestry tract to its known true source population [85].

For simulated scenarios with two distinct episodes of archaic admixture, introgressed tracts were extracted separately for each event, as the *ts_tracts* function extracts tracts based on the starting time of a gene flow event. In real genomic data, it is not possible to distinguish which tracts come from which admixture event. Therefore, when adjacent tracts were found on the same sampled chromosome, we merged them using the *reduce()* function from the *GenomicRanges* R package [86]. This means that, adjacent tracts originated from two different gene flow event, were reduced into a single tract by keeping the smallest start coordinate and the largest end coordinate.

2.1.2 Methods for encoding introgressed tracts into binary matrices

The four approaches we developed to encode tracts from a tabular format into binary matrices were developed in R using custom functions. Each approach led to a resulting binary matrix, where the columns in the matrix correspond to an individual, or haplotype, while the rows, depending on the approach, correspond to a tract or genomic bin. For instance, for the *uniqueness approach* a row correspond to a precise tract defined by the recombination breakpoints, for the *windows approach* the rows correspond to a genomic windows, and so on. Rows which entries were all zeros were removed from the matrices.

Each approach uses a slightly different algorithm:

- For the *uniqueness approach*, we selected all the existing tracts sampled in the population and we extracted only the unique tracts. Subsequently, we used the function *findOverlaps()* from the *GenomicRanges* R package to assign which tracts among the all set were present in an individual.

- For the *windows approach*, we divided the chromosome into non overlapping fixed-size bins (50 kb) using the *slidingWindows()* function from the *GenomicRanges* R package. Subsequently, we used the function *findOverlaps()* from the *GenomicRanges* R package to assign which tracts sampled in an individual overlap with a bin.
- For the *sub-tracts approach*, we extracted all the possible unique recombination breakpoints observed in the population. Subsequently, we used the coordinates of the breakpoints to define the genomic bins, and, as for the previous two approaches, we used *findOverlaps()* from the *GenomicRanges* R package to assign which bins are present in an individual.
- For the *recombination breakpoints approach*, we extracted all the possible unique recombination breakpoints observed in the population. Subsequently, we simply check which recombination breakpoints in an individual are present among the entire set of observed recombination breakpoints.

2.1.3 Distribution of introgressed tract lengths in simulations

To calculate the distribution of introgressed tract lengths in our models, we performed 10 simulation replicates per model. For each model, we aggregated the lengths of all extracted tracts in sampled chromosomes across replicated into a single vector. We then obtained the distribution using the base R function *hist()*, specifying the bins width to 20 kilo bases (kb), ranging from 0 to 7 Mega bases (Mb).

2.1.4 Site frequency spectrum (SFS) of introgression informative alleles

The site frequency spectrum (SFS) is a summary statistics widely used in population genetics that describes the distribution of sharing of derived alleles in a population sample.

The SFS is generally represented by a histogram, where each bin, indexed by i, encodes the count of alleles shared by i sampled chromosomes. The value of i is discrete and it ranges between 1 and $2N$, with N being the number of diploid sampled individuals.

In our study, we computed the SFS of introgression informative alleles which are derived in Neanderthals. We defined an allele as introgressed if its state in all sampled

African individuals is ancestral and different from its state in all sampled Neanderthals. In other words, we conditioned on sites which carried alleles that were fixed derived in Neanderthals but absent in any sampled Africans. Once we identified these sites, we computed the SFS for each of our models. Each model was simulated 10 times, and the resulting SFS correspond to the average value over the 10 replicates. For each frequency class (bin), we also computed the standard deviation across replicates.

For graphical purposes, we plotted the SFS as line graphs rather than histograms.

2.1.5 Tract frequency spectrum (TFS)

The tract frequency spectrum (TFS) is a summary statistics that we developed to study the sharing of introgressed tracts among individuals. It is conceptually equivalent to the SFS.

In our study, we computed the TFS for each of our models. Analogously to the SFS of introgressed alleles (see section 2.1.4), we simulated each model 10 times. The resulting TFS correspond to the mean and standard deviation across the 10 replicates.

To compute the TFS starting from the tract-encoding binary matrices (see section 2.2.1), we simply obtained a vector of the counts of shared entries among individual in the binary matrix using the base R function *rowSums()*. Subsequently, we used the base R function *table()* to count the frequency of each count class.

As for the SFS (see section 2.1.4), for graphical purposes, we plotted the TFS as line graphs rather than histograms.

2.1.6 Graph-based tract sharing data structure

To obtain a graph data structure, starting from the adjacency matrix described in section 2.2.7, we used the *graph_from_adjacency_matrix()* function of the *igraph* R package [87].

To evaluate whether our tract genealogy graph is a qualitatively reasonable approximation of the true genealogy, we simulated a simple model. Once again, for consistency with the overall project, we used as general structure of the model the Out-of-Africa event followed by the admixture with a Neanderthal population. However, we disregarded dates and parameters found in literature, as not relevant for our conceptual example. We sampled 5 individuals from each admixed population at different time points. For admixed sampled individuals, we extracted the introgressed tracts (see section 2.1.1). Focusing on tracts intersecting a specific genomic location, we obtained

the true genealogy of these tracts stored in the tree-sequence using the *ts_phylo* function of the *slendr* package [77].

A schema of the model (see Figure S8) and a table with the model parameters (see Table S4) can be found in the Supplementary Information.

2.1.7 Correlation of f_3 and Neanderthal tract-sharing

To evaluate how different introgression scenarios would affect the analysis first proposed by Iasi *et al.* [40], which compares the pairwise f_3 statistics with the pairwise correlation of Neanderthal tracts, we designed a toy demographic model.

For consistency with the overall project, the general topological structure of the model is again related with the Out-of-Africa event followed by introgression with Neanderthals.

The model includes an Out-of-Africa (OOA) population that splits into two populations POP1 and POP2, that subsequently diversify into other lineages. Specifically, POP1 divides into POP1a and POP1b, which later splits into POP1bx and POP1by. POP2 instead divides into POP3 and POP4, which later respectively branches into POP3a, POP3b and POP4a, POP4b. All the aforementioned populations had a fixed $N_e = 3000$. Table 2.1 shows the model parameters, with the timing of the aforementioned splits.

The comparison was made evaluating three different introgression scenarios:

- **Scenario 1**: a single extended pulse (from 66 to 61 kya) of Neanderthal introgression into the OOA population.
- **Scenario 2**: two independent pulses of Neanderthal introgression (from 55 to 51 kya) into POP1 and POP2.
- **Scenario 3**: private pulses of Neanderthal introgression (from 12 to 8 kya) into any of the final lineages.

Parameter	Description	Value	Unit
$T_{AFR\&NEA}$	AFR and NEA splits from ANC	650	kya
T_{OOA}	OOA branches from AFR	70	kya
$T_{POP1\&2}$	OOA splits into POP1 and POP2	62	kya
$T_{POP1a\&1b}$	POP1 splits into POP1a and POP1b	30	kya
$T_{POP1bx\&1by}$	POP1b splits into POP1bx and POP1by	15	kya
$T_{POP3\&4}$	POP2 splits into POP3 and POP4	51	kya
$T_{POP3a\&3b}$	POP3 splits into POP3a and POP3b	20	kya
$T_{POP4a\&4b}$	POP4 splits into POP4a and POP4b	12	kya
gf_rate	Any gene flow from NEA	3	%
N_{NEA}	Neanderthal population size	1000	
N_{OOA}	OOA population size	3000	
N_{POP}	Population size all POP	3000	
N_{AFR}	African population size	10000	
N_{ANC}	Ancestral population size	10000	
g	Generation time	30	years
Chr_{length}	Chromosome length	50	Mb
n_{samp}	Sample size for each sampling	5	diploid genomes
t_{samp}	Sampling time points	0	kya

Table 2.1: Parameters used in the toy model to compare pairwise f_3 statistics with the pairwise correlation of Neanderthal tracts.

2.2 Results

2.2.1 Approaches for encoding introgressed tracts into binary matrices

When extracting the coordinates of introgressed tracts from a set of admixed individuals using the *slendr* R package [77] (see section 2.1.1), the *ts_tracts()* function returns them in a tabular format (see Table 2.2).

Because the tabular format of Neanderthal tracts (see Table 2.2) is not ideal for computational analysis, we developed 4 different approaches to encode the tracts into a more manageable data structure. Each approach produced a binary matrix, that facilitated the application of the tract-based summary statistics that we developed in the next sections.

chrom	start	end	name
chr1	100	150	IND_1
chr1	100	150	IND_2
chr1	130	150	IND_3
chr1	90	170	IND_4
chr1	90	140	IND_5
chr1	50	80	IND_6

Table 2.2: Toy example of a tabular format of introgressed tracts. For simplicity, each individual carries a 200 base pair (bp) diploid genome, and only one tract per individual on a single chromosome is shown. The column *chrom* corresponds to the chromosome, *start* and *end* are the genomic coordinates of the tract, and *name* refers to the individual ID.

We defined the 4 different tract-encoding approaches as:

- *Uniqueness approach.*
- *Windows approach.*
- *Sub-tracts approach.*
- *Recombination breakpoints approach.*

Uniqueness approach

The *uniqueness approach* consists in obtaining all the unique tracts present in the pool of sampled introgressed tracts. In the resulting binary matrix, each row represents

a sampled chromosome from an individual, and each column of the matrix represent a unique tract. An entry in the matrix will be 1 if the corresponding chromosome have that tract, and 0 otherwise.

Figure 2.1 shows an example of this approach based on the toy example of Table 2.2.

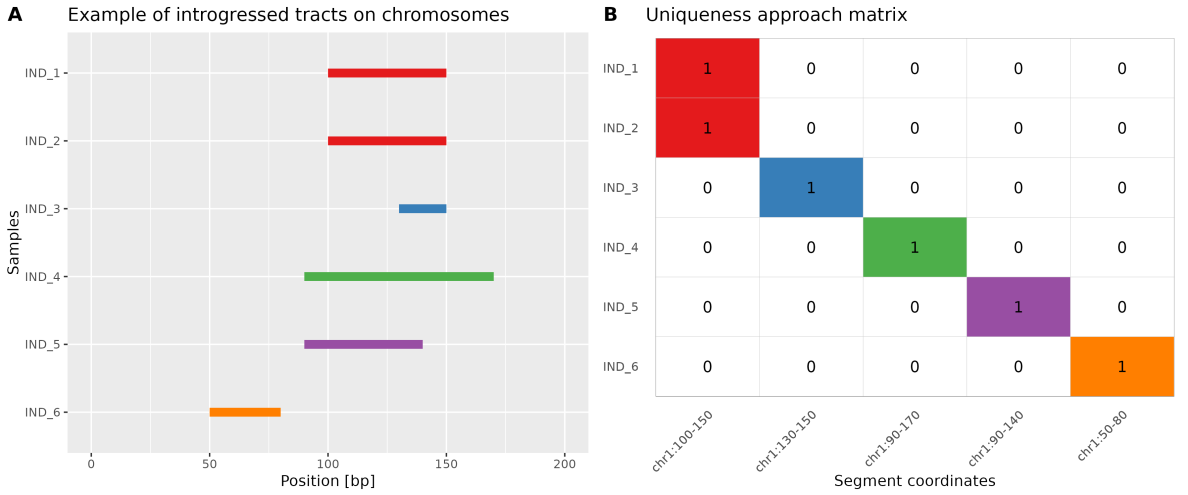

Figure 2.1: Visual example of the *uniqueness approach*. (**A**) shows a visual representation of the sampled tracts across sampled chromosomes. The x-axis represents the genomic position, the y-axis the sampled chromosomes. (**B**) shows the corresponding binary matrix representation, with the cell colors matching the tract colors in (**A**). *IND_1* and *IND_2* are the only two chromosomes that share one unique tract.

Windows approach

The *windows approach* consists in dividing the genome into fixed size, non overlapping bins. In the resulting binary matrix, each row corresponds to a sampled chromosome from an individual, and each column to a genomic bin. An entry in the matrix will be 1 if the corresponding chromosome have a tract overlapping that bin, and 0 otherwise.

Figure 2.2 shows an example of this approach based on the toy example of Table 2.2. The 200 bp genome have been divided into 8 non overlapping bins of 25 bp each.

When using this approach, we always arbitrarily set the window width to 50 kb.

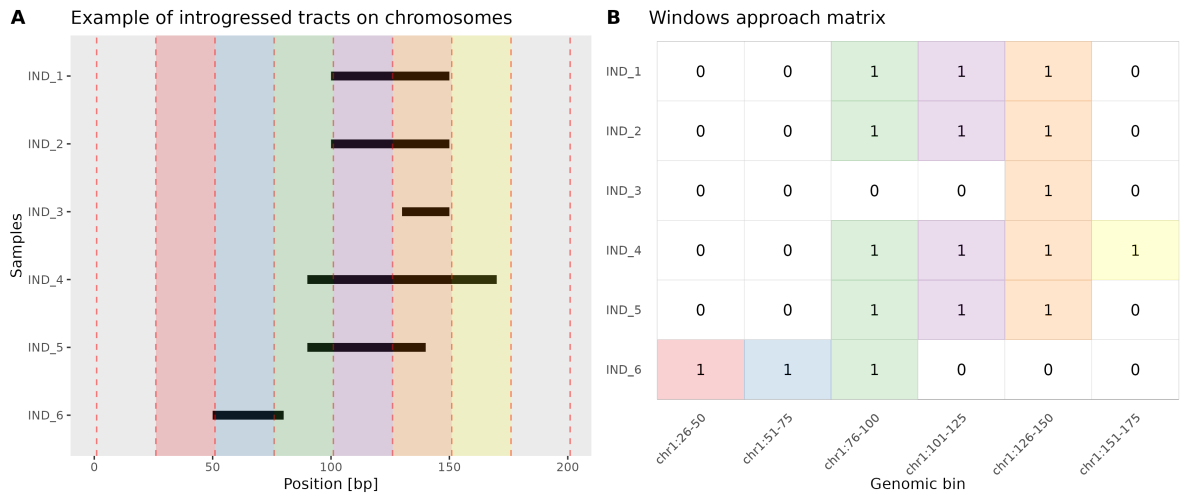

Figure 2.2: Visual example of the *windows approach*. (**A**) shows a visual representation of the sampled tracts across sampled chromosomes. The x-axis represents the genomic position, the y-axis the sampled chromosomes. The vertical red dashed lines represents the boundaries of the 25 bp genomic bins. (**B**) shows the corresponding binary matrix representation, with the cell colors matching the bin colors in (**A**). Note that a matrix entry is set to 1 if the tracts overlap the bin even by only 1 bp.

Sub-tracts approach

The *sub-tracts approach* consists in extracting all the possible sub-tracts from the pool of sampled introgressed tracts. Therefore, this is equivalent to divide the genome into variable size bins, where the bin boundaries are defined by all the existing start and end coordinates of the tracts. In the resulting binary matrix, each row corresponds to a sampled chromosome from an individual, and each column to one of these variable genomic bins. An entry in the matrix will be 1 if the corresponding chromosome have a tract overlapping that bin, and 0 otherwise. Note that a different pool of sampled tracts will result in a different set of variable size genomic bins.

Figure 2.3 shows an example of this approach based on the toy example of Table 2.2.

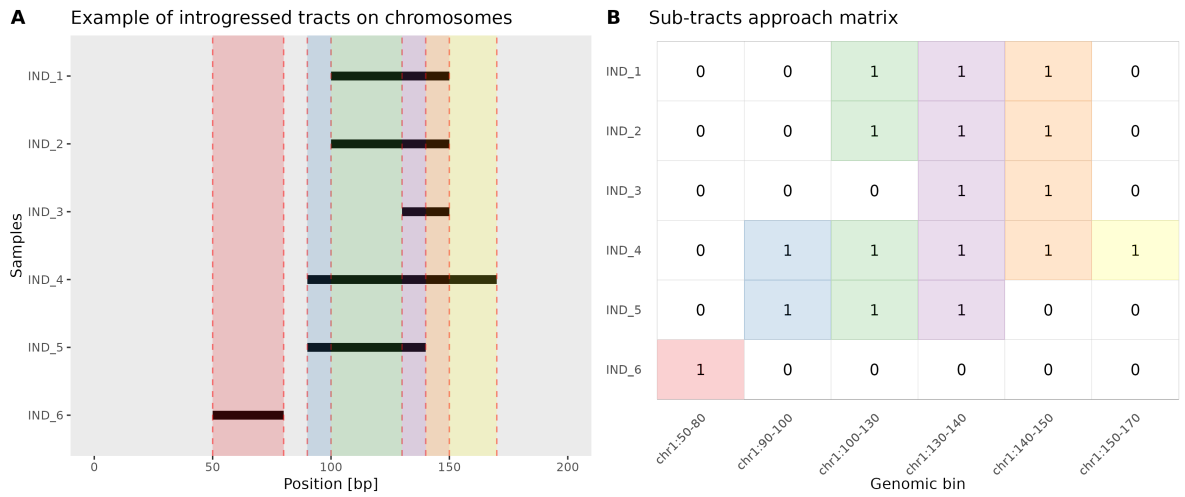

Figure 2.3: Visual example of the *sub-tracts approach*. (**A**) shows a visual representation of the sampled tracts across sampled chromosomes. The x-axis represents the genomic position, the y-axis the sampled chromosomes. The vertical red dashed lines represents the boundaries of the variable size genomic bins. (**B**) shows the corresponding binary matrix representation, with the cell colors matching the bin colors in (**A**).

Recombination breakpoints approach

The *recombination breakpoints approach* consists in extracting all unique start and end coordinates from the pool of sampled introgressed tracts. In other terms, this is equivalent to "record" all the possible recombination breakpoints that occurred. In the resulting binary matrix, each row corresponds to a sampled chromosome from an individual, and each column to one of these recombination breakpoints. An entry in the matrix will be 1 if the corresponding chromosome have a tract that has a start or end coordinate in that site, and 0 otherwise.

Figure 2.4 shows an example of this approach based on the toy example of Table 2.2.

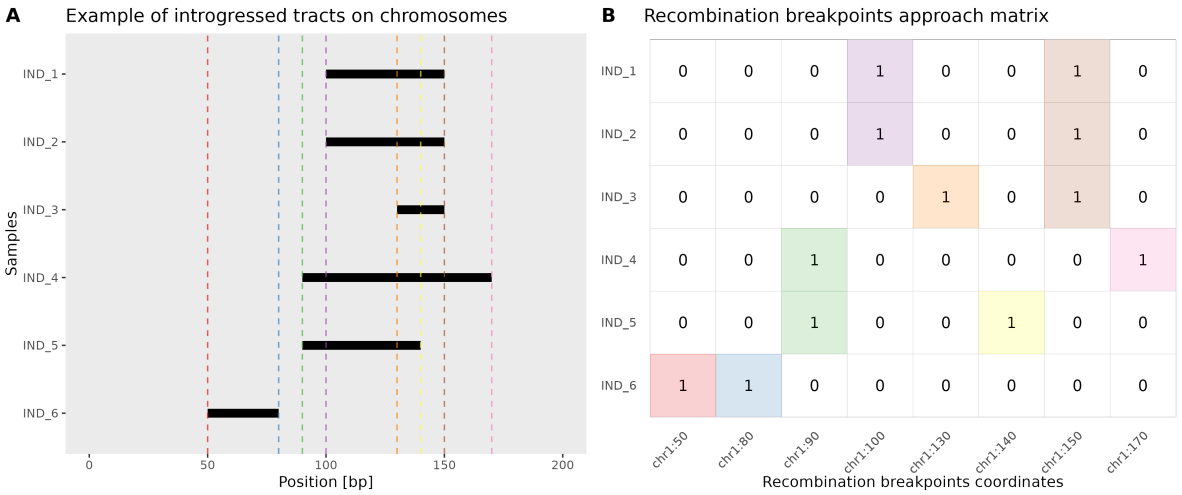

Figure 2.4: **Visual example of the *recombination breakpoints approach*.** (**A**) shows a visual representation of the sampled tracts across sampled chromosomes. The x-axis represents the genomic position, the y-axis the sampled chromosomes. The vertical colored dashed lines represents the all the observed recombination breakpoints. (**B**) shows the corresponding binary matrix representation, with the cell colors matching the dashed lines colors of the recombination breakpoints in (**A**).

2.2.2 Tract frequency spectrum (TFS)

The tract frequency spectrum (TFS) is a summary statistic that we developed to study the sharing of introgressed tracts in a population. It is conceptually equivalent to the site frequency spectrum (SFS), but instead of describing the distribution of sharing of derived alleles, it quantifies the distribution of shared introgressed tracts. Note that the meaning of "tracts" varies according to the encoding approach used (see 2.2.1). For instance, in the *windows approach*, we describe the sharing of tracts in the genomic bins defined by the windows. In the *recombination breakpoints approach*, we describe the sharing of recombination breakpoints, and so on.

The TFS is essentially a histogram, where each bin, indexed by i, encodes the count of tracts shared by i sampled chromosomes. The value of i is discrete and it ranges between 1 and $2N$, with N being the number of diploid sampled individuals. For comparison, instead of using the raw counts of "tracts", we normalized the TFS by considering the proportion of counts in each bin.

Figure 2.5 shows the TFS for the four different tracts encoding approaches based on the toy example of Table 2.2.

