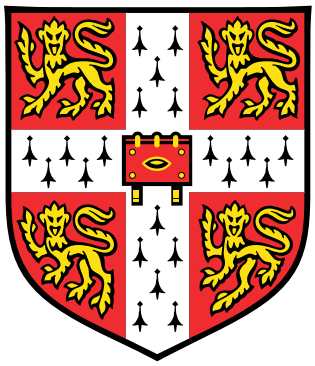
Phylogenetic inference using ancient environmental DNA



Bianca Diana De Sanctis

Department of Zoology, Department of Genetics

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Wolfson College

December 2022

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Bianca De Sanctis

Phylogenetic inference using ancient environmental DNA

Bianca De Sanctis

Ancient environmental DNA (aeDNA) has revolutionized our ability to describe and analyze biological communities in space and time by allowing for joint sequencing of entire ecosystems across thousands of years. However, because samples contain damaged, short fragments from multiple individuals or taxa, the field has been so far limited in its scope, and aeDNA has only been applied to population and phylogenetic studies in the last few years. In this thesis, I first build a theoretical coalescent framework to analyze error in supervised binning algorithms, which assign reads from environmental samples to individual taxa in a reference database. Under this framework, I determine the expected error rate under a wide range of parameters and the degradation in assignment accuracy as samples diverge from their closest reference sequence, and with incompleteness of reference sequences. Second, I describe a phylogenetic placement algorithm for non-recombining sequences such as mitochondria or chloroplast DNA, and apply this method to *Mammuthus* or mammoth and *Equus* or horse samples from an Arctic-wide aeDNA dataset spanning the last 50,000 years. This analysis demonstrates the potential existence of a previously undiscovered clade of mammoths, and extends the survival of an existing clade. Next, I report one of the first whole genome ancient environmental DNA studies, using DNA extracted from 14-16,000 year old cave soil with material from two closely related species, *Ursus arctos* or the American black bear and *Arctodus simus* or the extinct giant short-faced bear. By comparing the ancient sequence against a modern reference panel of black bears and a high-quality fossil giant short-faced bear reference, I infer evolutionary relationships between the Late Pleistocene populations and their modern relatives. Lastly, I molecularly date an ancient environmental *Betula* or birch tree chloroplast sequence from Northern Greenland, confirming that it was approximately 2 million years old, the oldest DNA to be successfully sequenced so far. All together, this work demonstrates the ability to infer phylogenies and population histories of individual taxa from ancient environmental DNA.

Acknowledgements

These past few years have genuinely been so much fun, and there's a lot of people to thank for that.

I somehow managed to get not one but three supportive, kind, and brilliant supervisors. Richard Durbin taught me to be precise and thorough, and is always handing out nuggets of wisdom. Once, during a particularly stressful week in lockdown when I wanted to give up on an analysis that didn't seem to be working, he told me "false pessimism is bad for science" – which helped so much that I still have it written on a sticky note on my desk. John Welch and I had countless spontaneous and productive meetings full of laughter. He also has an astonishing intuition, and listening to his thoughts over the years has helped train me to think more biologically. And there are no dull moments working with Eske Willerslev, who inspires me and everyone around him to think above and beyond.

I've been lucky to be a part of some amazing academic collaborations in the last few years. There are far too many people to list here, but I especially want to thank Hilde Schneemann, Yucheng Wang, Ruairidh Macleod, Mikkel Pedersen, Rui Martiniano, and Moritz Blumer. I also want to thank my local circus community for keeping me sane: Masha, Abbi, Rachael, Ilaria, Florencia, Jo, Verity and so many more. It's hard to remember any of your problems when you're upside down in the air and surrounded by friends.

Thank you to my partner Fred Johnstone, one of the kindest people I've ever met and my best friend in the whole world, for always keeping me grounded. Lastly, I'm so grateful for the most loving and supportive parents I could imagine. They have never stopped believing in me and cheering me on.