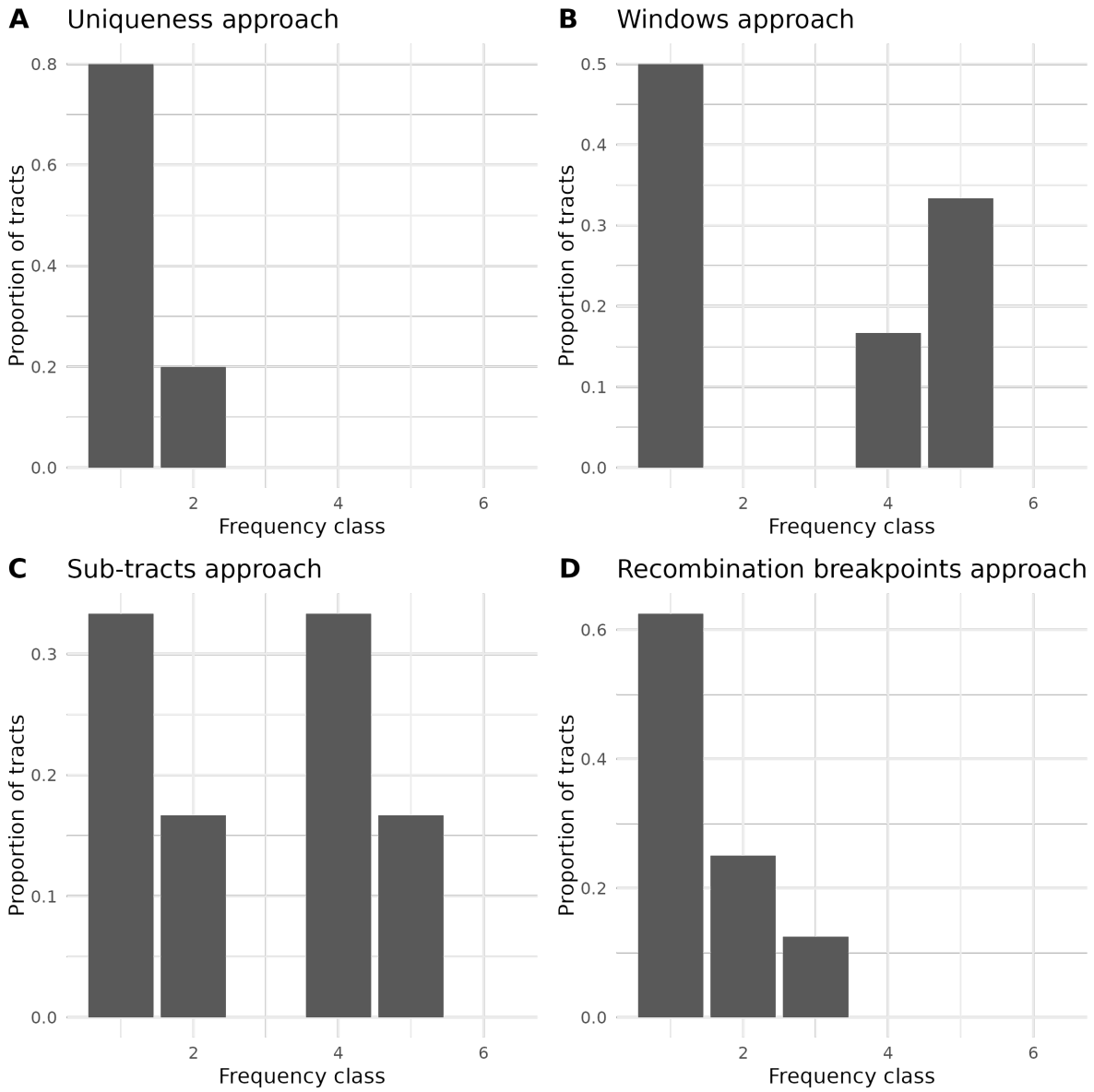

Figure 2.5: TFS for the toy example of Table 2.2. (**A**) corresponds to the TFS obtained using the uniqueness approach. (**B**) corresponds to the TFS obtained using the windows approach. (**C**) corresponds to the TFS obtained using the sub-tracts approach. (**D**) corresponds to the TFS obtained using the recombination breakpoints approach. Note how different tracts encoding approaches result in significantly different TFS.

For graphical purposes, we plotted the TFS as line graphs rather than histograms in the next plots of this manuscript.

2.2.3 Distribution of tract lengths

The distribution of introgressed tract lengths has been a commonly used summary statistics for studying introgression, mainly as a tool to infer the timing of admixture events [88, 89]. This is based on the idea thet after gene flow between two lineages occurs, the expected length of introgressed tracts decreases over generations because of the shuffling generated by the recombination process. Under some simplifying assumptions, the lengths of introgressed tracts then follows an exponential distribution.

Following this idea, we evaluated the tract length distribution resulted from each model scenario. In doing so, our goal here was not dating the admixture event per se, but rather to perform a sanity check of the simulation results and to investigate whether a potentially informative pattern arises between the different models.

Figure 2.6 shows the comparison between the distribution of tract lengths for the different models and variants. First of all, as expected, the distributions of tract lengths are not dramatically influenced by the value of the N_e of the OOA population that experienced admixture. For both the variants of introgression, **A** and **B** (respectively left and right columns in Figure 2.6), it is also clear how the distribution of tract lengths is invariant to the demographic history of the West Eurasian population as there are no significant differences between the different models. However, for the variant **B** it seems that the density of tracts at lower lengths is slightly lower. This might be due to the fact that second pulse of introgression occurring closer in time to the sampling point introduced a set of slightly longer tracts to the pool of tracts.

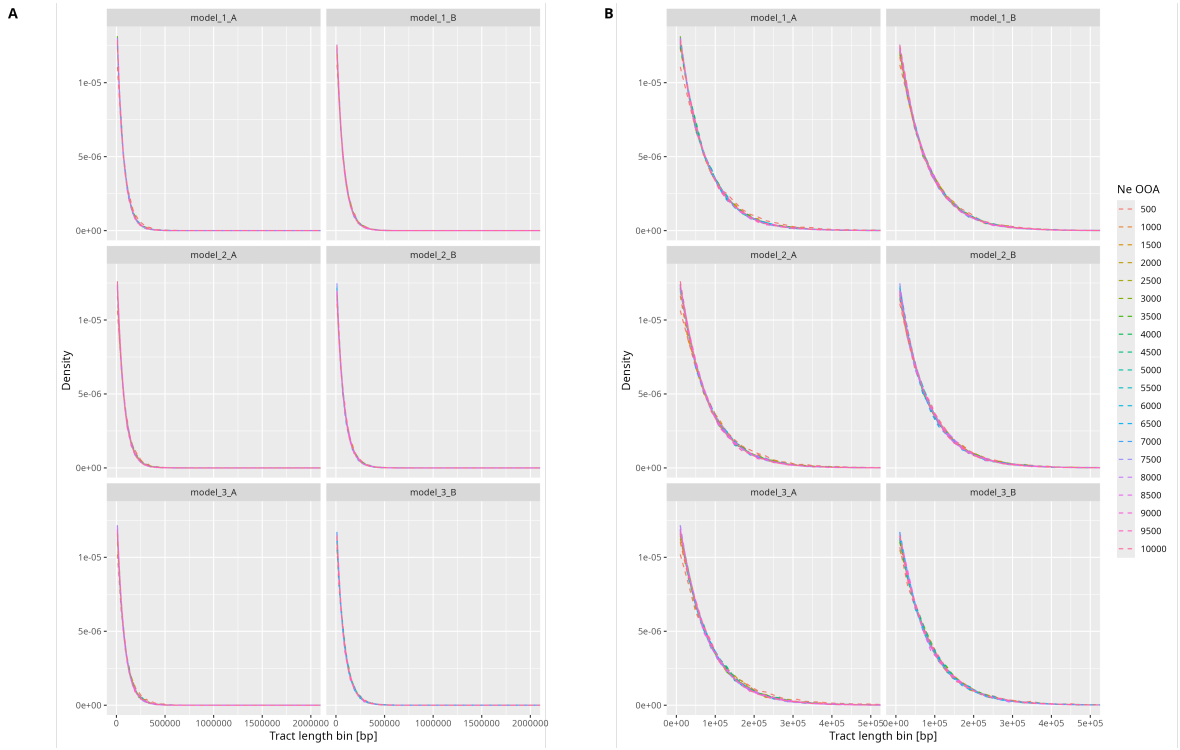

Figure 2.6: Comparison of the distributions of introgressed tract lengths between different models. (**A**) shows the full distribution of tract lengths between the different models. (**B**) corresponds to a zoomed version of **A**, where the x-axis has been limited between 0 and 500 kb. Each grid corresponds to the distribution of introgressed tract lengths for a certain model and variant. Each color of the curves corresponds to a different value of the N_e of the Out-of-Africa population ("OOA"). The x-axis represents the lengths of tracts in base-pairs (bp), while the y-axis represents the density.

2.2.4 Tract frequency spectrum on simulations

The tract frequency spectrum (TFS), previously introduced in section 2.2.2, is a summary statistics that we explored as a potential source of valuable information about the demographic history of the earliest Eurasian populations after the OOA event following the Neanderthal introgression, as early modern human populations spread to Eurasian (section 2.2.2). To describe the characteristics of TFS for this purpose, we computed it across all model scenarios, using each of the four tract-encoding approaches presented in section 2.2.1.

Figure 2.7 and 2.8 show the comparison of the TFS between different models and tract-encoding approaches. Unlike the distribution of tract lengths (see Figure 2.6), the

TFS varies substantially depending on the tract-encoding approach, model, introgression variant, and value of the N_e of the Out-of-Africa population that we consider.

First focusing on the comparison of TFS between different tract-encoding approaches, we found that methods based on the sharing of recombination events, such as the *uniqueness approach* and *recombination breakpoints approach*, are invariant to the value of the N_e of the OOA population that received the main pulse of admixture (see Figure 2.7). Their TFS seem following an exponential distribution, likely because these two approaches are strictly dependent mainly on the cascades of recombination events that generate the recorded tracts, rather than the sharing of introgressed genomic regions. This is consistent even when comparing these two approaches across the different models. Between the two approaches, the *uniqueness approach* tends to have more singletons because only exact tracts sharing both the recombination breakpoints are counted. This is unlike the approach based on sharing of recombination breakpoints, which considers the count of single recombination breakpoints shared.

On the other hand, the remaining two tract-encoding approaches, based on sharing of overlapping genomic windows, produce significantly different patterns of TFS (rows "*windows approach*" and "*sub-tracts approach*" in Figure 2.7). These approaches reveal much more sharing on higher frequencies across all the models and are sensitive to the value of N_e of the Out-of-Africa population. In particular, lower values of N_e tend to result in flatter TFS curves, with increased sharing of "tracts" and less proportions of rare "tracts" in the lower frequency classes, e.g. "tracts" shared by only one chromosome (singletons). Additionally, the two different variants of introgression, **A** and **B**, consistently give different results. Specifically, the variant **B** models always show an increased proportion of rare "tracts" at low frequency of sharing.

Finally, these two approaches show different shapes of their TFS. In particular, the sub-tracts approach tends to show much more sharing (i.e., higher proportion of common tracts at a given introgressed locus) compared to the windows approach, where the proportion of rarer alleles seem to decrease. This is likely due to how the genomic binning was defined, with the sub-tracts approach having a much finer scale that resulted in increased counts of shared tracts.

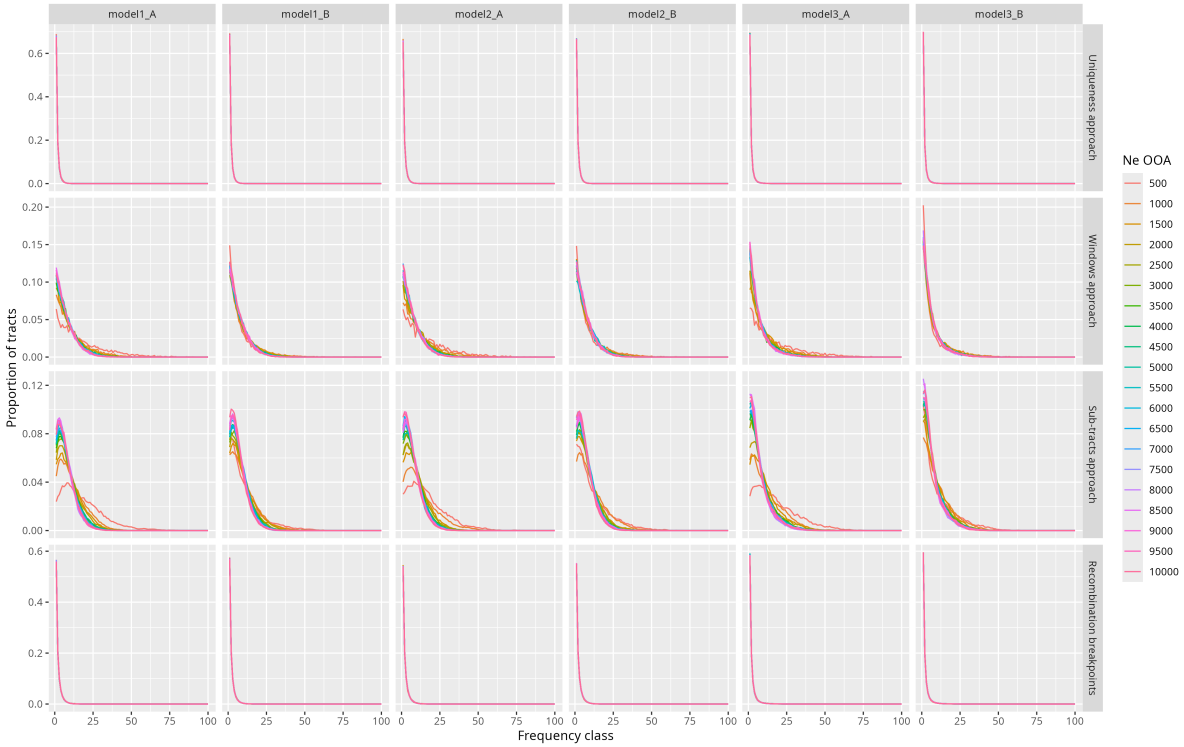

Figure 2.7: Comparison of the Tract Frequency Spectrum between different models and approaches colored by N_e of OOA. Each grid corresponds to the TFS computed for a certain model with a certain tract-encoding approach. Each color of the curves corresponds to a different value of the N_e of the Out-of-Africa population ("OOA"). Each column represents a different model for an introgression variant, **A** or **B**. Each row indicates the type of tract-encoding approach used. The x-axis represents the frequency class of sharing between chromosomes. The y-axis represents the proportion of counts for each frequency class.

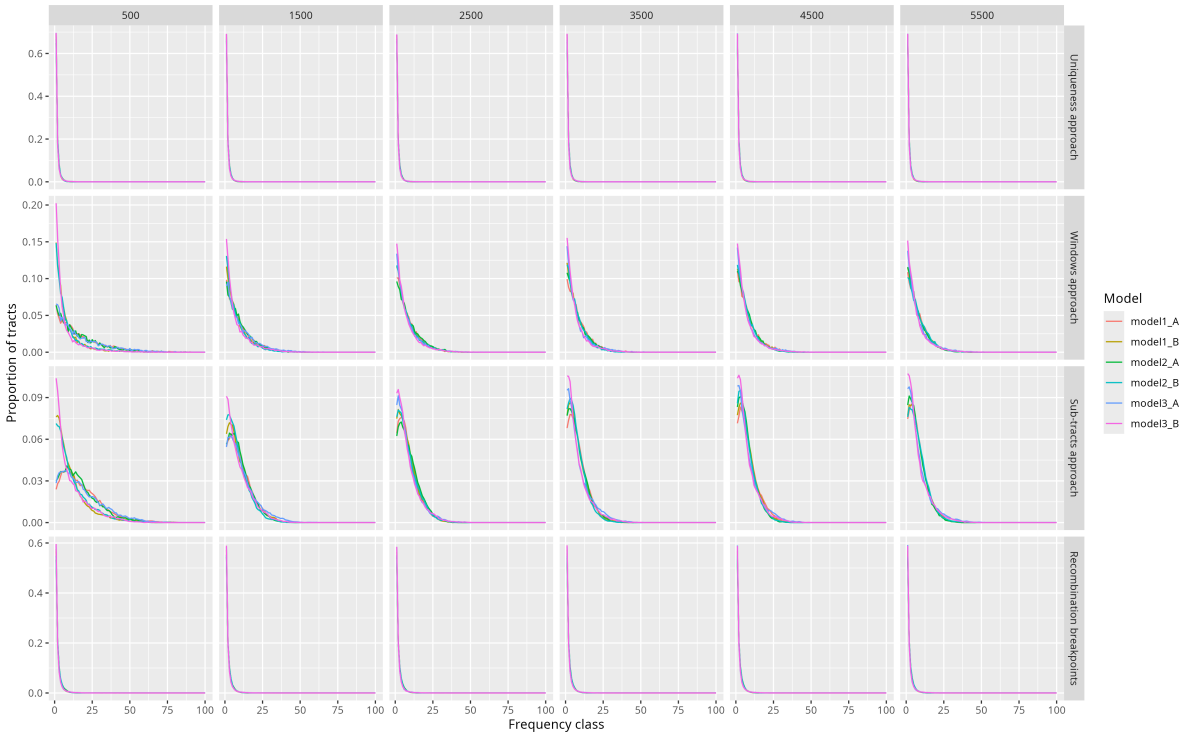

Figure 2.8: Comparison of the Tract Frequency Spectrum between different models and approaches colored by Model. Each grid corresponds to the TFS of the models computed for a certain value of the N_e of the Out-of-Africa population ("OOA") with a certain tract-encoding approach. Each color of the curves corresponds to a different model. Each column represents a different N_e of the OOA. Each row indicates the type of tract-encoding approach used. The x-axis represents the frequency class of sharing between chromosomes. The y-axis represents the proportion of counts for each frequency class.

2.2.5 Site frequency spectrum on introgression informative sites

Although containing a complete information about the spatio-temporal (between individuals across space and time) and genomic (within a genome) distributions of introgression, inferring Neanderthal tracts in real data is not a perfect process, and, as with any statistical procedure, necessarily involves some degree of error and uncertainty. To complement the tract-based approach with another, independent, method, we evaluated another introgression-based statistic computed on so-called "introgression informative sites". Briefly, this approach, also used in aDNA literature [35, 90], is based on filtering the genome for loci at which all individuals in a selected African group (such as Yoruba) all carry an ancestral allele but a set of Neanderthal genomes all carry a derived allele. As a result, estimating the proportion Neanderthal ancestry in a given individual involves computing the proportion of derived alleles at such conditioned loci. In a future work on developing a model inference procedure, we envisioned summary statistics based on these conditioned sites as a complementary source of information about human demography and introgression dynamics.

As a proof of concept, and to perform a comparative evaluation of the TFS metric as a summary statistic, we computed the site frequency spectrum (SFS) of alleles at introgression informative sites and compared it with the TFS results of the previous section. Details on how we computed the SFS of introgressed alleles are in section 2.1.4.

Figure 2.9 shows a comparison of the SFS of introgression informative alleles across the different models. The resulting plots are broadly consistent with the results in Figure 2.7, specifically with the TFS computed based on the *windows approach*. In particular, the shapes of the distributions across the models and parameters are generally consistent with those observed for the corresponding TFS for the *windows approach*: the sharing of alleles between individuals is consistently higher for lower values of the N_e of the Out-of-Africa population and **Model 3**, tend to produce steeper curves compared to **Model 1** and **2**. Moreover, additional pulses of admixture (variant **B**) increase the proportion of introgressed alleles at low frequency, a pattern which is displayed by all model scenarios.

It is important to point out an important difference in the SFS plots, which is the presence of a small peak at high frequency of shared alleles, that it is never present in any of the TFS. We hypothesize that this is likely due to sampling bias, and not a real sharing of introgressed alleles. This is likely cause by our site selection method

for determining the informative introgression sites, based on a small samples of only 5 African individuals (see section 2.1.4).

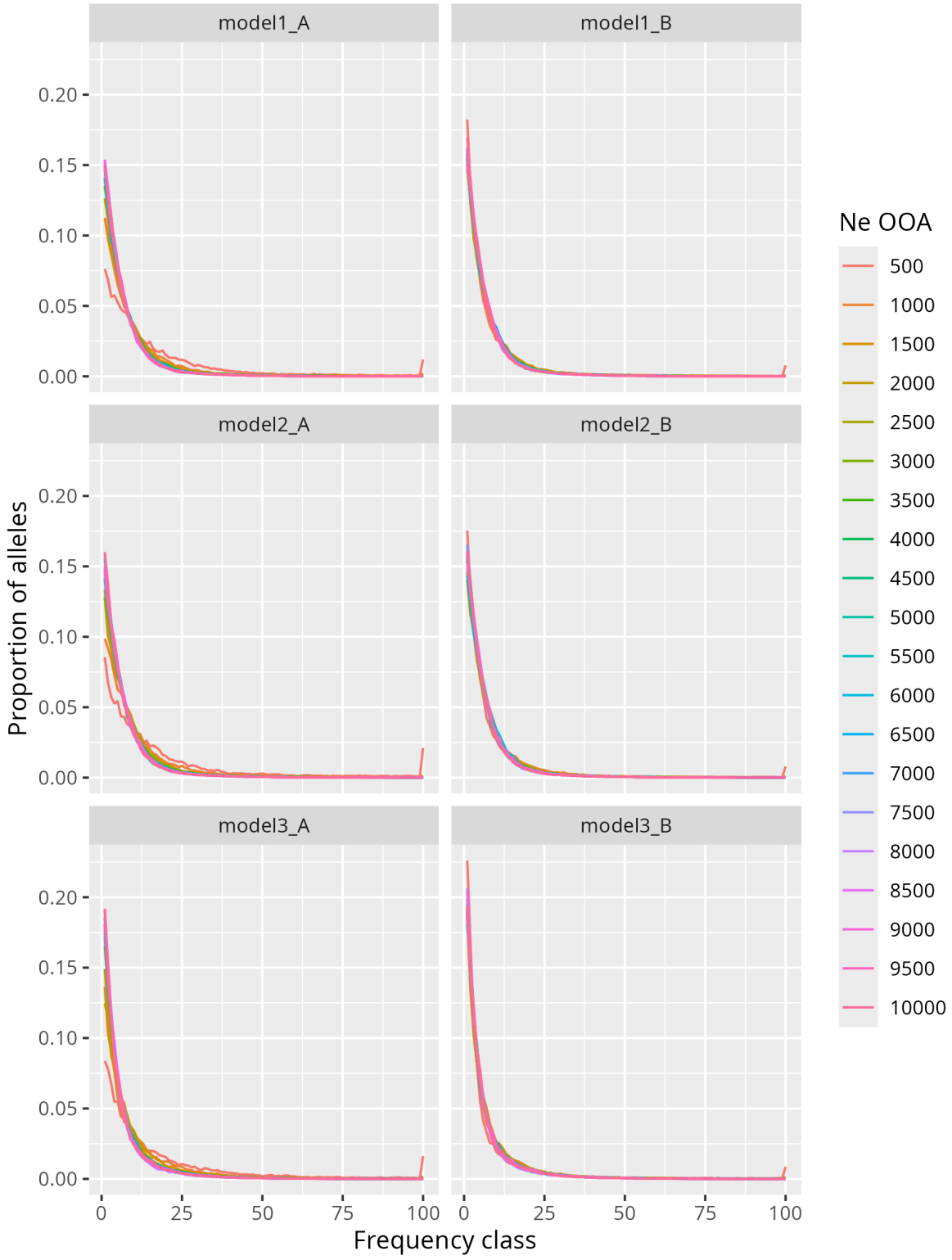

Figure 2.9: Comparison of SFS at introgression informative sites between different models. Each grid corresponds to the SFS of introgressed alleles computed for a certain model. Each color of the curves corresponds to a different value of the N_e of the Out-of-Africa population ("OOA"). Each column represents a different model for an introgression variant, **A** or **B**. The x-axis represents the frequency class of sharing between chromosomes. The y-axis represents the proportion of counts of introgressed alleles for each frequency class.