Contents

1 Introduction **6**

1.1 Environmental DNA 6
1.2 Ancient DNA 10
1.3 Ancient environmental DNA 14
1.4 Population genetics and phylogenetics in the context of aeDNA 20
1.5 Outline 26

2 A Theoretical Analysis of Taxonomic Binning Accuracy **28**

2.1 Introduction 29
2.2 Materials and Methods 32
2.2.1 Extensions to the Model 36
2.2.2 Modelling Incomplete Reference Sequences 39
2.2.3 Simulations 40
2.3 Results and discussion 41
2.3.1 Theoretical results 41
2.3.2 Simulation results 45
2.4 Conclusion 48

3 Phylogenetic placement of Arctic mammoth and horse from ancient environmental DNA **53**

3.1 Introduction 55
3.1.1 Evolutionary history of mammoths 55
3.1.2 Evolutionary history of horses 56
3.1.3 Phylogenetic placement with pathPhynder 57
3.2 Materials and Methods 60
3.2.1 Previously published reference genomes 60
3.2.2 Permafrost data 63
3.2.3 Workflow 66
3.3 Results and Discussion 70
3.3.1 Mammoths 70
3.3.2 Horses 77
3.4 Conclusion 82

4 Environmental Genomics of Late Pleistocene Black Bears and Giant Short-Faced Bears **84**
4.1 Introduction . . . 85
4.2 Methods . . . 86
4.2.1 Experimental methods and mapping pipeline . . . 86
4.2.2 Black bear analysis . . . 86
4.2.3 Giant short-faced bear fossil analysis . . . 91
4.2.4 Giant short-faced bear eDNA analyses . . . 92
4.3 Results and discussion . . . 93
4.3.1 Black bear . . . 93
4.3.2 Giant short-faced bear . . . 97
4.4 Conclusion . . . 101

5 Molecular dating of a *Betula* chloroplast aeDNA sequence from Northern Greenland **102**
5.1 Introduction . . . 103
5.2 Methods . . . 105
5.2.1 Extracting and mapping reads . . . 105
5.2.2 Phylogenetic placement . . . 107
5.2.3 Molecular dating with phylogenetic placement and SNP-counting . . . 108
5.2.4 Molecular dating with BEAST . . . 115
5.3 Results and discussion . . . 118
5.3.1 Phylogenetic placements . . . 118
5.3.2 Molecular dating . . . 123
5.4 Conclusion . . . 124

6 Conclusion **127**

References **131**

Appendix: Metadata for permafrost samples **152**

Appendix: Larger versions of selected figures **170**

1 Introduction

This work concerns phylogenetic and population genetic inferences made using ancient environmental DNA (aeDNA). Until recently, ancient (and modern) environmental DNA has mainly been used to detect the presence of species in an ecosystem. Ideally we would like to go beyond this, and exploit the sequence variation within species to make inferences about the evolutionary history of species in these ecosystems. However, this can be difficult because aeDNA data are low coverage, contaminated, fragmented, damaged, and represents a mixed sample, so that the number of genetic loci which can be recovered in the genome of any individual species is often insufficient for traditional population genetic and phylogenetic approaches. This difficulty is exacerbated by the use of capture methods, which are useful for species detection but only recover targeted genetic regions, and by the lack of necessary reference genomes. In the last few years, the use of shotgun sequencing, improved computational methods, clean lab techniques and reference databases, and the ability to sequence more fragments than ever has allowed us to sequence enough genetic loci in individual taxa to start to use phylogenetic and population genetic approaches for aeDNA. In this introductory chapter, I first review the fields of environmental DNA and ancient DNA separately, including their use, history and challenges, and then their convergence as ancient environmental DNA. Next, I examine phylogenetic and population genetic inference using aeDNA, covering the foundations of these fields, common algorithms, and their uses and computational challenges in aeDNA. Lastly, I outline the remainder of this thesis.

1.1 Environmental DNA

DNA extracted and sequenced from environmental sources such as soil or water is referred to as environmental DNA. Environmental DNA (eDNA) allows us to study the genetic makeup of an entire ecosystem using the molecular traces left behind by the organisms that inhabit it. Animal, plant and microbial species can leave DNA in their environment by the way of skin, mucous, saliva, sperm, secretions, eggs, feces, urine, feathers, blood, roots, leaves, fruit, pollen, rotting bodies, and more (Ruppert et al., 2019; Pedersen et al., 2015). We can now obtain eDNA from a wide variety of sources, including soil, water, ice, permafrost, and less conventional sources such as air, snow tracks and salt licks (Ruppert et al., 2019; Brys et al., 2020; Ishige et al., 2017).