2.2.6 Graph-based tract sharing data structure

Although potentially promising in terms of information about the demographic history and population processes that influenced the distribution and sharing of introgressed tracts, the TFS data still represents a relatively compressed summary of the genome-wide patterns of tract sharing between individuals. In an attempt to capture even more information about past demographic processes, we explored the possibility of designing a graph-based data structure which would capture the extent of tract sharing between individuals.

The rationale behind this graph-based data structure is that if two tracts share at least one end or start, then they must share a common ancestor. In other terms, two individuals that carry a tract defined by a shared recombination breakpoints have to share a common genealogy for that region of the genome. In fact, considering an infinite sites model [91], we can assume that the same recombination event will not occur twice in the same location of the genome.

Specifically, in the resulting graph the tracts (or alternatively individuals as one tract is carried by an individual) represent the nodes (or vertices) in the graph, while the edges (link in the graph) indicates the sharing of a recombination breakpoint.

The method begins by extracting all the observed tracts in a sampled population. This can be done either by considering all tracts in the population or only those intersecting a genomic location of interest. At this point, to each tract is assigned a unique ID (e.g. T_1, T_2, etc.) and they are stored in a tabular format. For simplicity, we will return to the toy example in Table 2.2, where we will assign a unique tract identifier to each entry of the table (see Table 2.3). In this specific example, since each sampled chromosome carries only one tract, using the tract identifier or the individual name is identical.

chrom	start	end	name	tract_id
chr1	100	150	IND_1	T_1
chr1	100	150	IND_2	T_2
chr1	130	150	IND_3	T_3
chr1	90	170	IND_4	T_4
chr1	90	140	IND_5	T_5
chr1	50	80	IND_6	T_6

Table 2.3: Toy example of a tabular format of introgressed tracts. The table is equivalent to Table 2.2, with a new column *tract_id* that stores the identifier of the tract.

Using the same *recombination breakpoints approach* described in Figure 2.2.1, it is now possible to represent Table 2.3 as a binary matrix M, with each row corresponding to a tract and each column to a recombination site (see Matrix 2.1).

$$M = \begin{array}{c} \\ \text{T\_1} \\ \text{T\_2} \\ \text{T\_3} \\ \text{T\_4} \\ \text{T\_5} \\ \text{T\_6} \end{array} \begin{array}{c} \begin{array}{cccccccc} \text{chr1:50} & \text{chr1:80} & \text{chr1:90} & \text{chr1:100} & \text{chr1:130} & \text{chr1:140} & \text{chr1:150} & \text{chr1:170} \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \tag{2.1}$$

Now, we can compute the adjacency matrix A by multiplying M by its transpose, M^T. In the resulting matrix A (see Equation 2.2), an entry is equal to 1 if two tracts share one end, and equal to 2 if they share both ends (if they are equal). An entry is equal to 0 for tracts that are not link.

$$A = MM^T = \begin{array}{c} \\ \text{T\_1} \\ \text{T\_2} \\ \text{T\_3} \\ \text{T\_4} \\ \text{T\_5} \\ \text{T\_6} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{T\_1} & \text{T\_2} & \text{T\_3} & \text{T\_4} & \text{T\_5} & \text{T\_6} \end{array} \\ \begin{bmatrix} 2 & 2 & 1 & 0 & 0 & 0 \\ 2 & 2 & 1 & 0 & 0 & 0 \\ 1 & 1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix} \end{array} \tag{2.2}$$

Since A is symmetric, we define A' (see Matrix 2.3) as its upper triangular part with the diagonal set to 0, as self-links are not informative in our case.

$$A' = \begin{array}{c} \\ \text{T\_1} \\ \text{T\_2} \\ \text{T\_3} \\ \text{T\_4} \\ \text{T\_5} \\ \text{T\_6} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{T\_1} & \text{T\_2} & \text{T\_3} & \text{T\_4} & \text{T\_5} & \text{T\_6} \end{array} \\ \begin{bmatrix} 0 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \tag{2.3}$$

At this point, we leveraged the property of the adjacency matrix to build a graph.

Figure 2.10 shows an example of graph-based tract sharing data structure for the tracts from the toy example of Table 2.3. A description of how we obtained the graph from the adjacency matrix can be found in section 2.1.6.

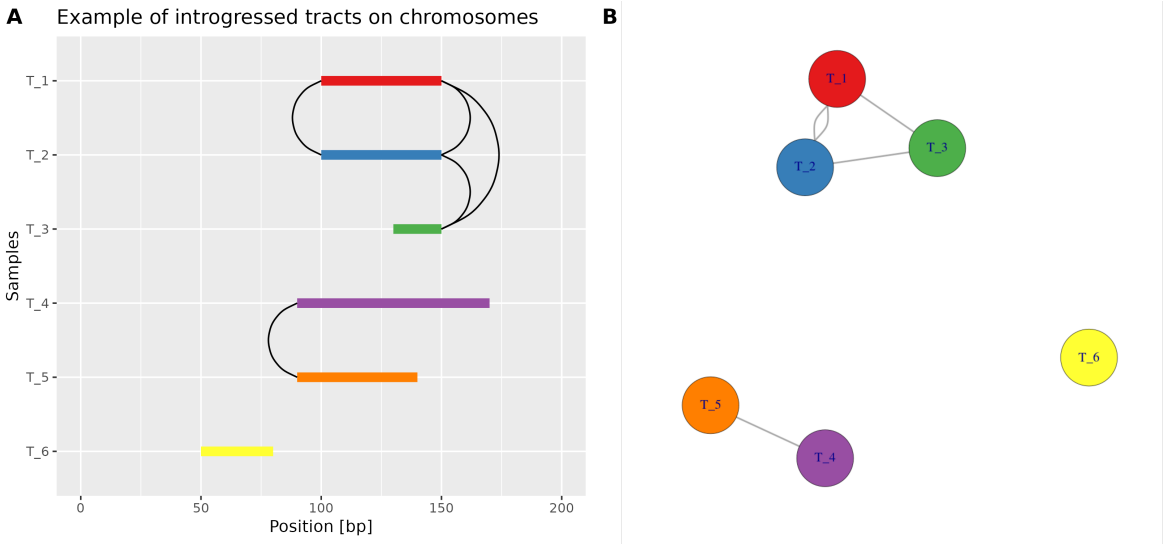

Figure 2.10: Visual example of the tract genealogy graph. (**A**) shows a visual representation of the sampled tracts across sampled chromosomes. The x-axis represents the genomic position, the y-axis the sampled chromosomes. The black curved lines represents the connections between the tracts due to common sharing of recombination breakpoints. (**B**) shows the corresponding graph built from the adjacency matrix 2.3, with the colors matching the tracts in (**A**). T_1 (red) and T_2 (blue) have two edges because they are identical tracts. T_6 (yellow) is not linked to any tract.

2.2.7 Tract-based genealogy graph

As we mentioned in the previous section when we introduced the graph-base data structure, two tracts sharing a recombination breakpoints have to be somehow related as generated by a common recombination event. In other terms, they should share a common genealogy.

To evaluate whether it is indeed possible to qualitatively infer the genealogy of tracts using the graph-base data structure, we simulated a simple scenario of Neanderthal admixture in an Out-of-Africa population (see section 2.2.6 and Figure S8). We sampled 5 diploid individuals from admixed populations at different time points and we identified the introgressed tracts. Figure 2.11 shows the Neanderthal introgressed tracts for the sampled individuals.

To explain the overall concept of the tract-based graph data structure on a simple concrete example, let us consider a locus on a simulated genome at a genomic location 1053331 bp (red dashed line in Figure 2.11). Due to the power of the tree-sequence data structure, we are able to reconstruct the true, known, underlying genealogy between individual haplotypes at this locus. After computing our tract-based graph (see section 2.2.6), we can compared it with the true genealogical tree at the locus.

Figure 2.12 shows the comparison between the true genealogy at the locus at the genomic coordinate 1053331 bp and the inferred tract-based graph. Although we note that reconstructing the exact genealogical relationship from tracts alone to the extent to which the true genealogical tree at a locus can be extracted is likely an impossible task, we can see that the topological structure of the graph does, indeed, reflect some of the features of the true genealogy. Specifically, we can see that the tracts form two distinct clusters (see Figure 2.12), those that do not share any recombination breakpoints with others remain isolated. Note that although two clusters are observed, it is not possible to infer the relationship to each other, and one could be nested into the other.

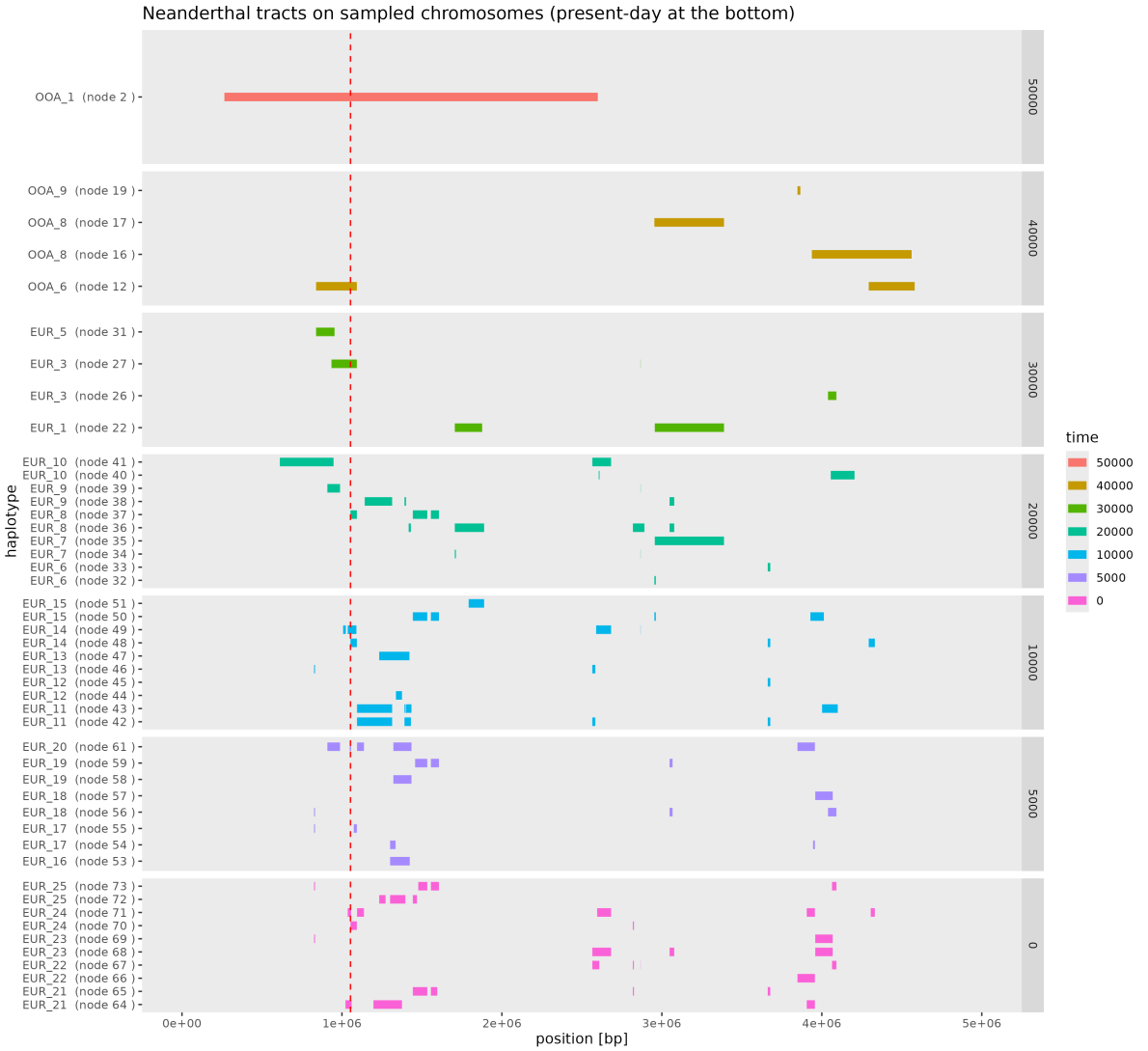

Figure 2.11: Visual representation of Neanderthal introgressed tracts in sampled chromosomes. The panel shows the coordinates of true Neanderthal introgressed tracts in samples from a single simulation run. The tracts are colored by the sampling time of each individual haplotype (top: oldest haplotype, bottom: present-day haplotype). For instance, the haplotype "OOA_1 (node 21)" in the top row indicates a haplotype number 2 (one chromosome of the individual "OOA_1"); haplotypes "EUR_10 (node 40)" and "EUR_10 (node 41)" indicates two haplotypes (i.e., two homologous chromosomes) of individual "EUR_10". The red dashed line marks the location 1053331 bp, at which a true genealogical tree was extracted from a simulated tree sequence, and a tract-based genealogy graph was reconstructed (Figure 2.12). Only haplotypes carrying an introgressed tract at a locus (out of potentially many more samples simulated) are shown in the figure.

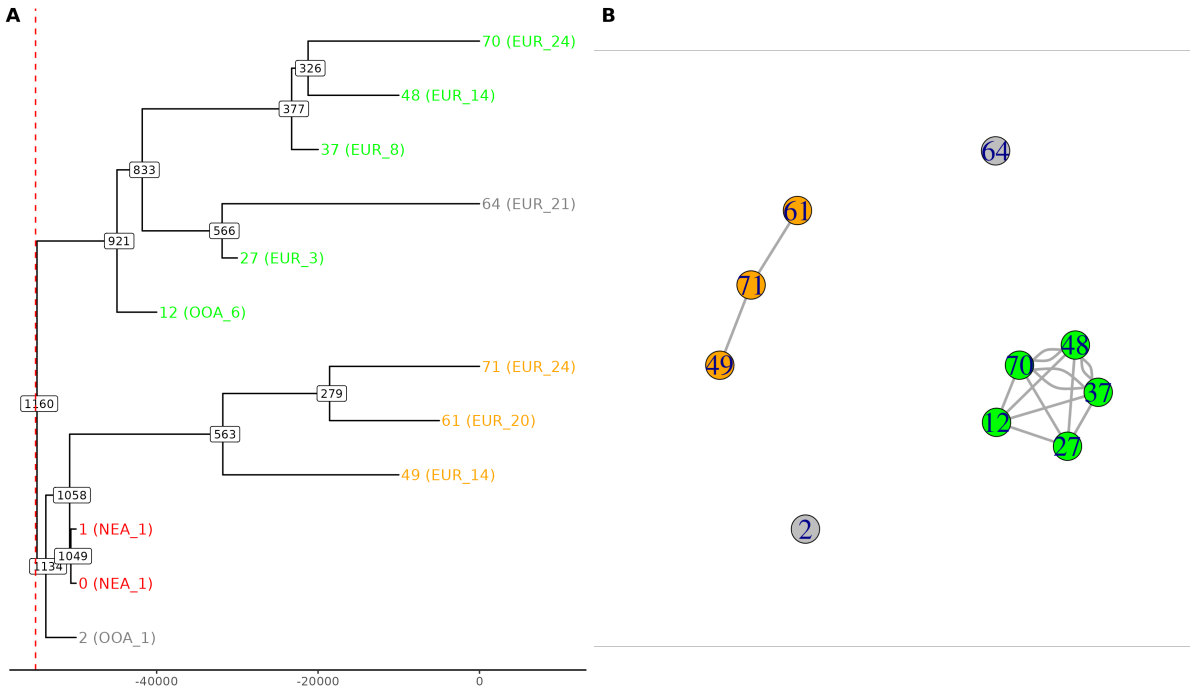

Figure 2.12: Comparison between the true genealogy and the inferred tract genealogy graph. (**A**) shows the true genealogy of the tracts at position 1053331 bp. The x-axis represents time, with the red dashed line corresponding with the beginning of the admixture event (55 kya). Each tip corresponds to a chromosome, labeled with its node ID in the tree-sequence. In parenthesis, it is the sampled individual name. The two red tips correspond to the two chromosomes of a sampled Neanderthal individual. (**B**) shows the corresponding tract genealogy graph. Vertices colors match the ones in (**A**). Vertices linked by two edges indicates that the two tracts are identical. Node 2 and 64 (in grey) do not share any tract recombination breakpoints with other tracts, staying then unconnected. The tracts investigated can be seen in Figure S9(**A**).

2.2.8 Correlation of f_3 and Neanderthal tract-sharing

A previous study [41] investigated whether the patterns of sharing of Neanderthal ancestry between pairs of individuals could be attributed exclusively to the demographic history of these individuals alone. The authors argued that if most of the Neanderthal ancestry of the individuals in their dataset are traced to the same gene-flow event and their distribution after introgression would be only affected by history of splits and migration events within anatomically modern humans, the pattern of pairwise sharing of introgressed Neanderthal tracts would mirror the population demography, as the Neanderthal tracts were subjected to the same demographic processes (and, particularly, genetic drift) as other part of the genome (section S4.3 in [41]). To investigate this hypothesis, the authors calculated the pairwise f_3 statistics between the samples in their study, a statistic which quantifies the degree of shared drift, and compared it with the pairwise correlation of introgressed Neanderthal genomic regions. Figure 2.13 shows the result of their analysis. Based on this result, they concluded that the overwhelming majority of introgressed tracts could be, indeed, traced to a single introgression event from Neanderthals into the ancestors of Eurasians. However, they could not entirely eliminate the possibility that other, perhaps minor, additional introgression events could have happened into some ancient Eurasian lineages, but not others.

In this section, we implemented a simulation-based approach recapitulating the f_3-vs-tracts correlation analysis using tree-sequences as an efficient means of calculation. In doing this, our goal was to develop another introgression-based summary statistic, alongside the TFS and tract-based genealogical graph described above, to be used in a future model fitting work. Additionally, given that [41] have not performed any simulation-based exploration of the expected behavior of their f_3-vs-tracts correlation analysis or characterizing its statistical power to reject the hypothesis of multiple introgression events, we were interested in exploring this statistic under different models of modern human demography and introgression scenarios. Particularly, our results in Figure 2.7 suggest potentially significant differences in how TFS behaves under different methods of binning the genome: how would this factor influence the f_3-vs-tracts correlation results on a purely technical basis?

To evaluate the f_3-vs-tracts correlation statistic, we implemented a set of slightly more complex, yet still sufficiently simple and abstract, demographic models intended to explore the behavior of this summary statistic given various models along the spectrum between a single introgression event into an ancestor of several European lineages and independent introgression events completely private to each lineage. In each scenario

we recorded five present-day individuals from each lineage in a tree-sequence output file, and computed the pairwise metrics of shared drift and correlation of introgressed tracts. The details for this model as well as the exact computation of both sets of statistics are described in section 2.1.7.

As a first case, we considered a scenario with a single extended pulse of introgression into an ancestral lineage followed by its diversification into multiple lineages, corresponding to the baseline scenario representing single-introgression event scenario proposed by [40]. Figure 2.14 shows the comparison between the pairwise f_3 values and the pairwise correlation of introgressed tracts for this model. In agreement with Figure 2.13, the result suggests that when a single admixture event occurred in an ancestral population the pairwise f_3 values closely mirrors the pairwise correlation of Neanderthal tracts. This general pattern is true regardless of which tract-encoding approach is used. However, as shown in the TFS results presented in this chapter (Figure 2.7), each tract-encoding approach produces a different pattern and differs in terms of inferential power. For instance, the recombination breakpoints approach guarantees that if two individuals share an entry in a tract encoding matrix, they are truly genealogically related at a locus. In particular, approaches based on overlapping of genomic regions (*sub-tracts* and *windows*) are able to retain more information about the splits between populations even for more phylogenetically distant lineages. On the other hand, approaches which are more precise (*uniqueness* and *recombination breakpoints*) about recent splits, are less accurate with distant phylogenetic relationships, precisely because a single recombination event at a locus at which two samples at some point in the past shared perfectly overlapping tracts will lead to a loss of the shared tract boundary between their descendants.

As a second case, we considered a scenario with two independent admixture events, each happening in one of two ancestral populations ("POP1" and "POP2") that originated from a common ancestor ("OOA"). Figure 2.15 shows the comparison between the pairwise f_3 values and the pairwise correlation of introgressed tracts obtained from tree sequences for this model. Broadly speaking, the overall pattern that we obtained is similar to Figure 2.14. However, a key difference is that in this demographic scenario, the pairwise correlation values between samples belonging to the same population, or closely related populations, appear to be more homogeneous. The potential reasons for this behavior might be numerous. For instance, the fact that Neanderthal introgression (66 kya in the previous model, 55 kya in this one) happened closer to the sampling time of the sampled individuals, indicating on average longer tracts shared between

them, might be one potential reason. Another, more technical, explanation is that two independent admixture events will generate two completely different sets of tracts. Taking this into account, when computing the binary matrix using our four tract-encoding approaches, the number of rows in each matrix will be on average two times the number of rows obtained in the first case with one single admixture event. Most of the rows corresponding to one admixture event will be zeros for samples that did not experience a given respective recombination event, and vice versa. As a result, when computing the correlation between closely related samples, the numbers of zero bins shared between them will outnumber the number of bins carrying the value 1 from the other recombination event. This might create a certain degree of uncertainty when using the Pearson correlation as a metric of tract sharing correlation between individuals. Other metrics might then be more advisable.

Finally, in the last scenario we considered independent pulses of introgression private to each of the lineages. Of course, given what we have learned about studying Neanderthal introgression patterns in people today [92, 42], this scenario is entirely unrealistic. However, it does carry value in capture the pattern expected from the f_3-vs-tract correlation analysis in the most extreme alternative to the single-pulse model Figure 2.16. **B** shows the f_3-vs-tract correlation heatmap resulting from this model. Interestingly, the strong pattern of correlation observed in the previous two examples is no longer present here. This is unexpected, as since the pulses of introgression are private to each population, we should expect high correlation of sharing of Neanderthal tracts between samples of the same population, and almost nonexistent, or even negative, correlation between two different populations (which can never share a tract recombination breakpoint under the assumption of the infinite sites model). However, correlation within-population is particularly weak for all the tract-encoding approaches, and tract-encoding approaches based on sharing of overlapping genomic regions still show correlation between-populations, even though the introgressed tracts between lineages did not share any drift.

A possible explanation for this surprising pattern can be found again in the timing of the introgression event which, in the case of these simple toy models, is always quite close to the sampling time of each recorded individual at the "present-day" (from 12 to 8 kya). This results in tracts which are, on average, much longer compared to the previous two models, increasing the probability of random overlap of introgressed regions between samples from different populations. Moreover, the short time since introgression might have not allowed sufficient recombination and genetic drift for

tracts to become differentiated within each population, reducing within-population correlation and increasing between population correlation. As suggested in the previous example, multiple independent introgression events may lead to a higher number of recombination breakpoints and generate larger tract-encoding matrices. This could potentially influence the results from a more technical point of view, and thus represents an important lesson to take for future work on more detailed simulation-based inference using the correlation metric presented by [40] as one of the summary statistics.

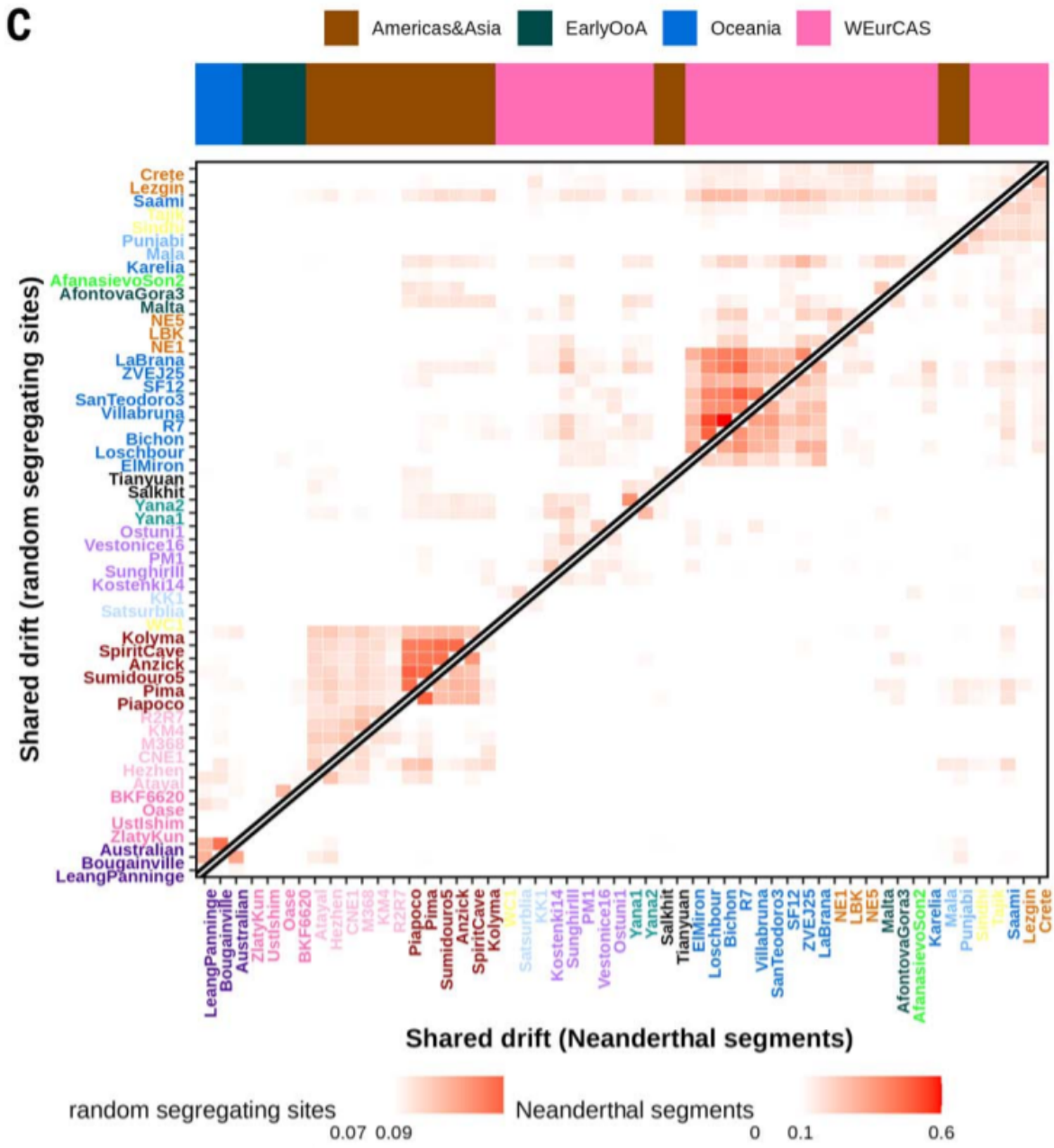

Figure 2.13: Comparison of the differences in pairwise f_3 values and the pairwise correlation of Neanderthal genomic regions. The figure was adapted from [41]. The upper part of the matrix corresponds the pairwise f_3 statistics among individuals in their dataset, while the lower part of the matrix corresponds to the pairwise correlation of introgressed Neanderthal genomic regions. The pattern of f_3 values resembles the pattern of correlation of Neanderthal regions.

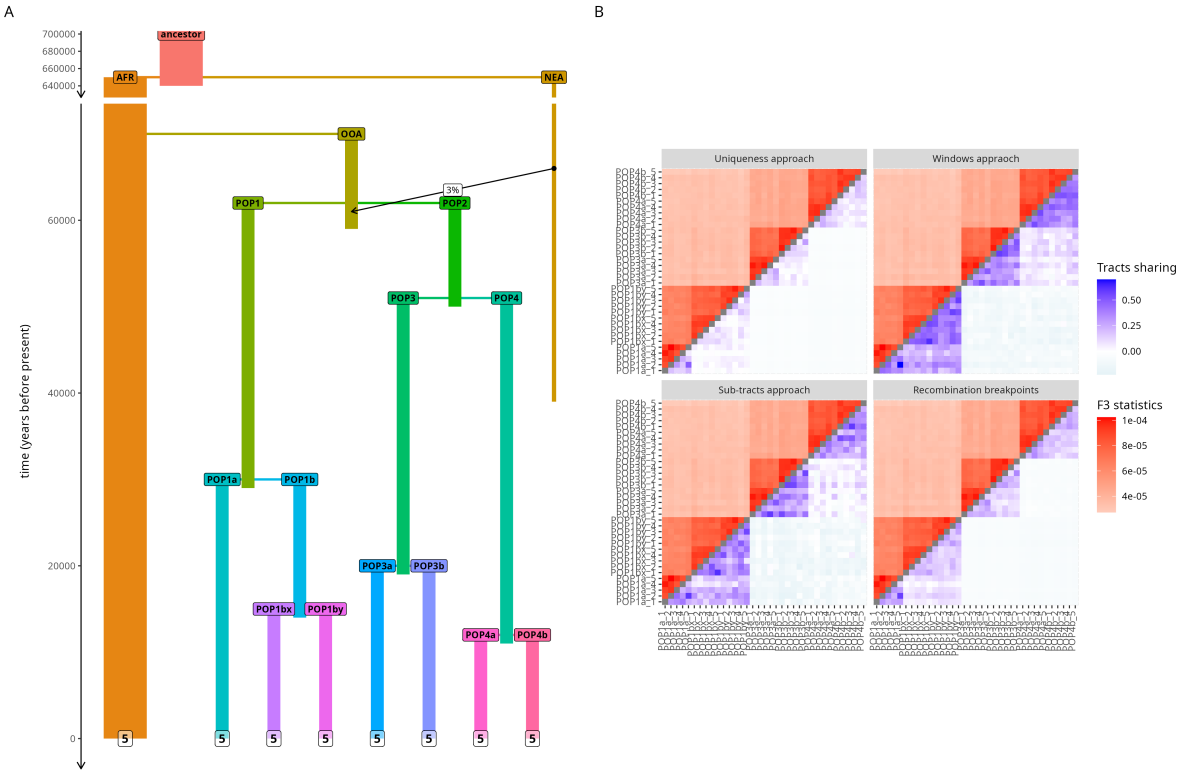

Figure 2.14: Comparison of the differences in pairwise f_3 values and the pairwise correlation of Neanderthal tracts for one pulse of introgression. (**A**) shows a visual representation of the model. (**B**) shows the resulting matrices for each tract-encoding approach. The upper part of the matrices corresponds to the pairwise f_3 values, while the lower part to the pairwise correlation of Neanderthal introgressed tracts. Five individuals were sampled from each population.

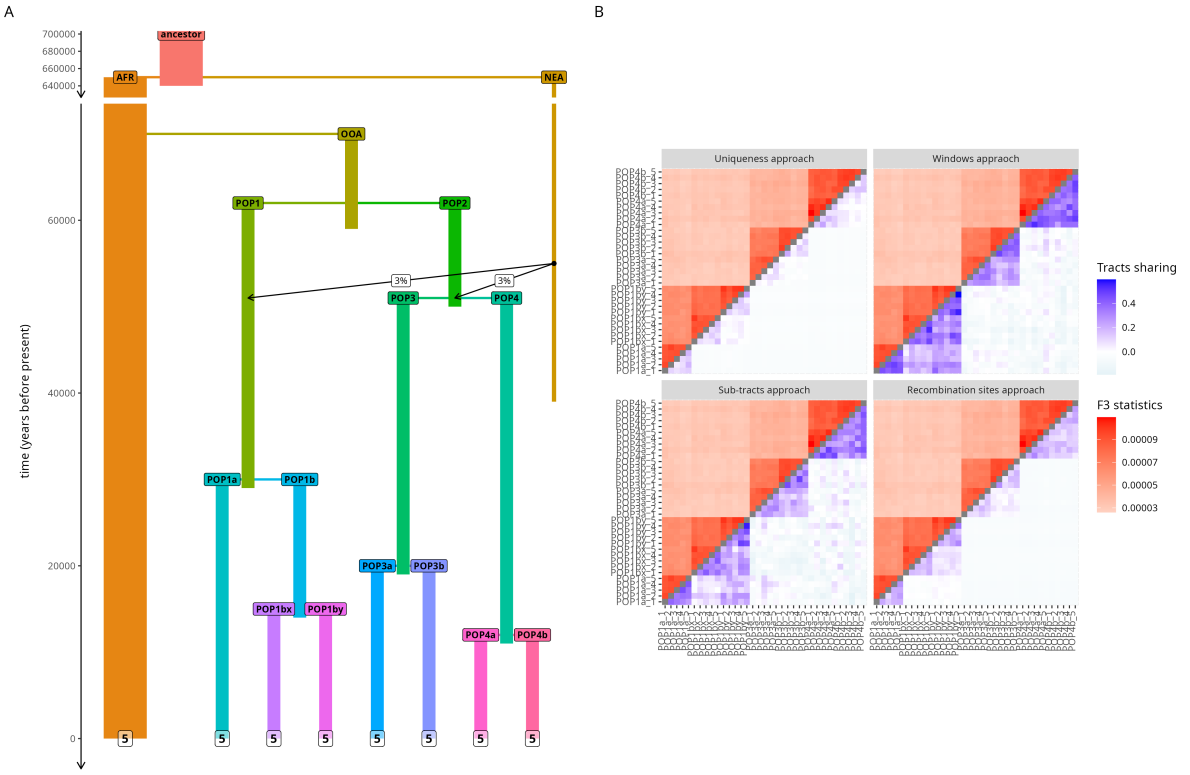

Figure 2.15: Comparison of the differences in pairwise f_3 values and the pairwise correlation of Neanderthal tracts for two pulses of introgression. (**A**) shows a visual representation of the model. (**B**) shows the resulting matrices for each tract-encoding approach. The upper part of the matrices corresponds to the pairwise f_3 values, while the lower part to the pairwise correlation of Neanderthal introgressed tracts. Five individuals were sampled from each population.

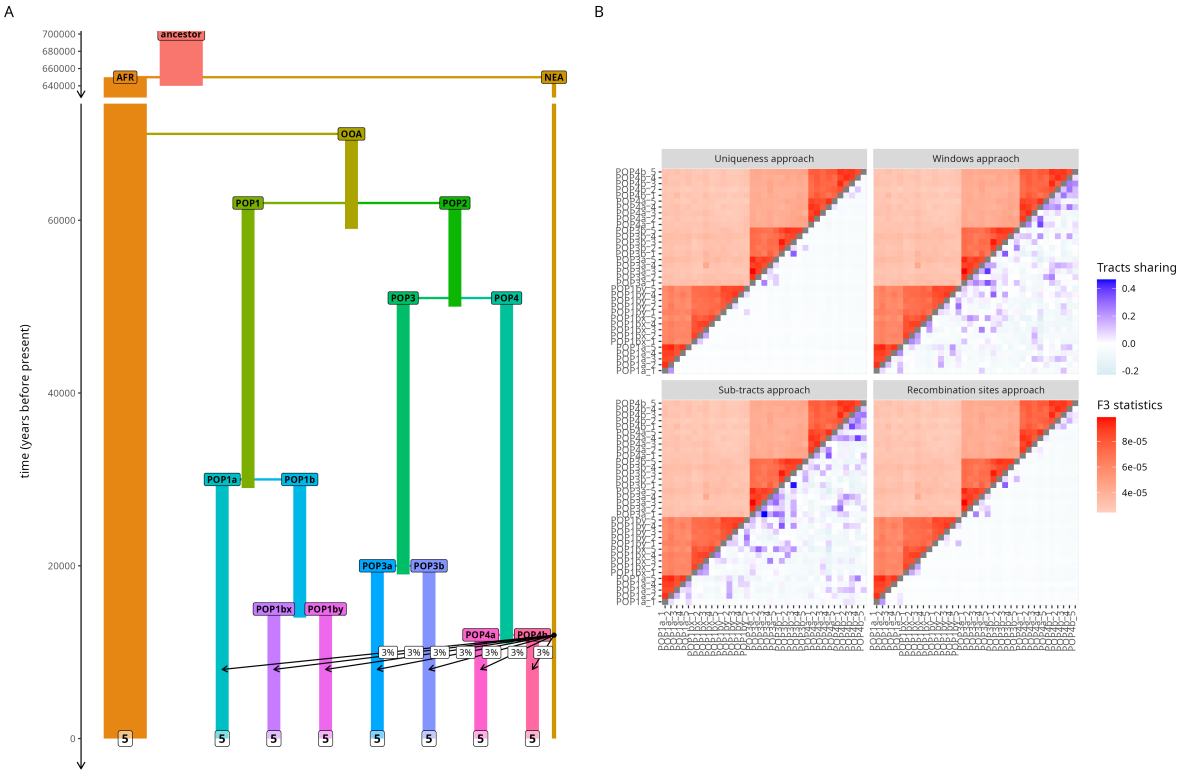

Figure 2.16: Comparison of the differences in pairwise f_3 values and the pairwise correlation of Neanderthal tracts for independent pulses of introgression. (**A**) shows a visual representation of the model. (**B**) shows the resulting matrices for each tract-encoding approach. The upper part of the matrices corresponds to the pairwise f_3 values, while the lower part to the pairwise correlation of Neanderthal introgressed tracts. Five individuals were sampled from each population.

2.3 Discussion

In the second chapter of this study, we developed and explored a set of metrics based on Neanderthal introgression tracts, developed as to study early human demography and dynamics of Neanderthal introgression, applying them both to the models implemented in Chapter 1 and to some other conceptual scenarios. Our primary motivation for creating new summary metrics based on introgressed tracts was that, unlike genome-wide variants, the time at which segments of introgressed Neanderthal DNA appeared on the modern human genetic background is fairly well established at around 49 kya [39]. Consequently, their distribution in present-day non-African people exclusively depend on the population demography of Eurasian populations in the last approximately 50 thousand years. Could the knowledge of the time since which introgressed tracts became influenced by various demographic forces of the earliest modern humans who received them, help us with more accurate inference of population and social dynamics in prehistoric Eurasia?

First, our results revealed that, as expected, the distribution of tract lengths is not a particularly informative metrics that can be helpful in distinguishing the different demographic scenarios that we designed (see Figure 2.6). That said, it did serve a purpose as a useful sanity check of the correctness of the extraction of true tracts using the slendr function *ts_tracts()* and the relatively new underlying Python algorithm *tspop* [85], particularly in scenarios involving multiple introgression events. Given that the distribution of tract lengths is commonly used for dating admixture event [88, 89], its invariance that we observed to the different demographic history of our models suggests that it is indeed a robust summary statistics for dating archaic introgression.

Second, we developed a summary statistics called the *tract frequency spectrum* (TFS), named analogously to the traditionally used site frequency spectrum (SFS), and explored its behavior on model scenarios developed in Chapter 1 through four different ways to encode a tabular format of introgressed tracts into binary matrices (see section 2.2.1). Our results clearly showed that each tract-encoding approach has different informative features (see Figure 2.7). Specifically, we can divide our tract-encoding approaches into two classes: the first class includes the *uniqueness* and *recombination breakpoints approaches*, which are both based on encoding information about the precise locations of the recombination breakpoints; the second class includes the *windows* and *sub-tracts approach*, which are based on encoding information about the genomic regions that are introgressed.

The TFS for the first class of approaches revealed to be invariant to the N_e of the OOA population and to the two introgression variants **A** and **B** (see Figure 2.7), but slightly capable of capturing differences in the demographic history of our models. Specifically, both the approaches belonging to this class produced a steep distribution, with high frequencies of singletons across all the models.

On the other hand, the TFS for the second class of approaches revealed producing significantly different results across different demographic history and model parameters in our simulations. Specifically, the distributions for these two approaches turned out to be affected both by the N_e of the Out-of-Africa (OOA) population, the demographic history of the West Eurasian (EUR) population, and the number of pulses of introgression. In particular, the OOA population which underwent a strong bottleneck always led to flatter distributions with more frequent sharing of tracts in the sampled population (see Figure 2.7). This is not surprising, as smaller populations experience more drift, which make certain introgressed regions reach higher frequencies. Additionally, **Model 3** which exhibits an extended past of population structured produced steeper TFS distributions, likely because of the merging of independent recombination histories. Interestingly, we observed that the TFS distribution of the *windows approach* was always steeper than the one of the *sub-tracts approach*. Potentially, this difference is traceable to the definition of the genomic bins among the two approaches. Specifically, a much finer binning for the *sub-tracts approach* might lead to a higher proportion of shared counts (not singletons). However, this hypothesis still needs to be evaluated. For this class of approaches, our results showed that the variant **B** of introgression (see Figure 2.7), involving two pulses of Neanderthal introgression, always resulted in a higher proportions of low frequency class of sharing. This is likely due to the introduction of a new set of tracts in the population at particularly low frequencies.

Although the different patterns captured by the TFS among the several model scenarios might be indeed produced by the differences in demography and population structure, we need to be cautious at drawing our conclusions at this stage. In fact, although all the models all featured the final EUR population (i.e., "present-day Europeans") with $N_e = 15000$, the demographic scenarios differ in how and when they reach this value. For instance, in **Model 1** the EUR population reaches $N_e = 15000$ at 42 kya experiencing a grow that depends on the parametrized N_e of the OOA population. **Model 2** reaches it at 3 kya growing from a population with $N_e = 5000$ to $N_e = 15000$, while **Model 3** reaches it at at 3 kya growing from a population with $N_e = 2000$ and the final EUR population experienced admixture with the restant EUR demes. The different

growth rates and ways that these populations experienced admixture are very likely to have a major influence on the distribution of the TFS rather than the structured history of the population itself. This is something that needs to be investigated in the future.

Our complementary analysis based on the site frequency spectrum (SFS) at introgression informative sites (see Figure 2.9) were broadly consistent with the ones for the TFS (see Figure 2.7). In particular, the shape of the SFS of introgressed alleles was more similar to the TFS obtained for the *windows approach*. A primary difference that we noticed in the SFS results is the presence of a small peak at high frequency of shared alleles, that it is never present in any of the TFS. We hypothesized that this behaviour is likely produced by a form of a sampling bias due to the fact that our sample of African genome used to establish the introgression informative sites comprised only 5 individuals. This increases the probability of misclassifying sites as private derived to Neanderthals. This general consistency suggests that the TFS might be a summary statistics well-suited for studying introgression

The fact that two independent admixture events generate two completely different cascades of recombination events and, therefore, tracts that share an recombination breakpoint must have a common origin under an infinite sites model [91] is a crucial detail that appears to carry a significant amount of evolutionary information. The graph-based tract sharing data structure (see Figure 2.10) that we developed in this chapter is capable of exploiting this powerful feature with promising potential for modeling the prehistory of early modern humans and the dynamics of Neanderthal introgression. Moreover, in the future, this graph-based data structure could be expanded to incorporate more quantitative details about individuals represented by vertices of the graph [93]. In fact, since real populations evolve in time and space, we could encode temporal and geographical distances in the edges that connect the nodes (individuals) of the graph.

To this end, recombination breakpoints might be used as a novel marker not only to genealogically link individuals, but to also infer introgression dynamics and and demographic parameters of the prehistoric populations. In fact, although it is commonly assumed that recombination breakpoints are generally not directly observable because they do not result in a observable changes in the sequence of DNA basepairs [94], the presence of divergent introgressed material in a population, such as the much more clearly delineated Neanderthal DNA, might make it easier to identify these recombination breakpoints. Taking this into account, it could be possible to try to

model the prehistory of early modern humans using the number of recombination events that have occured in the population [95].