Compared to traditional biodiversity monitoring approaches such as camera traps, casting nets or direct observation, eDNA provides a different type of information which, for example, is better suited to detect rare or endangered species which are not easily observable. Using genetic markers as opposed to visual inspection can also ease the process of species identification, since identi-

fying species by their phenotype alone can be difficult and often requires taxonomic expertise. Sometimes, differentiating related species phenotypically is nearly impossible (eg. Boddé et al. (2022)), or individuals belong to poorly studied taxonomic groups and are not easily recognizable (Carew et al., 2013). At times, the relevant expertise needed is simply not available. On the other hand, though species identification using eDNA remains non-trivial, specialized taxonomic expertise is not required, and even related species usually have well-defined differences in their genomes. Unlike traditional monitoring approaches, eDNA can also record fine-scale genetic variation in populations which might not have a phenotypic effect, yielding a better understanding of factors such as population structure, diversity and speciation. Lastly, with the falling cost of sequencing, eDNA analyses are becoming more rapid and cost-effective than traditional approaches, so that its popularity as a tool is rising quickly (Ruppert et al., 2019).

The main use of eDNA so far has concerned the questions of presence, that is, whether taxa of interest are present in a given environment. This is already useful in many ways. First of all, eDNA can inform conservation practices, such as in Mizumoto et al. (2020) who used eDNA from water to determine the population structure of a critically endangered salmonid fish in Japan by measuring presence-absence in 120 rivers. It can determine subspecies ranges, such as in Gorički et al. (2017) who used specific molecular sequences to distinguish between two colour morphs of endangered salamanders using eDNA from water in the Balkan Peninsula. It can help with invasive species monitoring, such as Hunter et al. (2015) who used eDNA to infer the leading edge of the distribution of the Burmese python. It has even been recently shown that eDNA can be obtained directly from air, in Lynggaard et al. (2022), where airborne eDNA was sequenced from a zoo in Copenhagen and led to the detection of 49 vertebrate zoo species.

Since a typical goal of many eDNA studies is to determine the presence of only one or a few desired taxa (eg. (Mizumoto et al., 2020; Gorički et al., 2017)), capture or enrichment approaches have been developed to amplify the amount of DNA from a target species before sequencing. One method is to use bulk amplification relying on conserved primer annealing sites, but this can give biased results due to PCR amplification biases Wilcox et al. (2018) (discussed further in the next section of this chapter). Another is to use hybridization capture, which uses probes to hybridize genetic regions associated with a target species. On the computational side, studies often use metabarcoding, which is the practice of identifying species in metagenomic samples with genetic barcodes, or short regions associated uniquely with that species. To this end, there is a vast public reference database of barcodes called the International Barcode of Life (iBOL, 2022). When interested in an entire ecosystem or a larger number of taxa, studies will often use shotgun sequencing instead, and map all filtered reads to a large reference database such as NCBI (Wang et al., 2022).

Environmental DNA comes with challenges. First, eDNA samples tend to yield a very small amount of total DNA, and therefore the resulting data are low copy number and hence low coverage. Low copy number issues can be partially overcome by capture methods when interested in specific taxa, as discussed above, or by sequencing more total DNA, which requires more funding. In the case when there is a single target taxon, one can also sample preferentially where the organism is known to have left more genetic material. For example, Parsons et al. (2018) sampled eDNA for harbour porpoises in their fluke prints, which are patches of calm water that indicate the recent presence of a marine animal, and Dugal et al. (2022) sampled water eDNA behind individuals sharks. For downstream population genetic analyses, this low copy number issue is compounded by the likely existence of multiple individuals in a sample. With low coverage, it can be difficult to determine if variation at a site across reads is due to genuine genetic variation in the species or due to error.

Since there are thousands of mitochondria and on the order of a hundred chloroplasts in a cell, as opposed to a single nucleus, eDNA samples will tend to have higher coverage of mitochondrial or chloroplast DNA than it will nuclear DNA. This is useful in some ways, since there exists more mitochondrial and chloroplast reference genomes than nuclear, and reference genomes are needed to reliably identify taxa in the sample (Howe et al., 2020). Additionally, mitochondrial and chloroplast DNA are generally non-recombining, which eases downstream phylogenetic analyses with reads obtained from eDNA (Ladoukakis and Zouros, 2017). On the other hand, mitochondrial and chloroplast genomes are much smaller than nuclear genomes, yield less information about species relationships, and can be very similar across related species.

In addition to its low copy number, sequencing DNA from an entire ecosystem at once presents the unique challenge of identifying the individual taxa in the ecosystem, or assigning individual reads to taxa. This technical process, often known as binning, is difficult to do with high accuracy since related taxa share long segments of their DNA with each other. Because of this, reads assigned to taxa are often biased towards higher coverage in genetic regions which are more unique to those taxa, or which contain more taxon-specific mutations or structural variants. The development of binning algorithms requires novel mathematical or computational techniques. There are now a wide variety of binning algorithms available, some meant specifically for microbial or for animal or plant taxa, but errors in taxon identification are still widespread and difficult to avoid entirely because of population genetic factors. Chapter 2 of this thesis is dedicated to quantifying this error rate, where this aspect of eDNA is discussed more comprehensively.

A further issue, which complicates the use of eDNA for anything more complicated than taxa detection such as studying genetic diversity or evolutionary history, is the uneven distribution of

genetic material across environments and between organisms. This affects any population genetic inferences that rely on allele frequencies, and means that even creating reliable abundance estimates from eDNA is nontrivial. Organisms do not shed equal amounts of DNA, and so studies using relative abundances need to account for differences based on body size, feeding habits and habitat use, among other factors (Ruppert et al., 2019). Variable sources of DNA may also degrade at different rates. For example, Goldberg et al. (2016) found that eDNA in water is undetectable after 1 day to 8 weeks depending on the system, but DNA which is bound to sediment in water degrades at a much slower rate. A single sample may not accurately represent the contents of the surrounding ecosystem, so independent samples should be taken and cross-validated with each other to ensure reliability if an understanding of the entire ecosystem is wanted. For example, a sample using soil outside a bear den might capture DNA from bears but not wolves. One could even account for seasonality, as some organisms will consume and shed more during the warmer months, and DNA preservation may additionally be temperature dependent (Ruppert et al., 2019). Along with the low coverage of eDNA, this means it is even tough to claim the absence of a species in an environment by the absence of its DNA in an environmental sample. However, only a subset of these factors will convolute attempts to study abundance of an individual taxa over time, and so inferences about abundance changes in a single taxa might be considered more reliable than comparisons between taxa.

All of these difficulties have mostly limited the field to species detection until this point (Sigsgaard et al., 2020). However, eDNA carries significantly more possibilities than this. If we could reliably separate out species and sequence sufficiently many reads from multiple samples within a single environment, we could learn about the evolutionary history and diversity of every species within it. Furthermore, doing this over a series of time points could yield a real-time understanding of ecosystem development, such as responses to climate change, and of the relationships of species to one another. The potential of this is vast. For example, it has been suggested that eDNA could be used to study all possible environmental biotic exposures implicated in human diseases (Thakur and Roy, 2020). For example, Brennan et al. (2019) studied airborne human allergens by sequencing DNA by capturing pollen in air, which could be extended to study the change in airborne allergens over time by sequencing airborne eDNA, directly informing healthcare efforts. Another study investigated the potential of eDNA to reduce or prevent malaria and other mosquito-borne infectious diseases through the tracking of insect larvae DNA (Sakata et al., 2022). In theory, even in a single timepoint, eDNA reads assigned to individual taxa could be used on the same scale as if one had sequenced only that organism, and so could be used to infer an extensive amount about a species, including functional mutations, population size, sex ratio, demographic history,

hybridization, introgression, selection, and population structure.

In reality, the use of eDNA has only very recently been extended beyond presence-absence into the realm of population-level or functional inferences. For example, last year, eDNA from soil was used to study the impact and function of mutations in individual microbial species after forest fires (Köster et al., 2021). In Djurhuus et al. (2020), they used correlations between species from eDNA over an eighteen month period to infer a community network graph of predator prey relationships and to identify biological predictors of ecosystem changes. Dugal et al. (2022) was able to assign shark reads from eDNA to six different haplotypes, and confirmed their analysis using tissue samples from the actual sharks. Though eDNA research has been underway for several decades now, we are just beginning to uncover the true potential of this field.