Recently, Iasi *et al.* [40] developed a method which compares the pairwise f_3 statistics with the pairwise correlation of Neanderthal tracts (see Figure 2.13) and, based on their results, concluded that the set of introgressed tracts observed today derived from a single extended pulse of admixture with Neanderthals. As the authors did not formally test this hypothesis (or, for that matter, the power of this statistic to reject alternative scenarios), we implemented their analytical framework under different introgression scenarios, and included it in our simulation workflow. In doing so, we encountered several limitations of this summary statistic that will need to be taken into account in future work. On one hand, our results showed that a scenario involving a single pulse of Neanderthal introgression, consistently to what proposed by Iasi *et al.*, results in a pattern comparable to theirs (see Figure 2.14), regardless of an exact tract-encoding approach used (see section 2.2.1). On the other hand, we observed quite a similar result for an alternative scenario with two independent pulses of Neanderthal introgression (see Figure 2.15). Moreover, for the scenario involving multiple independent pulses of Neanderthal introgression, we obtained a resulting correlation matrix showing almost no correlation between individual sampled within the same population, independently on the exact tract-encoding approach used (see Figure 2.16).

It is important to note that as several factors might play a role in shaping the observed patterns. For instance, an important aspect to consider is the timing of the introgression. Populations which received Neanderthal DNA during independent introgression events might still share a significative amount of overlapping introgressed regions, specifically if sampled close to the admixture date. This is inherently due to the fact that tracts sampled closer to the admixture event tend to be longer,and, consequently, the probability of observing overlapping tracts increases. This factor might provide an explanation for why in Figure 2.16, despite the fact that all the populations have experienced private introgression events, individuals sampled from different populations are correlated. As a technical point, we note that using the traditional Pearson correlation might not be appropriated in this case and other metrics specifically designed for binary vectors might be more advisable [96]. Moreover, when comparing two individuals, the fact that they share a row in a tract-encoding binary matrix, or that neither share it, might have a different weight.

Chapter 3: Exploratory applications to empirical data

3

In the previous two chapters, we have developed a set of models of human prehistory as proof-of-concept alternatives to traditionally single-population panmictic models used in the literature and implemented a number of summary statistics based on information encoded in introgressed tracts. In doing so, our overall goal has been to develop draft implementations of components of a future simulation-based inference pipeline, including Approximate Bayesian Computation (ABC) [97], which will attempt to infer the parameters of prehistoric population and social dynamics beyond standard panmictic population genetic models.

Although implementing this inference pipeline in full is beyond the scope of this thesis, we have dedicated the third chapter of this study to an exploratory application of some of the tract-based summary statistics and concepts developed in Chapter 2 to a data set of Neanderthal introgressed tracts in a panel of ancient and present-day modern human genomes (see section 3.1.1) as inferred in a previous project in our group (Refoyo-Martínez, *et al.*, in prep.) using the IBDmix software [52]. In doing so, our primary motivation was to evaluate the summary statistics presented in Chapter 2 in a real-world setting, because unlike the perfect simulated data encoding the ground truth without error, statistical inference on real data involves uncertainty and error and often involves filtering steps that will necessarily change some of its statistical properties. To this end, and to achieve a more fair comparison with the results obtained from the simulations we modified the simulated data accordingly to match the filtering performed on empirical data (see section 3.1.2). Therefore, all results shown in this chapter are those between empirical data and the processed data from the simulated scenarios (see section 1.2.1).

First, we assessed the degree to which the results of our tract-based summary statistics on empirical present-day data are consistent with the results we obtained on our simulated scenarios (see section 2.2.4). In particular, we computed and compared the distribution of introgressed tract lengths, TFS, and SFS of introgressed alleles for present-day West Eurasians to those derived from the simulations. Again, as noted in the previous chapter, our alternative demographic scenarios of prehistoric structure have not been designed based on rigorous statistical modeling. Nevertheless, our

intention was to assess in which way might more complex models fit some aspects of the empirical data better compared to a single-population panmictic model.

Finally, we demonstrate how the tract genealogy graph structure (which by definition carries spatio-temporal information about relationships between individuals) could be applied and visualized in a geographical context. As a conceptual demonstration, we computed the graph for a selected introgressed genomic region shared by several ancient samples and visualized it on the geographical map.

3.1 Materials and Methods

3.1.1 Neanderthal tracts inference in modern human genomes

In order to compare the tract-based summary statistics derived from our models with empirical data from modern human genomes, we used a large panel of both present-day [98] and ancient imputed genomes [26] where the coordinates of the introgressed Neanderthal tracts were inferred using the IBDmix software [52] through a pipeline developed by our group (Refoyo-Martínez, *et al.*, in prep.). Briefly, the authors of this prior work, identified Neanderthal tracts using the Vindija Neanderthal genome as a reference [79]. Only tracts longer than 50 kilobases (kb) and with a logarithm of the odds (LOD) score greater than 4 were kept, to minimize the number of false positives. Moreover, regions that were called as Denisovan sequences in at least one African sample and are also present as Neanderthal-like sequences in any other non-African individuals were filtered out because potential candidate of incomplete lineage sorting (ILS). Although the genomes used for predicting the tracts were phased, IBDmix returns tracts at the individual rather than at the chromosome level.

3.1.2 Comparison between simulations and empirical data

When working with empirical data, the detection of Neanderthal introgressed tracts cannot be as precise as in the simulations. In particular, the IBDmix software [52] used to indentify Neanderthal introgressed tracts in modern human genomes presents some limitations. This method outputs only one set of introgressed tracts per individual, regardless if the data were phased. Moreover, tracts shorter than 50 kb were removed to minimize the number of false positives.

To account for these characteristics of the data when comparing empirical against simulated data, we applied similar filtering to the tracts extracted from our simulations to achieve a more fair comparison. Specifically, for each sampled individual we merged overlapping tracts between the two homologous chromosomes (which are, by definition, "phased" in the simulations) using the *reduce()* function of the *GenomicRanges* R package, thus approximating the non-phased aspect of tracts inferred by IBDmix. Additionally, we removed tracts shorter than 50 kb.

To compare the distributions of tract lengths we proceed as described in section 2.1.3. However, because tracts shorter than 50 kb were not present, we defined bins of 20 kb of widths, ranging from 50 kb to 7 Mb. The empirical distribution corresponds to the tract lengths identified in the full dataset of present-day West Eurasians (n=503).

Since IBDmix outputs one only set of tracts per individual, when comparing the TFS between empirical and simulated data, the length of the TFS is equivalent to the number of sampled individuals N, rather than $2N$. This does not apply to the SFS of introgressed alleles, as we have accessed to both the homologous chromosomes.

When computing SFS and TFS, we sampled 50 out of 503 West Eurasians individuals, repeating the sampling independently 10 times. We then computed the mean and standard deviation across the 10 sampling.

3.1.3 Tract genealogy graph in space

To qualitatively explore the tract genealogy graph described in section 2.2.6 in a geographical setting, we filtered the dataset to include only West Eurasians individuals dated between 10 and 12 kya. This subset includes 12 ancient modern human samples.

For these samples, we encoded their introgressed tracts using the *sub-tracts approach* and we selected the genomic region on chromosome 4 at position 35491388 bp because identified as of Neanderthal origin in 11 of the 12 samples. We then extracted the tracts from these individuals intersecting this region and we computed our tract genealogy graph as described in section 2.2.6.

To visualize the graph on a geographical map we used the *sf* [99] and *sfnetworks* [100] R packages.

3.1.4 Code availability

The source code used for performing the work in this project, including the dataset with the IBDmix predictions of Neanderthal tracts, can be found in a publicly accessible GitHub repository (https://github.com/fil-tel/MSc-thesis).

3.2 Results

3.2.1 Distribution of introgressed tract lengths in present-day West Eurasians

In a first step, we computed the distribution of tract lengths in present-day West Eurasians and we compared it against the simulated distributions from our three demographic model scenarios (see section 1.2.1).

In Figure 3.1, we can see that the distribution corresponding to empirical data shows a higher density for shorter tracts compared to the simulations. This is slightly surprising, because on an intuitive level, we would likely expect the IBDmix software to have more difficulty to infer the shortest fragments. A detailed analysis of this would require a more detailed investigation which we will attempt in future work. Some potential explanations are that this could be a technical artifact of the IBDmix inference procedure, an incorrect date of Neanderthal introgression at 49 kya used in the simulations based on the literature [39], or a non-uniform recombination rate in real genomes.

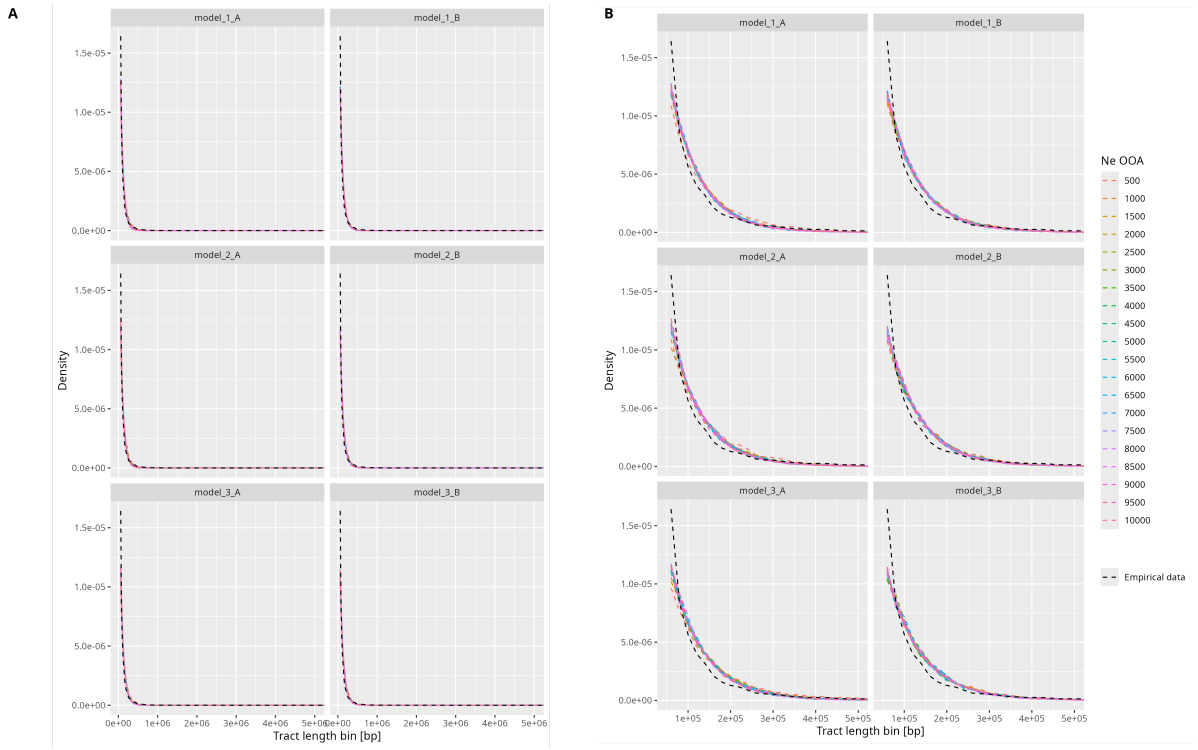

Figure 3.1: Comparison of the distribution of tract lengths between empirical data and our models. (**A**) shows the full distribution of tract lengths between empirical data and our models. (**B**) corresponds to a zoomed version of **A**, where the x-axis has been limited between 50 and 500 kb. Each grid corresponds to the distribution of introgressed tract lengths for a certain model and variant. Each color of the curves corresponds to a different value of the N_e of the Out-of-Africa population ("OOA"). The black dashed line corresponds to the distribution of tract lengths of tracts inferred in 503 present-day West Eurasians. The x-axis represents the lengths of tracts in basepairs (bp), while the y-axis represents the density. Note that tracts shorter than 50 kb have been removed for both empirical and simulated data.

3.2.2 Tract frequency spectrum for present-day West Eurasians

To evaluate whether the patterns of sharing of Neanderthal introgressed tracts between present-day West Eurasian individuals align (even just qualitatively) with any of the prehistoric scenarios of the Out-of-Africa expansion that we simulated in Chapter 2, we computed the TFS (see section 2.2.2) for introgressed tracts inferred in a sample of 50 present-day West Eurasian individuals in the empirical data and compared the results with the equivalent set of simulation results.

Figure 3.2 and 3.3 show the comparison of the TFS between empirical and simulated data for each of the four tract-encoding approaches developed in Chapter 2 (see section 2.2.1). We can see that the empirical TFS shows a steep distribution across all the tract-encoding approaches, with a distinctive peak in the singleton frequency class. For a visualization of the empirical TFS alone, please refer to Figure S10.

When comparing the empirical TFS to the TFS obtained for simulated models, it is clear that all the scenarios appear to be quite incompatible with the empirical data across all four tract-encoding approaches. However, **Model 3** show a qualitatively much better alignment. In particular, the variant **B** of **Model 3** is the most similar. This pattern might suggest that more complex models with population structure and migration may be better in capturing the complexity of past modern human demography and Neanderthal introgression compared to idealized panmictic models, like those exemplified by **Model 1**. Of course, we need to note again that all of our models are proof-of-concept idealized scenarios and have not been inferred by rigorous model fitting procedures. As such, other models could fit equally well. Similarly, the TFS represent only one of many potentially useful summary statistics which should be used in a future model estimating step. Indeed, as our results in Chapter 1 show (see Figure 1.4), even **Model 3B** fails to capture important patterns of present-day heterozygosity.

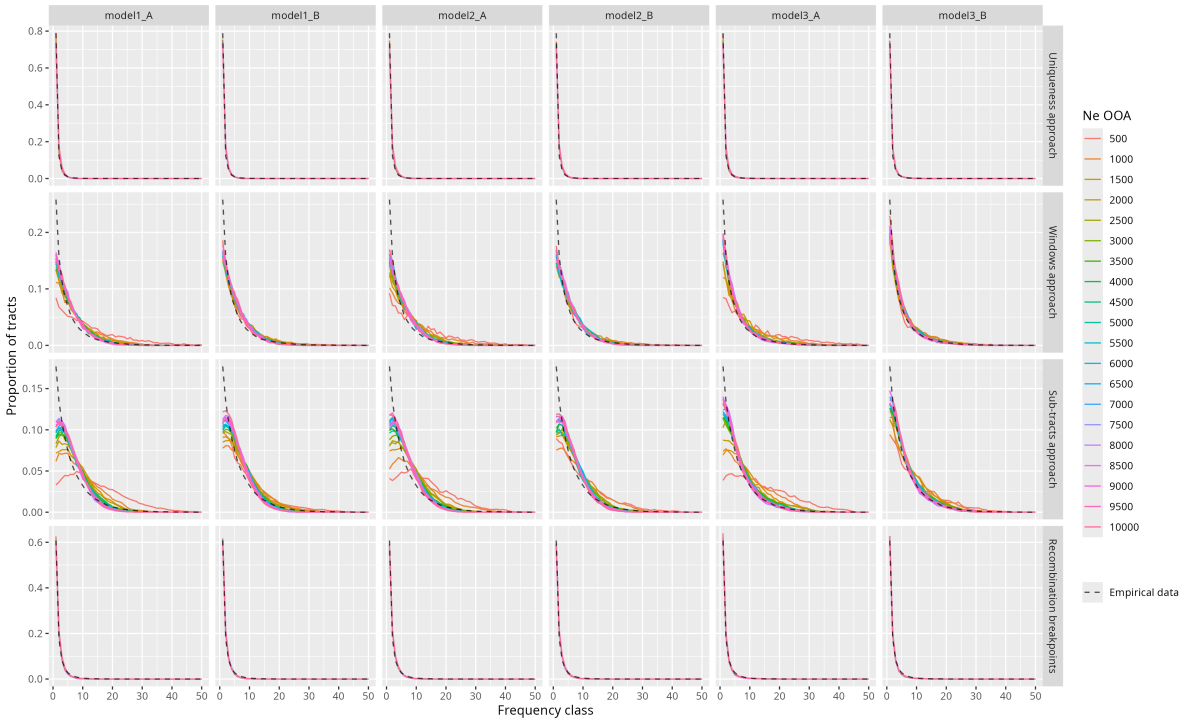

Figure 3.2: Comparison of the TFS between empirical data and models for each tract-encoding approach colored by the N_e OOA. The panel shows a comparison between the TFS of 50 present-day West Eurasians and our models. The TFS shown corresponds to the average across 10 runs of simulations. Each column corresponds to a different model and each row to a different tract-encoding approach. The different colors correspond to different N_e of the Out-of-Africa population, while the black dashed line represents the TFS for empirical data.

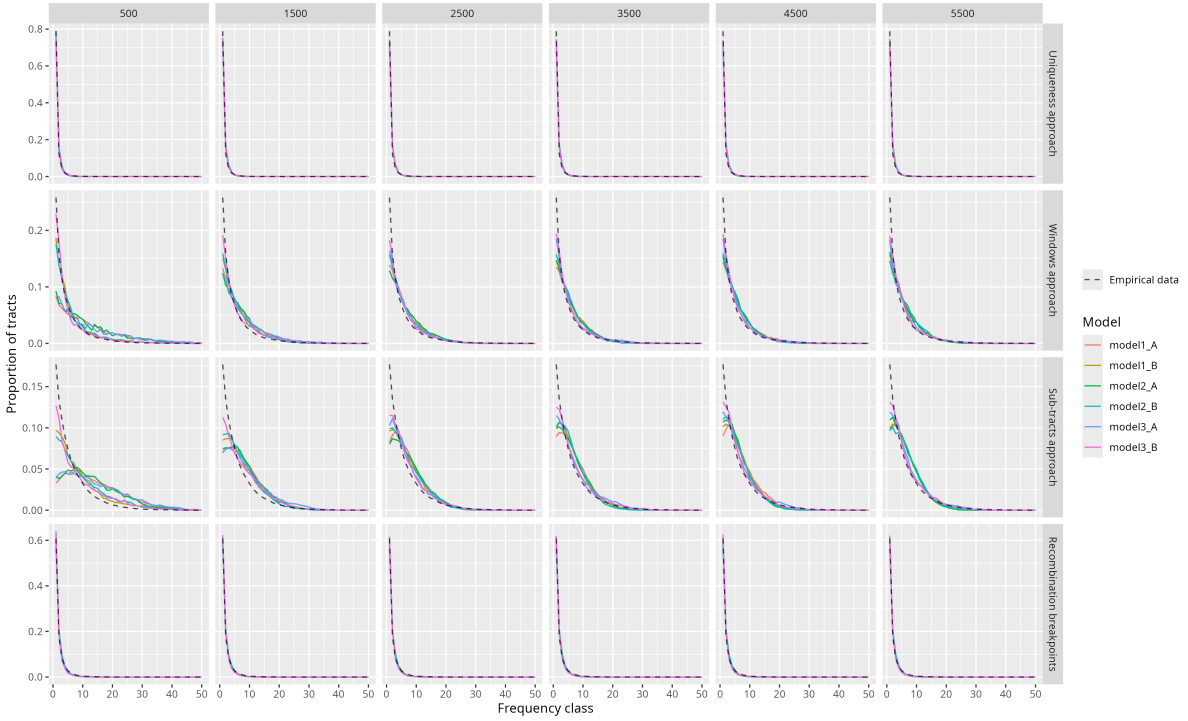

Figure 3.3: Comparison of the TFS between empirical data and models for each tract-encoding approach colored by Model. The panel shows a comparison between the TFS of 50 present-day West Eurasians and our models. The TFS shown corresponds to the average across 10 runs of simulations. Each column corresponds to a different N_e of the Out-of-Africa population and each row to a different tract-encoding approach. The TFS are colored according to the model they represent. The black dashed line represents the TFS for empirical data.

3.2.3 Site frequency spectrum of introgressed alleles for present-day West Eurasians

In this section we performed a complementary analysis to section 2.2.5, computing the SFS of introgression informative alleles on empirical data. InFigure 3.4, we can see that, similarly to the TFS results presented in Figure 3.2, **Models 3** matches the empirical TFS distributions more closely than **Model 1** and **2**. Specifically, the variant **B** for **Model 3** appears to provide the closest match to the empirical distribution. We note that unlike the simulated distributions, the SFS of introgression informative alleles for empirical data do not show any peak at high frequencies (which we earlier speculated might be the result of incomplete lineage sorting).

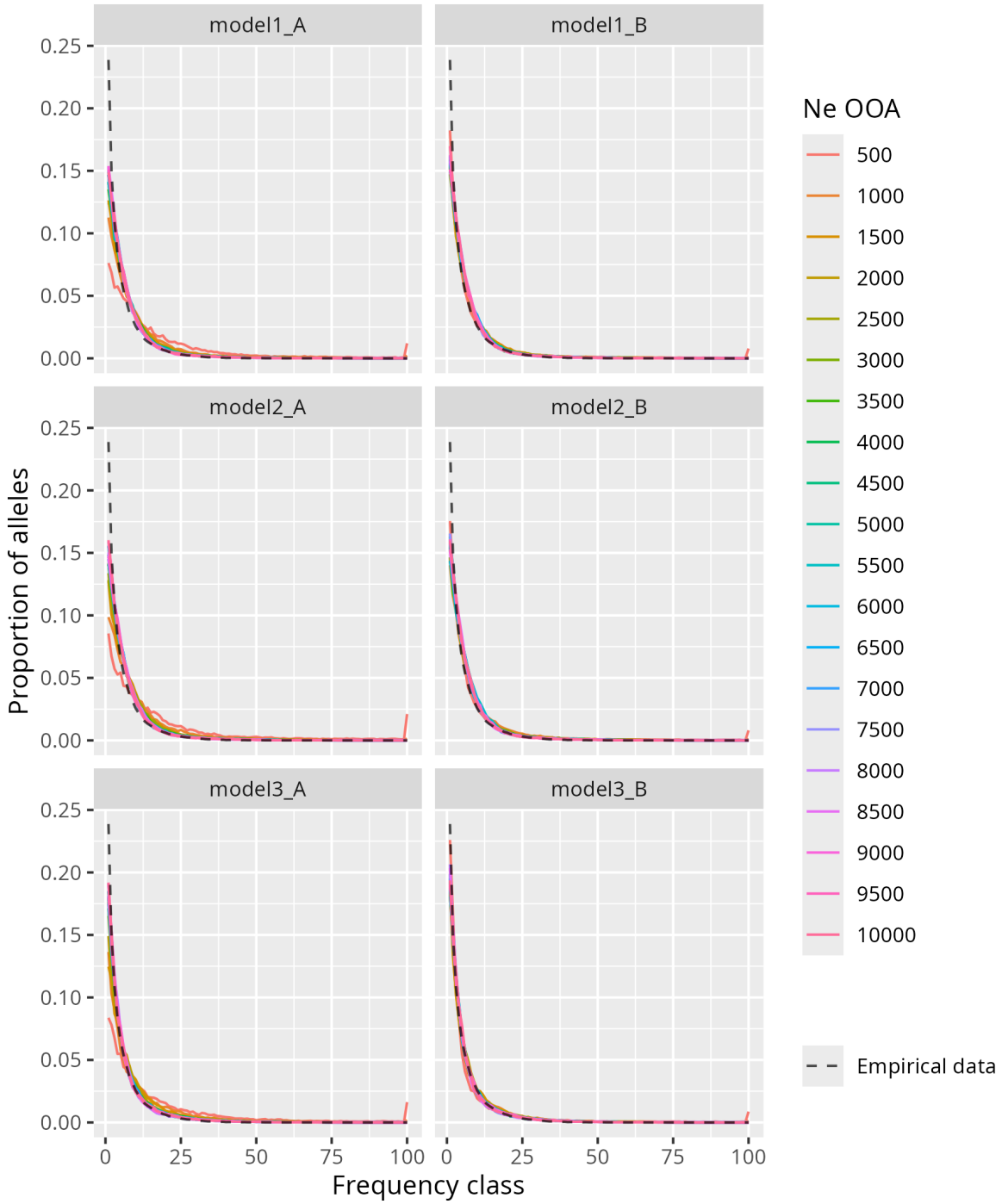

Figure 3.4: Comparison of the SFS of introgression informative sites between empirical data and simulations. The panel compares the SFS of introgressed alleles in 50 present-day West Eurasians with those simulated from our models. For simulations, each distribution is averaged over 10 independent runs, while the empirical distribution is averaged over 10 independent sampling. Each grid cell corresponds to a different model, with the distributions colored according to the N_e of the Out-of-Africa population. The black dashed line indicates the empirical distribution.

3.2.4 Spatial graph of tract sharing

Real populations evolve and migrate within a geographical context. Following a project focused on spatial Identity-by-Descent graph structures, currently under development in our group, we explored the potential of our tract-based genealogy graph as a means to visualize the relationships between individuals across space and time. Could this represent a useful tool to visualize the migration of people, and, eventually, also contribute another introgression-based metric for future simulation-based inference?

We investigated the tract genealogy of tracts intersecting the position 35491388 bp on chromosome 4 in a set of 12 ancient West Eurasian samples dated approximately between 10 and 12 kya, as described in section 3.1.3.

Figure 3.5 shows the map of West Eurasia with the locations of the sampled individuals, with the resulting tract genealogy graph overlaid on top of this map to show the inferred genealogical relationships in a geographical context. Two clusters are visible: one linking individuals sampled in eastern Europe, and one linking southern and northern European individuals. Three out of twelve of the individuals do not belong to any cluster. The only European farmer (pink) sample not connected to any cluster is the only individual that does not carry an inferred Neanderthal introgressed tract in that genomic location.

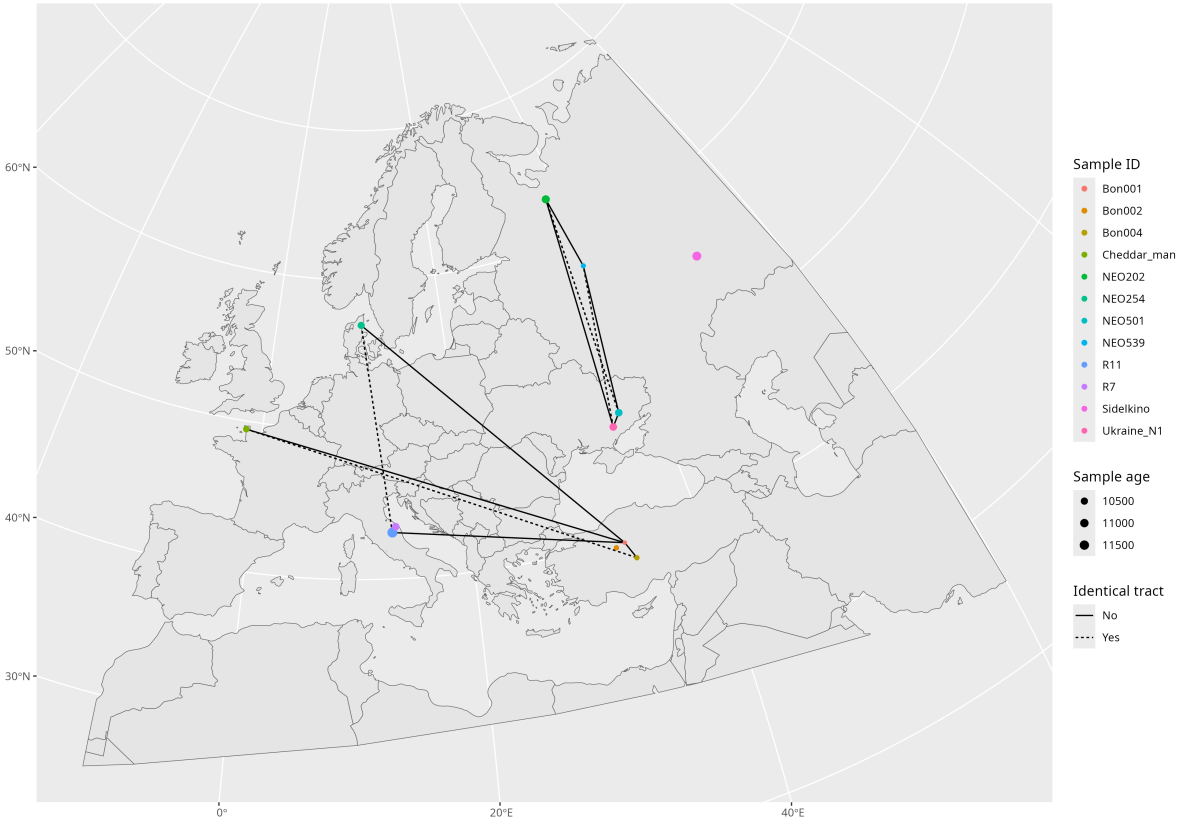

Figure 3.5: Map showing genealogical relationships between tracts in a geographical context. Map of West Eurasia showing as colored dots the sampling locations of 12 ancient modern human samples dated between 10 and 12 kya. The dots are colored according to the sample ID of the individuals. The size of the tracts correspond to the sample age. The lines connecting the dots correspond to shared recombination breakpoints between the tracts. Dashed lines indicate that the two connected individuals share the same tract.

3.3 Discussion

In the last chapter of this study, we applied several of the tract-based summary statistics introduced in Chapter 2 to the empirical data, where the coordinates of introgressed tracts were previously inferred in our group (see section 3.1.1). Despite not being able to perform full-scale simulation-based inference in this thesis, we were motivated by evaluating, on a qualitative level, how the results obtained in the previous chapter for the different demographic models of the prehistory of early modern humans compares to empirical data, and identify potential caveats and limitations of our summary statistics. As expected, we found that the distributions obtained from empirical data were quite different the proof-of-concept models we designed in Chapter 1 (see Figure 3.1, 3.2, 3.4). Nevertheless, these comparisons did provide the opportunities for several interesting reflections.

For instance, when considering the distribution of tract lengths which section 2.2.3 showed to be not particularly informative about the demographic history of the simulated populations, the empirical distribution revealed to be steeper than the distributions obtained from the simulations. This suggests that there is a higher density of shorter tracts compared to the models. The reasons for this might be numerous, but as already mentioned, given that the distribution of tract lengths is commonly used to date admixture events [88], this could mean that the introgression date we used in our simulations (introgression between 49 to 45 kya [39]) is slightly off compared to reality, and dates inferred by other studies might be more correct (from 55 to 50 kya [79]).

The results obtained for the TFS and SFS of introgression informative sites, which in Chapter 2 we concluded are summaries which are quite informative about the demographic history of a population, indicate that introducing a certain degree of population structure in the prehistoric population (**Model 2** and in particular **Model 3**) lead to a better alignment between simulated and empirical distributions. This observation might suggest that indeed, more complex demographic models of the history of the West Eurasian hunter-gatherer populations might indeed be closer to the truth. Of course, more confident conclusion will require much more careful modeling, which we intend to do in follow up work. On a similar note, reality is always much more complex than any model, involving many more factors that we did not consider. For example, genomes are not subjected to homogeneous recombination rate, and the presence of hot-spots might have an influence on the results. Moreover, it is believed that as the Neanderthal DNA entered the modern human genomic background have

been subjected to both positive and negative selection [42, 45]. This might lead certain regions to be more shared than others, and how this would look in our TFS have to be investigated. Additionally, although our ultimate aim is to describe population and social dynamics in more interpretable terms, such as moving from the traditionally used concept of N_e towards real census sizes, for computational efficiency, we ran all our models with the coalescent *msprime* back-end simulation engine of *slendr*. In a follow-up work, we intend to expand our models to use the forward-time, individual-based engine of *slendr* written in SLiM [101]. This will not only allow the simulation of individuals as discrete "agents", but also specify the parameters of social dynamics within each deme (or, a "tribe"). Similarly, these full-scale *slendr* simulations would also allow spatio-temporal modeling of population and social dynamics on real geographic landscapes (see Figure 3.6)[77].

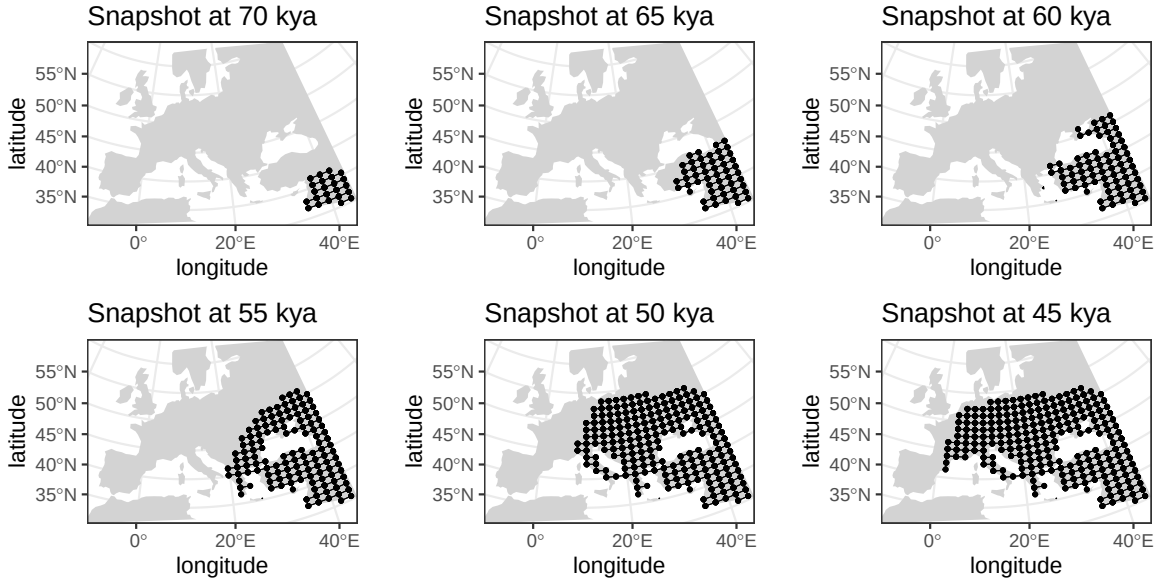

Figure 3.6: A visualization of a hypothetical spatially explicit *slendr* model of a prehistoric expansion of early modern humans into Eurasia. Each point on the map represents a deme (a "tribe" of early modern humans) connected via continuous gene flow of a certain rate with neighboring demes. As modern humans expand into Eurasia (represented by snapshots of models in a temporal sequence of individual panels), new demes are being created on the frontier of expansion. Introgression with Neanderthals could be specified either uniformly at each deme coordinate, or heterogeneously, depending on the assumed geographical distribution of Neanderthals. Adapted from Petr et al. (in prep.).

Another essential aspect to consider is the nature of the tracts inferred from empirical genomes. When working with empirical data, we need to rely on statistical software

to predict the coordinates of introgressed tracts [52], just as our group did when generating the empirical tracts used in this study (Refoyo-Martínez, *et al.*, in prep.). As a result, the tracts that we are analyzing that come with some statistical uncertainty and errors, an aspect of the data we are safe to disregard when dealing with true tracts extracted from simulations. Consequently, this introduces a certain degree of uncertainty and potential artifacts that could be reflected in the distributions that we obtained.

However, the results that we obtained for the graph-based tract genealogy in space (see Figure 3.5) are suggesting that even inferred tracts have a relatively encouraging degree of precision, and that relying on information such as the coordinates of tract recombination breakpoints might be appropriate.

Although the example of graph-based tract genealogy graph in a geographical context that we obtained in Figure 3.5 was intended as a visual proof-of-concept of how such spatio-temporal graphs might be used in a geographical context, it nevertheless promises several avenues in which this approach could be exploited in the future. In fact, as already mentioned in section 2.3, a graph is not only a data structure that links vertices through edges, but it can incorporate much more quantitative information [93]. For instance, the edges connecting the vertices (that corresponds to individuals) in the spatial graph, represent real geographical distances. Moreover, these individuals were sampled at different time points. Because such tract-based relationship graphs naturally incorporate this quantitative, spatio-temporal information, this opens up the possibility for new kinds of analyses to study genealogy and migrations in time and space. In fact, an ongoing work in our group is currently using an extension of this methodology to build similar graphs based on Identity-by-Descent sharing for studying migration on more recent time-scales.

However, it is important to be cautious before drawing any conclusion from inferred tracts in empirical data, and further filtering is definitely needed. In fact, as other studies pointed out [102], reference based method such as IBDmix [52] might call as Neanderthal introgressed regions of the genome that belong to older introgression in the other direction, from modern humans into Neanderthals [11, 12, 103]. Therefore, it would be ideal to identify and filter out these regions.

Conclusion and future perspectives

In this study, we introduced a new set of summary statistics which leverage the information about demographic history of early modern humans (EMH) reflected in the distribution tracts introgressed from Neanderthals. Our main goal was to evaluate the potential of these statistics for simulation-based inference of population and social dynamics of prehistoric population, following the Out-of-Africa expansion and Neanderthal introgression.

Our primarily motivation was to explore the possibilities of moving from the simplified, single-population panmictic demographic histories of modern humans traditionally used in population genetic research, such as the class of models exemplified by the model by Gravel *et al.* [58], to more detailed models which include prehistoric population structure and social dynamics. To this end, we decided to evaluate the potential of Neanderthal introgressed tracts as a unique source of information, because they appeared on the modern human genetic background at a relatively well-estimated time point of around 49 kya [39]. This implies that their spatio-temporal and genomic distribution observed today in the present-day population (and, for that matter, in EMH genomes over time) has been shaped by the various demographic during the last 50 thousand years.

In this regard, the tract frequency spectrum (TFS), a summary statistics introduced in this study, summarizing the degree of sharing of introgressed tracts between individuals, revealed to be a useful source of information potentially able to capture some aspects of the population structure and dynamics of the demographic history of ancient and present-day West Eurasians. Specifically, although more rigorous statistical modeling is certainly required, a qualitative comparison of TFS betweeen simulated and empirical data suggests that more complex scenarios of prehistoric population substructure may fit the distribution of introgressed tracts observed today better compared to the

traditionally used single-population panmictic models. However, a more detailed evaluation of the properties of TFS will be needed to understand its behaviour under different scenarios, particularly in contrast to other summaries of genetic diversity used in previous studies, such as the classical site frequency spectrum.

Another promising observation is the potential of using the the coordinates of recombination breakpoints of introgressed tracts as markers for fitting the parameters of the dynamics of Neanderthal introgression and potentially reconstructing the tract-based graph of relationships of individuals across space and time. In particular, as suggested by our tract-based genealogy graph analysis, using the precise coordinates of tract recombination breakpoints rather than considering the simple overlap of introgressed regions, guarantees, that two individuals indeed share an introgressed tract at a given locus, and are thus genealogically related. A future step in this regard will be incorporating spatiotemporal information in the graph data structure, which could significantly expand the statistical power of the graph structure to study mobility.

Overall, this proof-of-concept study evaluating the potential of using introgressed tracts to model the dynamics ofintrogression and demographic history of the EMH offers interesting outlooks towards future follow-up work.

For instance, the general simulation based workflow developed for this project, which includes variants of alternative demographic models as well as tree-sequence based summary statistics, will serve a useful basis to provide more detailed evidence to address the question of multiple pulses of introgression into the ancestors of Eurasians. Although it is generally assumed that the vast majority of Neanderthal introgressed tracts in people today can be traced to a single extended pulse of introgression from a single Neanderthal population [41], to our knowledge, no rigorous modeling has been performed to truly reject the hypothesis of additional minor pulses of introgression, or quantify the degree to which these additional pulses contributed to introgressed tracts observed today. It is possible that summary statistics that have been used to date are not sensitive enough to detect these minor introgression contributions, especially if they occurred close in time to the first major introgression pulse [40]. In this study, we showed that two overlapping introgressed regions do not necessarily need to belong to the same introgression event. To this end, we can envision the computation of the graph genealogy of introgressed tracts along the genome, and compare the SNPs carried by these tracts. For instance, if a given region of the genome results in two cluster of tracts in the graph and the archaic SNPs in these two clusters are significantly different, this might suggest that these clusters of tracts originate to introgression from two different

Neanderthal groups. Similarly, it has been suggested that certain ancient individuals (older than 40 kya) carry evidence of additional pulses of introgression, but it seemed that these individuals did not contribute any ancestry to present-day individuals [31, 35, 36]. However, this does not necessarily imply that these additional pulses of introgression in the history of these ancient individuals are not found in present-day individuals today. In particular, it might be possible to test whether recombination tract breakpoints from longer tracts detected in these ancient individuals (i.e, in tracts likely resulting from these additional pulses), are present in later ancient individuals, or even present-day individuals today.

Another potentially promising analysis is the study of migration by using introgressed tracts as spatio-temporal markers of specific ancient populations. For instance, if some tracts drifted and become exclusive to a certain group belonging to a specific geographical area at a certain time in distant past, observing tracts at later points in time in individuals in a different geographical area could be served as a spatio-temporal signal of a migration of people between these two regions in the distant past.

Finally, we propose that the summary metrics developed in this study could be useful to study not only demographic and historical processes, but also to answer questions about adaptive introgression. For instance, by computing the frequency trajectories of different introgressed regions across space and time, one could identify time points at which some introgressed loci significantly increase in frequency, thus indicating positive selection. To this end, we suggest that the *sub-tracts approach* that we developed to encode tracts into binary matrices could be the most valuable source of information for this analysis, as it provides the possibility to partition the introgressed tracts along the genome into bins at the finest scale.

Bibliography

[1] Léo Gabounia, Marie-Antoinette de Lumley, Abesalom Vekua, David Lordkipanidze, and Henry de Lumley. „Découverte d'un nouvel hominidé à Dmanissi (Transcaucasie, Géorgie)“. In: *Comptes Rendus Palevol* 1.4 (Sept. 2002), pp. 243–253.

[2] *Age of the Earliest Known Hominids in Java, Indonesia | Science*.

[3] Richard E. Green, Johannes Krause, Adrian W. Briggs, *et al.* „A Draft Sequence of the Neandertal Genome“. In: *Science* 328.5979 (May 2010). Publisher: American Association for the Advancement of Science, pp. 710–722.

[4] David Reich, Richard E. Green, Martin Kircher, *et al.* „Genetic history of an archaic hominin group from Denisova Cave in Siberia“. en. In: *Nature* 468.7327 (Dec. 2010). Publisher: Nature Publishing Group, pp. 1053–1060.

[5] Peter M. Yaworsky, Emil S. Nielsen, and Trine K. Nielsen. „The Neanderthal niche space of Western Eurasia 145 ka to 30 ka ago“. en. In: *Scientific Reports* 14.1 (Apr. 2024). Publisher: Nature Publishing Group, p. 7788.

[6] Samantha Brown, Diyendo Massilani, Maxim B. Kozlikin, *et al.* „The earliest Denisovans and their cultural adaptation“. eng. In: *Nature Ecology & Evolution* 6.1 (Jan. 2022), pp. 28–35.

[7] Jean-Jacques Hublin, Abdelouahed Ben-Ncer, Shara E. Bailey, *et al.* „New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens“. en. In: *Nature* 546.7657 (June 2017). Publisher: Nature Publishing Group, pp. 289–292.

[8] Eleanor M. L. Scerri, Mark G. Thomas, Andrea Manica, *et al.* „Did Our Species Evolve in Subdivided Populations across Africa, and Why Does It Matter?“ English. In: *Trends in Ecology & Evolution* 33.8 (Aug. 2018). Publisher: Elsevier, pp. 582–594.

[9] Tom Higham, Katerina Douka, Rachel Wood, *et al.* „The timing and spatiotemporal patterning of Neanderthal disappearance“. en. In: *Nature* 512.7514 (Aug. 2014). Publisher: Nature Publishing Group, pp. 306–309.

[10] Viviane Slon, Fabrizio Mafessoni, Benjamin Vernot, *et al.* „The genome of the offspring of a Neanderthal mother and a Denisovan father“. en. In: *Nature* 561.7721 (Sept. 2018). Publisher: Nature Publishing Group, pp. 113–116.

[11] Cosimo Posth, Christoph Wißing, Keiko Kitagawa, *et al.* „Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals“. en. In: *Nature Communications* 8.1 (July 2017). Publisher: Nature Publishing Group, p. 16046.

[12] Daniel N. Harris, Alexander Platt, Matthew E. B. Hansen, *et al.* „Diverse African genomes reveal selection on ancient modern human introgressions in Neanderthals“. In: *Current Biology* 33.22 (Nov. 2023), 4905–4916.e5.