1.2 Ancient DNA

DNA that originates from an ancient source, called ancient DNA, was first sequenced in 1984 from the bone of a quagga, an extinct relative of the zebra (Higuchi et al., 1984) (some literature differentiates between historical and ancient DNA, but we choose not to here). The field exploded, both scientifically and in the public eye, with claims of DNA extracted from >10 million year old specimens, which were later proven to be false and stemming from contamination (e.g. Sidow et al. (1991); Woodward et al. (1994)). Two decades later, the technological advance of next generation sequencing changed the field dramatically, allowing many more reads to be sequenced at once and giving a more comprehensive picture of which fragments originated from real, ancient sources as opposed to contaminants. An in-depth historical review of the field can be found in Jones (2022). This section reviews where the field of ancient DNA is now, its common uses, and its remaining associated challenges.

Ancient DNA has been crucial in understanding human demography and our relationship to other hominids such as Neanderthals and Denisovans. The first mitochondrial Neanderthal genome was published in 2009 (Green et al., 2008), a draft genome in 2010 (Green et al., 2010) and a complete genome in 2013 (Prüfer et al., 2013), the last of which led to the first Nobel prize for ancient DNA, awarded a few months ago to Svante Paabo. Since then, we have obtained many more Neanderthal and Denisovan genomes. These have been used to estimate population split times and to understand admixture timing and location. Though it was long thought that Neanderthals and Homo sapiens did not interbreed due to a lack of archaeological evidence (Slatkin and Racimo, 2016), ancient DNA has conclusively proven the exact opposite. It is now accepted that non-African humans inherit an average of 2% of their DNA from Neanderthal origins (Sankararaman et al., 2014). Furthermore, a recent study sequenced an ancient hominid from Denisova Cave in Russia,

and found that the individual had a Neanderthal mother and a Denisovan father (Slon et al., 2018). These ancient hominid genomes have also illuminated some interesting functional adaptations, both in other species and in our own. A classic example is the introgression of a high-altitude tolerance related gene, EPAS1, from Denisovans into modern Tibetans (Huerta-Sánchez et al., 2014). Zeberg and Pääbo (2020) showed that a major genetic risk factor for post-infection respiratory failure after severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was inherited from Neanderthals, and is present in about 50% of South Asians. Similarly, there are human alleles with Neanderthal origins that are relevant to Crohn's disease, lupus and type 2 diabetes (Sankararaman et al., 2014).

How extinct species lived and why they might have gone extinct can also be clarified by using ancient DNA. This includes functional adaptations of extinct species which have since been lost. For example, Duc et al. (2022) sequenced 12 whole genomes of the extinct Steller's sea cow *Hydrodamalis gigas*, and found candidate genes for cold adaptation in marine environments. Lipoxygenase genes, which when inactivated cause a disease in humans characterized by thick, rough skin, are also inactivated in Steller's sea cow, potentially implicating these genes in their thick, bark-like skin. For a classic example of using ancient DNA to determine why extinctions may have occurred, we can turn to mammoths. There is an ongoing debate as to whether mammoths went extinct due to human impact, climate change, or a mix or the two, and many mammoths genomes have been sequenced to this end (eg. (Wang et al., 2021; van der Valk et al., 2021)). Another example can be found in the extinct passenger pigeon, which rapidly went extinct in the 19th century. This extinction was previously thought to be because of population instability, but its genome sequence appears to contradict these results (Murray et al., 2017).