[13] Montgomery Slatkin and Fernando Racimo. „Ancient DNA and human history“. In: *Proceedings of the National Academy of Sciences* 113.23 (June 2016). Publisher: Proceedings of the National Academy of Sciences, pp. 6380–6387.

[14] Adrian W. Briggs, Udo Stenzel, Philip L. F. Johnson, *et al.* „Patterns of damage in genomic DNA sequences from a Neandertal“. eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 104.37 (Sept. 2007), pp. 14616–14621.

[15] Russell Higuchi, Barbara Bowman, Mary Freiberger, Oliver A. Ryder, and Allan C. Wilson. „DNA sequences from the quagga, an extinct member of the horse family“. en. In: *Nature* 312.5991 (Nov. 1984). Publisher: Nature Publishing Group, pp. 282–284.

[16] K. Mullis, F. Faloona, S. Scharf, R. Saiki, G. Horn, and H. Erlich. „Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction“. eng. In: *Cold Spring Harbor Symposia on Quantitative Biology* 51 Pt 1 (1986), pp. 263–273.

[17] James P. Noonan, Michael Hofreiter, Doug Smith, James R. Priest, Nadin Rohland, Gernot Rabeder, Johannes Krause, J. Chris Detter, Svante Pääbo, and Edward M. Rubin. „Genomic Sequencing of Pleistocene Cave Bears“. In: *Science* 309.5734 (July 2005). Publisher: American Association for the Advancement of Science, pp. 597–599.

[18] Marcel Margulies, Michael Egholm, William E. Altman, *et al.* „Genome Sequencing in Open Microfabricated High Density Picoliter Reactors“. In: *Nature* 437.7057 (Sept. 2005), pp. 376–380.

[19] Hendrik N. Poinar, Carsten Schwarz, Ji Qi, *et al.* „Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA“. In: *Science* 311.5759 (Jan. 2006). Publisher: American Association for the Advancement of Science, pp. 392–394.

[20] Pontus Skoglund, Bernd H. Northoff, Michael V. Shunkov, Anatoli P. Derevianko, Svante Pääbo, Johannes Krause, and Mattias Jakobsson. „Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal“. In: *Proceedings of the National Academy of Sciences* 111.6 (Feb. 2014). Publisher: Proceedings of the National Academy of Sciences, pp. 2229–2234.

[21] Jesse Dabney, Matthias Meyer, and Svante Pääbo. „Ancient DNA Damage“. In: *Cold Spring Harbor Perspectives in Biology* 5.7 (July 2013), a012567.

[22] Kurt H. Kjær, Mikkel Winther Pedersen, Bianca De Sanctis, *et al.* „A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA“. en. In: *Nature* 612.7939 (Dec. 2022). Publisher: Nature Publishing Group, pp. 283–291.

[23] Jolijn A M Erven, Amelie Scheu, Marta Pereira Verdugo, Lara Cassidy, Ningbo Chen, Birgit Gehlen, Martin Street, Ole Madsen, and Victoria E Mullin. „A High-Coverage Mesolithic Aurochs Genome and Effective Leveraging of Ancient Cattle Genomes Using Whole Genome Imputation“. In: *Molecular Biology and Evolution* 41.5 (May 2024), msae076.

[24] Fabrizio Mafessoni, Steffi Grote, Cesare de Filippo, *et al.* „A high-coverage Neandertal genome from Chagyrskaya Cave“. In: *Proceedings of the National Academy of Sciences* 117.26 (June 2020). Publisher: Proceedings of the National Academy of Sciences, pp. 15132–15136.

[25] Ke Wang, Kay Prüfer, Ben Krause-Kyora, Ainash Childebayeva, Verena J. Schuenemann, Valentina Coia, Frank Maixner, Albert Zink, Stephan Schiffels, and Johannes Krause. „High-coverage genome of the Tyrolean Iceman reveals unusually high Anatolian farmer ancestry“. In: *Cell Genomics* 3.9 (Sept. 2023), p. 100377.

[26] Morten E. Allentoft, Martin Sikora, Alba Refoyo-Martínez, *et al.* „Population genomics of post-glacial western Eurasia“. en. In: *Nature* 625.7994 (Jan. 2024). Publisher: Nature Publishing Group, pp. 301–311.

[27] Thomaz Pinotti, Michael A. Adler, Richard Mermejo, *et al.* „Picuris Pueblo oral history and genomics reveal continuity in US Southwest“. en. In: *Nature* 642.8066 (June 2025). Publisher: Nature Publishing Group, pp. 125–132.

[28] Mário Vicente and Carina M Schlebusch. „African population history: an ancient DNA perspective“. In: *Current Opinion in Genetics & Development*. Genetics of Human Origin 62 (June 2020), pp. 8–15.

[29] Kay Prüfer, Fernando Racimo, Nick Patterson, *et al.* „The complete genome sequence of a Neanderthal from the Altai Mountains“. en. In: *Nature* 505.7481 (Jan. 2014). Publisher: Nature Publishing Group, pp. 43–49.

[30] Matthias Meyer, Martin Kircher, Marie-Theres Gansauge, *et al.* „A High-Coverage Genome Sequence from an Archaic Denisovan Individual“. In: *Science* 338.6104 (Oct. 2012). Publisher: American Association for the Advancement of Science, pp. 222–226.

[31] Qiaomei Fu, Heng Li, Priya Moorjani, *et al.* „Genome sequence of a 45,000-year-old modern human from western Siberia“. en. In: *Nature* 514.7523 (Oct. 2014). Publisher: Nature Publishing Group, pp. 445–449.

[32] David Reich, Nick Patterson, Martin Kircher, *et al.* „Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania“. In: *American Journal of Human Genetics* 89.4 (Oct. 2011), pp. 516–528.

[33] Davide M. Vespasiani, Guy S. Jacobs, Laura E. Cook, Nicolas Brucato, Matthew Leavesley, Christopher Kinipi, François-Xavier Ricaut, Murray P. Cox, and Irene Gallego Romero. „Denisovan introgression has shaped the immune system of present-day Papuans“. en. In: *PLOS Genetics* 18.12 (Dec. 2022), e1010470.

[34] Sriram Sankararaman, Nick Patterson, Heng Li, Svante Pääbo, and David Reich. „The Date of Interbreeding between Neandertals and Modern Humans“. In: *PLoS Genetics* 8.10 (Oct. 2012), e1002947.

[35] Qiaomei Fu, Mateja Hajdinjak, Oana Teodora Moldovan, *et al.* „An early modern human from Romania with a recent Neanderthal ancestor“. In: *Nature* 524.7564 (Aug. 2015), pp. 216–219.

[36] Mateja Hajdinjak, Fabrizio Mafessoni, Laurits Skov, *et al.* „Initial Upper Palaeolithic humans in Europe had recent Neanderthal ancestry“. en. In: *Nature* 592.7853 (Apr. 2021). Publisher: Nature Publishing Group, pp. 253–257.

[37] Israel Hershkovitz, Ofer Marder, Avner Ayalon, *et al.* „Levantine cranium from Manot Cave (Israel) foreshadows the first European modern humans“. en. In: *Nature* 520.7546 (Apr. 2015). Publisher: Nature Publishing Group, pp. 216–219.

[38] Rasmus Nielsen, Joshua M. Akey, Mattias Jakobsson, Jonathan K. Pritchard, Sarah Tishkoff, and Eske Willerslev. „Tracing the peopling of the world through genomics“. In: *Nature* 541.7637 (Jan. 2017), pp. 302–310.

[39] Arev P. Sümer, Hélène Rougier, Vanessa Villalba-Mouco, *et al.* „Earliest modern human genomes constrain timing of Neanderthal admixture“. en. In: *Nature* 638.8051 (Feb. 2025). Publisher: Nature Publishing Group, pp. 711–717.

[40] Leonardo N M Iasi, Harald Ringbauer, and Benjamin M Peter. „An Extended Admixture Pulse Model Reveals the Limitations to Human–Neandertal Introgression Dating“. In: *Molecular Biology and Evolution* 38.11 (Nov. 2021), pp. 5156–5174.

[41] Leonardo N. M. Iasi, Manjusha Chintalapati, Laurits Skov, Alba Bossoms Mesa, Mateja Hajdinjak, Benjamin M. Peter, and Priya Moorjani. „Neanderthal ancestry through time: Insights from genomes of ancient and present-day humans“. In: *Science* 386.6727 (Dec. 2024). Publisher: American Association for the Advancement of Science, eadq3010.

[42] Sriram Sankararaman, Swapan Mallick, Michael Dannemann, Kay Prüfer, Janet Kelso, Svante Pääbo, Nick Patterson, and David Reich. „The genomic landscape of Neanderthal ancestry in present-day humans“. en. In: *Nature* 507.7492 (Mar. 2014). Publisher: Nature Publishing Group, pp. 354–357.

[43] *Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals | Science*.

[44] Sriram Sankararaman, Swapan Mallick, Nick Patterson, and David Reich. „The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans“. English. In: *Current Biology* 26.9 (May 2016). Publisher: Elsevier, pp. 1241–1247.

[45] Laurent Abi-Rached, Matthew J. Jobin, Subhash Kulkarni, *et al.* „The Shaping of Modern Human Immune Systems by Multiregional Admixture with Archaic Humans“. In: *Science* 334.6052 (Oct. 2011). Publisher: American Association for the Advancement of Science, pp. 89–94.

[46] Emilia Huerta-Sánchez, Xin Jin, Asan, *et al.* „Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA“. en. In: *Nature* 512.7513 (Aug. 2014). Publisher: Nature Publishing Group, pp. 194–197.

[47] Evonne McArthur, David C. Rinker, and John A. Capra. „Quantifying the contribution of Neanderthal introgression to the heritability of complex traits“. en. In: *Nature Communications* 12.1 (July 2021). Publisher: Nature Publishing Group, p. 4481.

[48] Fernando Racimo, Davide Marnetto, and Emilia Huerta-Sánchez. „Signatures of Archaic Adaptive Introgression in Present-Day Human Populations“. In: *Molecular Biology and Evolution* 34.2 (Feb. 2017), pp. 296–317.

[49] J. J. Hublin. „The origin of Neandertals“. In: *Proceedings of the National Academy of Sciences* 106.38 (Sept. 2009). Publisher: Proceedings of the National Academy of Sciences, pp. 16022–16027.

[50] Patrick F. Reilly, Audrey Tjahjadi, Samantha L. Miller, Joshua M. Akey, and Serena Tucci. „The contribution of Neanderthal introgression to modern human traits“. In: *Current Biology* 32.18 (Sept. 2022), R970–R983.

[51] Laurits Skov, Ruoyun Hui, Vladimir Shchur, Asger Hobolth, Aylwyn Scally, Mikkel Heide Schierup, and Richard Durbin. „Detecting archaic introgression using an unadmixed outgroup“. en. In: *PLOS Genetics* 14.9 (Sept. 2018). Publisher: Public Library of Science, e1007641.

[52] Lu Chen, Aaron B. Wolf, Wenqing Fu, Liming Li, and Joshua M. Akey. „Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals“. In: *Cell* 180.4 (Feb. 2020), 677–687.e16.

[53] Sharon R. Browning, Brian L. Browning, Ying Zhou, Serena Tucci, and Joshua M. Akey. „Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture“. English. In: *Cell* 173.1 (Mar. 2018). Publisher: Elsevier, 53–61.e9.

[54] Simone Rubinacci, Diogo M. Ribeiro, Robin J. Hofmeister, and Olivier Delaneau. „Efficient phasing and imputation of low-coverage sequencing data using large reference panels“. en. In: *Nature Genetics* 53.1 (Jan. 2021). Publisher: Nature Publishing Group, pp. 120–126.

[55] Tim D. White, Berhane Asfaw, David DeGusta, Henry Gilbert, Gary D. Richards, Gen Suwa, and F. Clark Howell. „Pleistocene Homo sapiens from Middle Awash, Ethiopia“. en. In: *Nature* 423.6941 (June 2003). Publisher: Nature Publishing Group, pp. 742–747.

[56] Rebecca L. Cann, Mark Stoneking, and Allan C. Wilson. „Mitochondrial DNA and human evolution“. en. In: *Nature* 325.6099 (Jan. 1987). Publisher: Nature Publishing Group, pp. 31–36.

[57] Anders Bergström, Chris Stringer, Mateja Hajdinjak, Eleanor M. L. Scerri, and Pontus Skoglund. „Origins of modern human ancestry“. en. In: *Nature* 590.7845 (Feb. 2021). Publisher: Nature Publishing Group, pp. 229–237.

[58] Simon Gravel, Brenna M. Henn, Ryan N. Gutenkunst, *et al.* „Demographic history and rare allele sharing among human populations“. In: *Proceedings of the National Academy of Sciences* 108.29 (July 2011). Publisher: Proceedings of the National Academy of Sciences, pp. 11983–11988.

[59] Heng Li and Richard Durbin. „Inference of human population history from individual whole-genome sequences“. en. In: *Nature* 475.7357 (July 2011). Publisher: Nature Publishing Group, pp. 493–496.

[60] Maria Teresa Vizzari, Andrea Benazzo, Guido Barbujani, and Silvia Ghirotto. „A Revised Model of Anatomically Modern Human Expansions Out of Africa through a Machine Learning Approximate Bayesian Computation Approach“. In: *Genes* 11.12 (Dec. 2020), p. 1510.

[61] Saioa López, Lucy van Dorp, and Garrett Hellenthal. „Human Dispersal Out of Africa: A Lasting Debate“. In: *Evolutionary Bioinformatics Online* 11.Suppl 2 (Apr. 2016), pp. 57–68.

[62] Stephan Schiffels and Richard Durbin. „Inferring human population size and separation history from multiple genome sequences“. en. In: *Nature Genetics* 46.8 (Aug. 2014). Publisher: Nature Publishing Group, pp. 919–925.

[63] Robin S Waples. „What Is Ne, Anyway?“ In: *Journal of Heredity* 113.4 (July 2022), pp. 371–379.

[64] Motoo Kimura James Franklin Crow. *An Introduction To Population Genetics Theory*.

[65] G. H. Hardy. „Mendelian Proportions in a Mixed Population“. In: *Science* 28.706 (July 1908). Publisher: American Association for the Advancement of Science, pp. 49–50.

[66] R. A. Fisher. „On the dominance ratio“. In: *Bulletin of Mathematical Biology* 52.1 (Jan. 1990), pp. 297–318.

[67] Sewall Wright. „Evolution in Mendelian Populations“. In: *Genetics* 16.2 (Mar. 1931), pp. 97–159.

[68] O. Mazet, W. Rodríguez, S. Grusea, S. Boitard, and L. Chikhi. „On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference?“ en. In: *Heredity* 116.4 (Apr. 2016). Publisher: Nature Publishing Group, pp. 362–371.

[69] Nils Ryman, Linda Laikre, and Ola Hössjer. „Do estimates of contemporary effective population size tell us what we want to know?“ en. In: *Molecular Ecology* 28.8 (2019). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.15027, pp. 1904–1918.

[70] Sara Kurland, Nils Ryman, Ola Hössjer, and Linda Laikre. „Effects of subpopulation extinction on effective size (Ne) of metapopulations“. en. In: *Conservation Genetics* 24.4 (Aug. 2023), pp. 417–433.

[71] Brian P. McEvoy, Joseph E. Powell, Michael E. Goddard, and Peter M. Visscher. „Human population dispersal “Out of Africa” estimated from linkage disequilibrium and allele frequencies of SNPs“. In: *Genome Research* 21.6 (June 2011), pp. 821–829.

[72] Martin Sikora, Andaine Seguin-Orlando, Vitor C. Sousa, *et al.* „Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers“. In: *Science* 358.6363 (Nov. 2017). Publisher: American Association for the Advancement of Science, pp. 659–662.

[73] Alan R. Rogers. „Genetic Evidence for Geographic Structure within the Neanderthal Population“. fr. In: *Peer Community Journal* 4 (2024).

[74] Mark A. Beaumont. „Approximate Bayesian Computation in Evolution and Ecology“. en. In: *Annual Review of Ecology, Evolution, and Systematics* 41.Volume 41, 2010 (Dec. 2010). Publisher: Annual Reviews, pp. 379–406.

[75] Pascale Gerbault, Robin G. Allaby, Nicole Boivin, *et al.* „Storytelling and story testing in domestication“. In: *Proceedings of the National Academy of Sciences* 111.17 (Apr. 2014). Publisher: Proceedings of the National Academy of Sciences, pp. 6159–6164.

[76] Liisa Loog. „Sometimes hidden but always there: the assumptions underlying genetic inference of demographic histories“. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 376.1816 (Nov. 2020). Publisher: Royal Society, p. 20190719.

[77] Martin Petr, Benjamin C. Haller, Peter L. Ralph, and Fernando Racimo. „*slendr*: a framework for spatio-temporal population genomic simulations on geographic landscapes“. fr. In: *Peer Community Journal* 3 (2023).

[78] Franz Baumdicker, Gertjan Bisschop, Daniel Goldstein, *et al.* „Efficient ancestry and mutation simulation with msprime 1.0“. In: *Genetics* 220.3 (Mar. 2022), iyab229.

[79] Kay Prüfer, Cesare de Filippo, Steffi Grote, *et al.* „A high-coverage Neandertal genome from Vindija Cave in Croatia“. In: *Science* 358.6363 (Nov. 2017). Publisher: American Association for the Advancement of Science, pp. 655–658.

[80] Aaron P. Ragsdale and Simon Gravel. „Models of archaic admixture and recent history from two-locus statistics“. en. In: *PLOS Genetics* 15.6 (June 2019). Publisher: Public Library of Science, e1008204.

[81] Iosif Lazaridis. „The evolutionary history of human populations in Europe“. In: *Current Opinion in Genetics & Development*. Genetics of Human Origins 53 (Dec. 2018), pp. 21–27.

[82] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[83] Swapan Mallick, Heng Li, Mark Lipson, *et al.* „The Simons Genome Diversity Project: 300 genomes from 142 diverse populations“. en. In: *Nature* 538.7624 (Oct. 2016). Publisher: Nature Publishing Group, pp. 201–206.

[84] Shuangbin Xu, Meijun Chen, Tingze Feng, Li Zhan, Lang Zhou, and Guangchuang Yu. „Use ggbreak to Effectively Utilize Plotting Space to Deal With Large Datasets and Outliers“. English. In: *Frontiers in Genetics* 12 (Nov. 2021). Publisher: Frontiers.

[85] Georgia Tsambos, Jerome Kelleher, Peter Ralph, Stephen Leslie, and Damjan Vukcevic. „link-ancestors: fast simulation of local ancestry with tree sequence software“. In: *Bioinformatics Advances* 3.1 (Jan. 2023), vbad163.

[86] Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. „Software for Computing and Annotating Genomic Ranges“. en. In: *PLOS Computational Biology* 9.8 (Aug. 2013). Publisher: Public Library of Science, e1003118.

[87] Gabor Csardi and Tamas Nepusz. „The igraph software package for complex network research“. In: *InterJournal* Complex Systems (2006), p. 1695.

[88] Mason Liang and Rasmus Nielsen. „The Lengths of Admixture Tracts“. In: *Genetics* 197.3 (July 2014), pp. 953–967.

[89] Lionel N Di Santo, Claudio S Quilodrán, and Mathias Currat. „Temporal Variation in Introgressed Segments' Length Statistics Computed from a Limited Number of Ancient Genomes Sheds Light on Past Admixture Pulses“. In: *Molecular Biology and Evolution* 40.12 (Dec. 2023), msad252.

[90] Qiaomei Fu, Cosimo Posth, Mateja Hajdinjak, *et al.* „The genetic history of Ice Age Europe“. en. In: *Nature* 534.7606 (June 2016). Publisher: Nature Publishing Group, pp. 200–205.

[91] Motoo Kimura. „The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population Due to Steady Flux of Mutations“. In: *Genetics* 61.4 (Apr. 1969), pp. 893–903.

[92] Benjamin Vernot and Joshua M. Akey. „Resurrecting surviving Neandertal lineages from modern human genomes“. eng. In: *Science (New York, N.Y.)* 343.6174 (Feb. 2014), pp. 1017–1021.

[93] Matthias Dehmer, Frank Emmert-Streib, and Yongtang Shi. „Quantitative Graph Theory: A new branch of graph theory and network science“. In: *Information Sciences* 418-419 (Dec. 2017), pp. 575–580.

[94] J. Claiborne Stephens. „On the Frequency of Undetectable Recombination Events“. In: *Genetics* 112.4 (Apr. 1986), pp. 923–926.

[95] Richard R. Hudson and Norman L. Kaplan. „Statistical Properties of the Number of Recombination Events in the History of a Sample of DNA Sequences“. In: *Genetics* 111.1 (Sept. 1985), pp. 147–164.

[96] Bin Zhang and S. Srihari. „Properties of Binary Vector Dissimilarity Measures“. In: 2003.

[97] Katalin Csilléry, Michael G. B. Blum, Oscar E. Gaggiotti, and Olivier François. „Approximate Bayesian Computation (ABC) in practice“. eng. In: *Trends in Ecology & Evolution* 25.7 (July 2010), pp. 410–418.

[98] „A global reference for human genetic variation“. In: *Nature* 526.7571 (2015), pp. 68–74.

[99] Edzer Pebesma. „Simple Features for R: Standardized Support for Spatial Vector Data“. In: *The R Journal* 10.1 (2018), pp. 439–446.

[100] Lucas van der Meer, Lorena Abad, Andrea Gilardi, and Robin Lovelace. *sfnetworks: Tidy Geospatial Networks*. R package version 0.6.5, https://github.com/luukvdmeer/sfne 2024.

[101] Benjamin C. Haller, Philipp W. Messer, and Peter L. Ralph. *SLiM 5: Eco-evolutionary simulations across multiple chromosomes and full genomes*. en. ISSN: 2692-8205 Pages: 2025.08.07.669155 Section: New Results. Aug. 2025.

[102] Liming Li, Troy J. Comi, Rob F. Bierman, and Joshua M. Akey. „Recurrent gene flow between Neanderthals and modern humans over the past 200,000 years“. In: *Science* 385.6705 (July 2024). Publisher: American Association for the Advancement of Science, eadi1768.