Ancient DNA can inform and direct modern conservation practices in a number of important ways. First, DNA from ancient specimens can provide a valuable baseline of past genetic information and diversity, showing if, how quickly and even why populations have declined. This is especially valuable in the case of endangered or threatened species. Genetic diversity within a species is a recognized form of biodiversity according to the Convention on Biological Diversity, and showing its decline in a species can directly impact government policy (Jensen et al., 2022). For example, Mondol et al. (2013) found that most mitochondrial DNA variants in ancient tiger DNA samples are not present in modern tigers, leading them to implicate habitat loss due to humans in this loss of genetic diversity and highlighting the unsuitability of modern conservation policies for tigers. Similarly, ancient DNA can help to determine the extent of accumulated deleterious mutations or genetic load, and when applied to extinct species it could help clarify the relationship between extinction probability and genetic load (Bertorelle et al., 2022). Second, ancient DNA can be used to help resolve species delimitation questions, which is important because conservation

status is usually determined on a per-species basis. For example, in Palkopoulou et al. (2018), they used ancient DNA to justify the classification of forest and savanna elephants, *Loxodonta africana* and *Loxodonta cyclotis*, as separate species, leading to their conservation statuses to be updated to endangered and critically endangered, respectively. On the other hand, Mikheyev et al. (2017) used ancient DNA to confirm that an extinct population of stick insect was the same species as a modern population living on a nearby island, supporting recolonization efforts using the living population. Third, documenting past ranges and dynamics of populations, sometimes in relationship to paleoclimate data, can help forecast future movement and changes in response to a warming planet or habitat loss. This can in turn guide policies concerning which habitats to protect and where efforts should focus. For example, Casas-Marce et al. (2017) used ancient DNA from the endangered Iberian lynx to reconstruct past population dynamics, and showed that populations were much more connected and shared more gene flow thousands of years ago than their modern, structured subpopulation groups. This led them to suggest that modern individuals could be translocated between subpopulations as a conservation policy. Similarly Reynolds and Klavitter (2006) sequenced ancient DNA from critically endangered Laysan duck bones outside of their present-day range, which was used as evidence to support reintroductions.

Since microbial, human, or other sources of contamination often make up the majority of ancient samples (de Filippo et al., 2018; Ginolhac et al., 2011), any kind of ancient DNA can reasonably be viewed as a type of environmental DNA, and will share many of the same considerations. Like with eDNA, reads need be assigned to individual taxa in order to proceed with downstream analyses. Contaminating DNA can be present in the original sample or introduced throughout the sampling process. The latter type can be minimized with strict sampling and laboratory protocols, and both can be accounted for post-sequencing with bioinformatic steps, especially when high quality references are available and when the target taxon is already known, which is often the case in sequencing bones. Though protocols are not strictly standardized across groups (but see Fulton and Shapiro (2019)), DNA is usually extracted under strict conditions including UV radiation, filtered air systems, with personnel in body suits, shoe covers, masks, and gloves, and bleached instruments and surfaces (Slatkin and Racimo, 2016). Even the number of copies in a single tube of DNA that becomes aerosolized when opened can outnumber the amount of DNA in an ancient sample (Fulton and Shapiro, 2019), so extraction protocols should take place in a separate laboratory from amplification, ideally one that does not share air with the latter. Studies should also contain negative controls and multiple replicates, both in the original laboratory and ideally in a second, independent laboratory, among other considerations. Dedicated ancient DNA laboratories are expensive to create and maintain, and as of this year, ISOGG only records fewer than 30 in the entire world

(ISOGG, 2022).

DNA degrades over time, becoming fragmented and accumulating damage that would have been fixed by repair mechanisms in a live organism. Hydrolysis-induced deamination transforms unmethylated cytosine into uracil and methylated cytosine into thymine over time, and when this occurs on a single stranded 5' overhang, DNA polymerase introduces the complementary adenine (Gokhman et al., 2014). When sequenced, this type of damage appears as C-to-T miscoding transitions on the 5' termini of molecules and G-to-A miscoding transitions on the corresponding 3' termini (Fulton and Shapiro, 2019; Ginolhac et al., 2011). Another type of postmortem degradation is the removal of entire bases via hydrolytic depurination (Briggs et al., 2007). Other sources of damage include oxidation or cross-links, which can block polymerases. Overall, this means most sequenced ancient DNA segments are shorter than 100 base pairs and contain misincorporated C-to-Ts and G-to-As near their termini. The extent of fragmentation and damage is environment and age-dependent, and is affected by age, temperature, pH, and humidity (Fulton and Shapiro, 2019). Many ancient DNA studies have therefore focused on specimens from caves or permafrost areas, which provide more stable environments with good long-term preservation.