[103] Martin Petr, Mateja Hajdinjak, Qiaomei Fu, *et al.* „The evolutionary history of Neanderthal and Denisovan Y chromosomes“. In: *Science* 369.6511 (Sept. 2020). Publisher: American Association for the Advancement of Science, pp. 1653–1656.

Supplementary Information

S4 Chapter 1: Demographic scenarios for an Out-of-Africa expansion of modern humans

S4.1 Models parameters

Parameter	Description	Value	Unit
$T_{admix\_1\_start}$	Start of OOA–NEA admixture	49	kya
$T_{admix\_1\_end}$	End of OOA–NEA admixture	45	kya
$T_{admix\_2\_start}$	Start of NEA–EUR admixture	42	kya
$T_{admix\_2\_end}$	End of NEA–EUR admixture	40	kya
gf_rate_1	Gene flow from NEA to OOA	**A:** 3; **B:** 2	%
gf_rate_2	Gene flow from NEA to EUR	**A:** 0; **B:** 1	%
T_{split}	EUR splits from OOA	42	kya
N_{NEA}	Neanderthal population size	2000	
N_{OOA}	OOA population size (parametrized)	500–10000 (step 500)	
N_{EUR}	EUR population size	15000	
N_{AFR}	African population size	15000	
N_{ANC}	Ancestral population size	15000	
T_{OOA}	OOA splits from AFR	70	kya
$T_{AFR\&NEA}$	AFR and NEA splits from ANC	650	kya
g	Generation time	30	years
Chr_{length}	Chromosome length	100	Mb
n_{NEA}	NEA sample size	2	diploid genomes
n_{AFR}	AFR sample size	5	diploid genomes
n_{EUR}	EUR sample size	50	diploid genomes
t_{EUR}	EUR sampling time points	0	kya
t_{NEA}	NEA sampling time points	50	kya
t_{AFR}	AFR sampling time points	0	kya

Table S1: Parameters used in Model 1.

Parameter	Description	Value	Unit
$T_{admix\_1\_start}$	Start of OOA–NEA admixture	49	kya
$T_{admix\_1\_end}$	End of OOA–NEA admixture	45	kya
$T_{admix\_2\_start}$	Start of NEA–EUR_x admixture	42	kya
$T_{admix\_2\_end}$	End of NEA–EUR_x admixture	40	kya
$T_{admix\_ANA\_start}$	Start of ANA-WHG admixture	8.5	kya
$T_{admix\_ANA\_end}$	End of ANA-WHG admixture	3	kya
$T_{admix\_YAM\_start}$	Start of YAM-WHG admixture	5.5	kya
$T_{admix\_YAM\_end}$	End of YAM-WHG admixture	3	kya
gf_rate_1	Gene flow from NEA to OOA	**A:** 3; **B:** 2	%
gf_rate_2	Gene flow from NEA to each EUR_x	**A:** 0; **B:** 1	%
gf_rate_ANA	Gene flow from ANA to WHG	50	%
gf_rate_YAM	Gene flow from YAM to WHG	50	%
T_{split}	EUR_x splits from OOA	42	kya
T_{EUR}	WHG becomes EUR	3	kya
N_{NEA}	Neanderthal population size	2000	
N_{OOA}	OOA population size (parametrized)	500–10000 (step 500)	
N_{EUR_x}	EUR_x population size (demes)	2000	
N_{WHG}	WHG population size	5000	
N_{ANA}	ANA population size	5000	
N_{YAM}	YAM population size	5000	
N_{EUR}	EUR population size	15000	
N_{AFR}	African population size	15000	
N_{ANC}	Ancestral population size	15000	
T_{OOA}	OOA splits from AFR	70	kya
$T_{AFR\&NEA}$	AFR and NEA splits from ANC	650	kya
g	Generation time	30	years
Chr_{length}	Chromosome length	100	Mb
n_{NEA}	NEA sample size	2	diploid genomes
n_{AFR}	AFR sample size	5	diploid genomes
n_{EUR}	EUR sample size	50	diploid genomes
t_{EUR}	EUR sampling time points	0	kya
t_{NEA}	NEA sampling time points	50	kya
t_{AFR}	AFR sampling time points	0	kya

Table S2: Parameters used in Model 2.

Parameter	Description	Value	Unit
$T_{admix\_1\_start}$	Start of OOA–NEA admixture	49	kya
$T_{admix\_1\_end}$	End of OOA–NEA admixture	45	kya
$T_{admix\_2\_start}$	Start of NEA–EUR_x admixture	42	kya
$T_{admix\_2\_end}$	End of NEA–EUR_x admixture	40	kya
$T_{admix\_3\_start}$	Start of EUR_x-EUR_(x+1) admixture	30	kya
$T_{admix\_3\_end}$	End of EUR_x-EUR_(x+1) admixture	3	kya
gf_rate_1	Gene flow from NEA to OOA	**A:** 3; **B:** 2	%
gf_rate_2	Gene flow from NEA to each EUR_x	**A:** 0; **B:** 1	%
gf_rate_3	Gene flow from EUR_x to EUR_(x+1)	11	%
T_{split}	EUR_x splits from OOA	42	kya
T_{merge}	EUR_x demes merge into EUR	3	kya
N_{NEA}	Neanderthal population size	2000	
N_{OOA}	OOA population size (parametrized)	500–10000 (step 500)	
N_{EUR_x}	EUR_x population size (demes)	2000	
N_{EUR}	EUR population size	15000	
N_{AFR}	African population size	15000	
N_{ANC}	Ancestral population size	15000	
T_{OOA}	OOA splits from AFR	70	kya
$T_{AFR\&NEA}$	AFR and NEA splits from ANC	650	kya
g	Generation time	30	years
Chr_{length}	Chromosome length	100	Mb
n_{NEA}	NEA sample size	2	diploid genomes
n_{AFR}	AFR sample size	5	diploid genomes
n_{EUR}	EUR sample size	50	diploid genomes
t_{EUR}	EUR sampling time points	0	kya
t_{NEA}	NEA sampling time points	50	kya
t_{AFR}	AFR sampling time points	0	kya

Table S3: Parameters used in Model 3.

S4.2 Visualization of the demographic history of the models

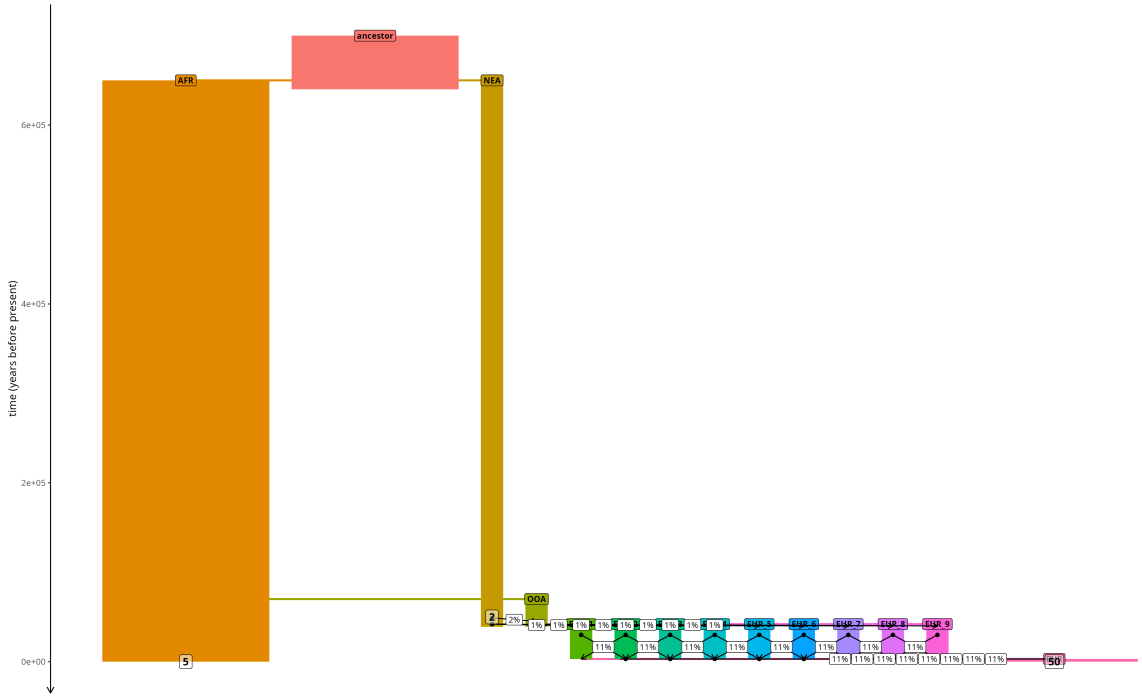

Figure S7: Example of a cluttered representation of the demographic history of a more complex model. The figure was generated using the default parameters of the *plot_model* function in the *slendr* R package.

S5 Chapter 2: Developing tract-based summary statistics

S5.1 Tract genealogy graph

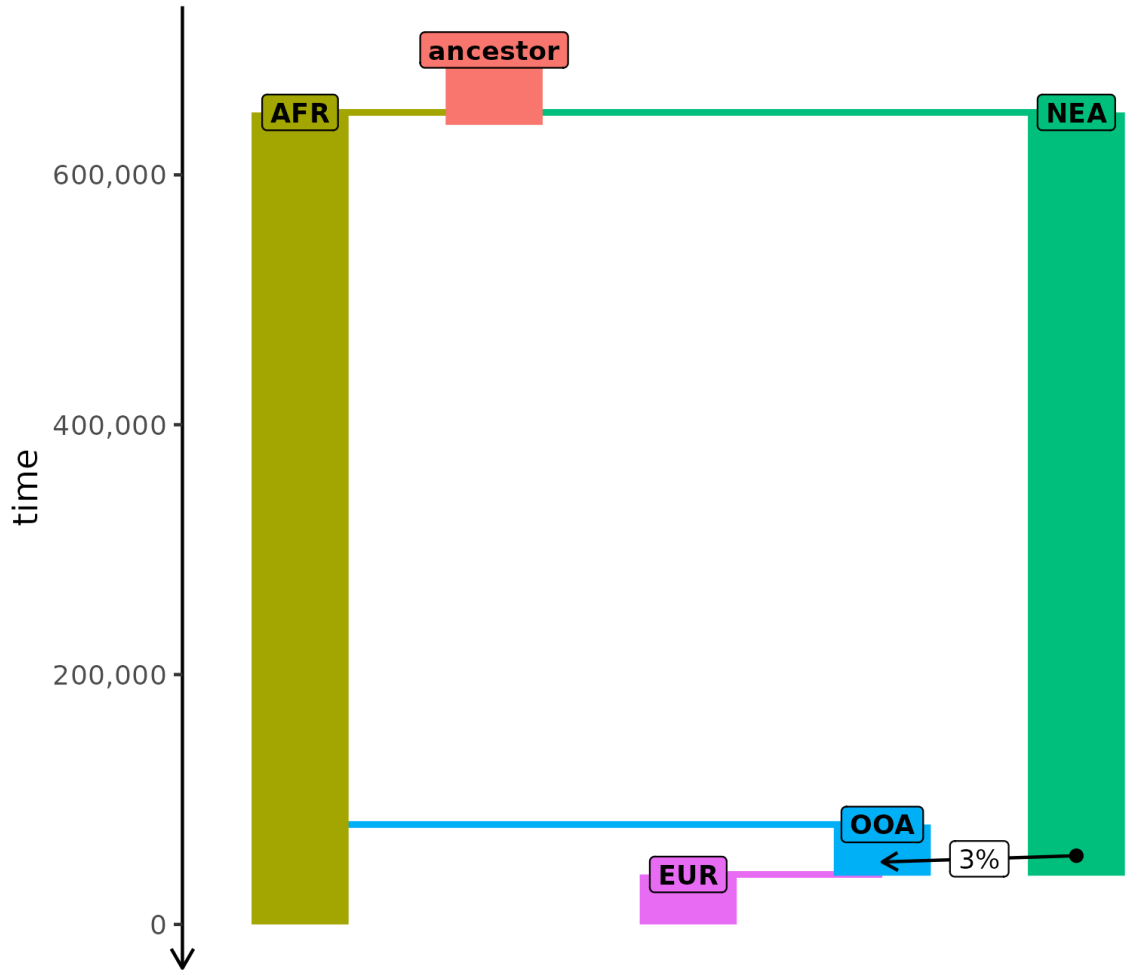

Figure S8: Schema of the model used for evaluating our tract genealogy graph.

Parameter	Description	Value	Unit
T_{admix_start}	Start of OOA–NEA admixture	55	kya
T_{admix_end}	End of OOA–NEA admixture	50	kya
gf_rate	Gene flow from NEA to OOA	3	%
T_{split}	EUR splits from OOA	40	kya
N_{NEA}	Neanderthal population size	10	
N_{OOA}	Out-of-Africa population size	5000	
N_{EUR}	EUR population size	5000	
N_{AFR}	African population size	10000	
N_{ANC}	Ancestral population size	10000	
T_{OOA}	OOA splits from AFR	80	kya
$T_{AFR\&NEA}$	AFR and NEA splits from ANC	650	kya
g	Generation time	30	years
Chr_{length}	Chromosome length	5	Mb
n_{NEA}	NEA sample size	1	diploid genomes
n_{AFR}	AFR sample size	1	diploid genomes
n_{EUR}	EUR sample size	5	diploid genomes
n_{OOA}	OOA sample size	5	diploid genomes
t_{OOA}	OOA sampling time points	50, 40	kya
t_{EUR}	EUR sampling time points	30, 20, 10, 5, 0	kya
t_{NEA}	NEA sampling time points	50	kya
t_{AFR}	AFR sampling time points	0	kya

Table S4: Parameters used in the model used for evaluating our tract genealogy graph. The sample sizes are for sampling time points. Note the extreme low Neanderthal population size used to force coalescence.

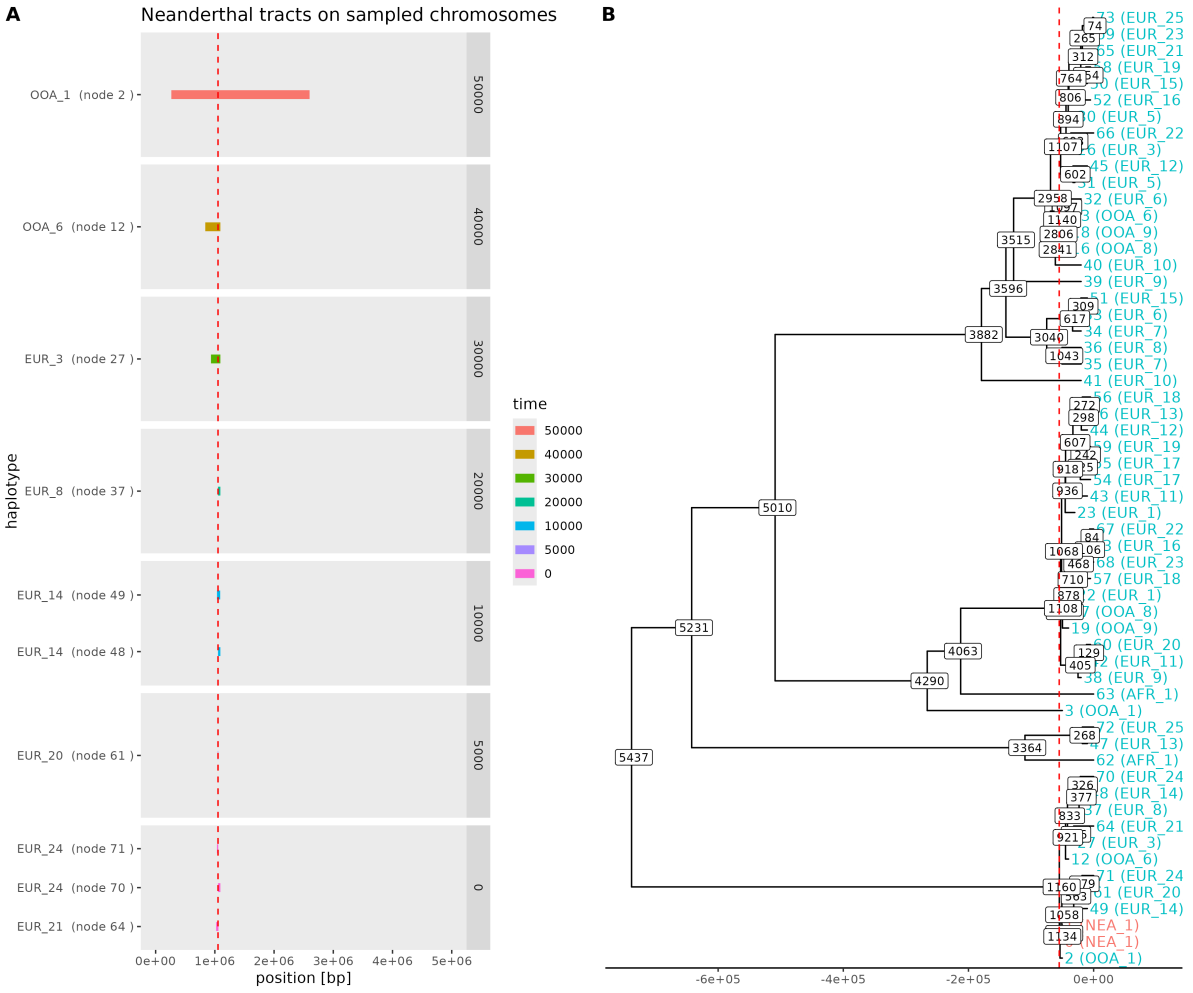

Figure S9: True genealogy of introgressed tracts for a genomic location. (**A**) shows a visual representation of introgressed tracts in sampled genomes. Only the tracts that intersect that specific genomic location (red dashed line) are shown. The tracts are colored according to their sampling time point. (**B**) shows the true genealogy of that genomic location for all the sampled chromosomes. The red dashed line corresponds to the beginning of the admixture (55 kya). Note how the samples with an introgressed tract cluster with the Neanderthal sample (red), while the others cluster with the African sample.

S6 Chapter 3: Proof-of-concept application on empirical data

S6.1 Tract frequency spectrum for present-day West Eurasians

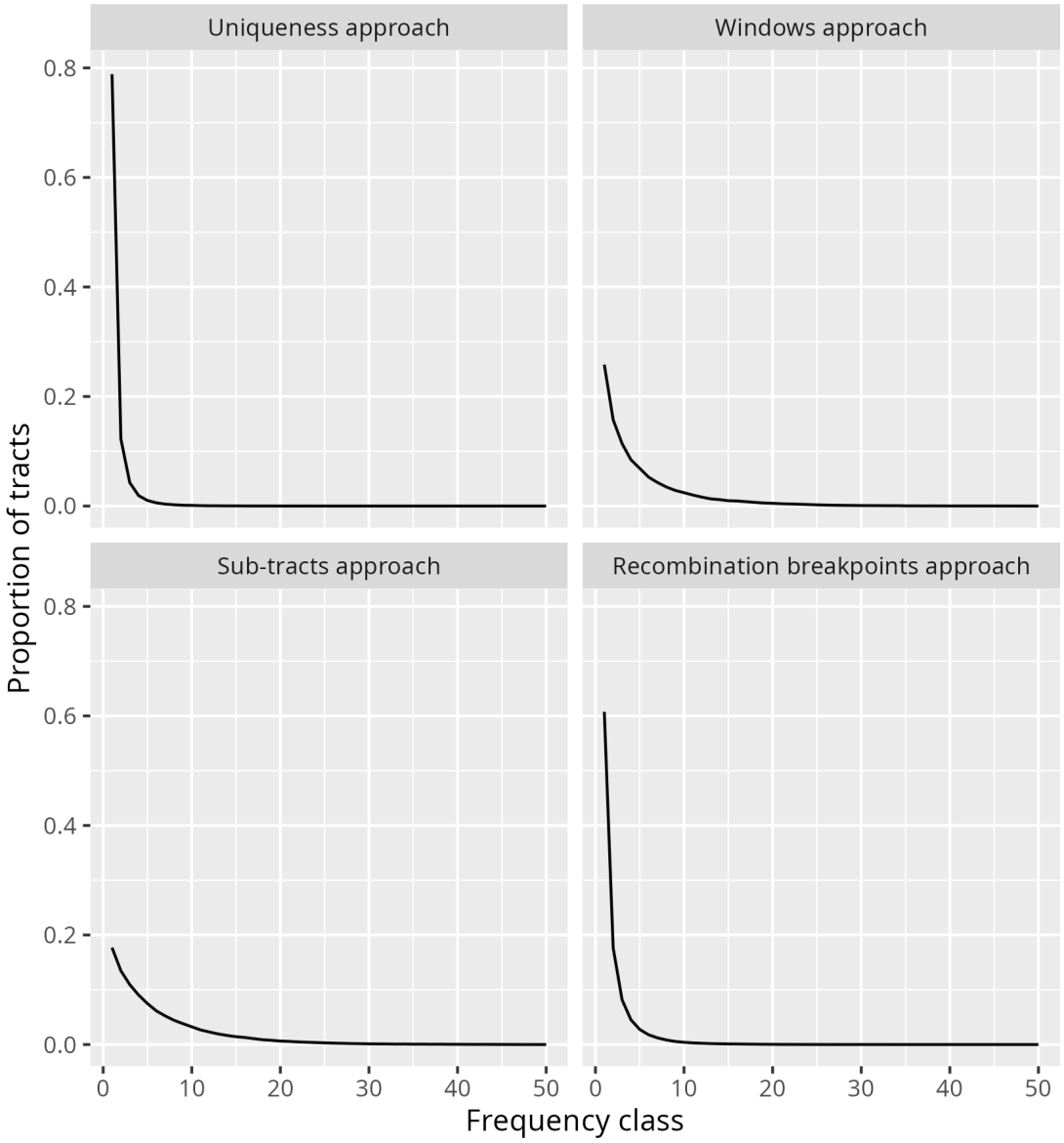

Figure S10: Tract frequency spectrum for present-day West Eurasians with the various tract-encoding approaches. Each grid corresponds to the TFS for 50 present-day West Eurasians where Neanderthal tracts were inferred using a different tract-encoding approach. The TFS correspond to the average over 10 independent sampling of 50 individuals. Alyhough not visible, the plots also show the standard deviation computed across the 10 sampling as a light gray area around the curve.