For the field of ancient DNA, deamination is not all bad. These C-to-T (and complementary G-to-A) misincorporations on the ends of ancient DNA fragments are so reliably present that they are almost ubiquitously used as proof of authenticity, since modern contaminants will not show these damage patterns. U-shaped graphs showing damage patterns occurring on the ends of reads (such as those produced by the software mapDamage (Ginolhac et al., 2011)) can be found in most ancient DNA studies nowadays. New ancient DNA extraction and library prep protocols are only considered to be reliable once authentication via damage patterns has taken place (Kapp et al., 2021). Identification of miscoding C-to-T transitions requires knowledge of the original sequence, and therefore cannot be done until reads have undergone filtering, quality control, and mapping to a high quality reference sequence, so even proving the authenticity of ancient DNA generally requires a bioinformatician. DNA damage can be used to our benefit in other ways as well. For example, the different transformations of methylated and unmethylated cytosine into thymine and uracil respectively has been exploited to reconstruct DNA methylation map of Neanderthals and Denisovans (Gokhman et al., 2014).

Ancient DNA has now been sequenced from many sources, including fossils (Sankararaman et al., 2014), resin (Peris et al., 2020), and ancient wood (Wagner et al., 2018). In 2003, the first ancient DNA was sequenced from an environmental source (Willerslev et al., 2003). This combination of the fields of ancient and environmental DNA is, naturally, referred to as ancient environmental DNA (aeDNA), or sometimes sedimentary ancient DNA (sedaDNA) when specifically

extracted from sediment.

1.3 Ancient environmental DNA

Ancient environmental DNA was first successfully extracted and sequenced in 2003 from permafrost cores in Siberia and a cave in New Zealand and mapped to over 40 taxa, yielding the first genetic records of these ancient environments (Willerslev et al., 2003). In the last two decades, aeDNA has proven to be an impressive tool with which to study ancient environments. The extraction of DNA from environmental sources means that it can be correlated with other variables such as paleoclimate and pollen records to give a more complete understanding of these ecosystems. Accordingly, ancient environmental DNA has been used for a wide variety of purposes so far, such as inferring a more recent extinction date than previously thought for mainland mammoths (Wang et al., 2021), uncovering how ancient ecosystems changed through time and responded to climate changes (Dussex et al., 2021), tracking correlations between species (Seeber et al., 2021), understanding human demography either through human DNA or through animal and plant proxies (Pedersen et al., 2016; Zavala et al., 2021), and more. Importantly, there is a possibility that DNA can persist longer when bound to molecules in environmental sources as opposed to in fossils, because DNA degradation enzymes could have reduced molecular recognition due to adsorption at mineral surfaces (Cai et al., 2006; Vandeventer et al., 2013). The number of publications including ancient environmental DNA is now more than 20 per year (Capo et al., 2021).

Sequencing ancient DNA from environmental sources can be particularly useful to detect and study rare species which may have not left many fossils, and to avoid the destruction of those fossils which do exist. In particular, the youngest fossil find of species need not represent its last surviving member, and so fossil dates can only be used as maximum age constraints for extinction or "latest appearance" dates. Ancient environmental DNA has already been used to extend the latest appearance date of ancient populations, such as the extinction time of the mammoth both in Alaska and in Northern Siberia (Wang et al., 2021). In the former case, this aeDNA evidence proved that the existence of mammoths in Alaska overlapped with human occupation, and that their extinction in the area could have been soon after the first human arrival. Fossils may also be rare in cases when species are newly populating an area, and so aeDNA may be better suited to resolve questions about migrations. For example, Gilbert et al. (2008) used coprolites from a cave in Oregon to show that humans must have been present in this area by $\sim 12,300$ thousand years ago. Some taxa, such as plants or bacteria, form fossils especially rarely compared to those from teeth or bones and which can be difficult to classify, and so are well-suited to analysis using aeDNA. Using aeDNA, Parducci et al. (2012) detected a rare mitochondrial haplotype of spruce trees in

Scandinavia, suggesting that these conifers survived in a refugia during the last glaciation, where it was previously thought they were locally extinct. Most importantly, aeDNA has the ability to detect genetic information from entire ecosystems at once, whereas fossils from many taxa are not usually found in the same place or time. Like modern eDNA, the comprehensive nature of aeDNA allows it to inform a detailed understanding of the evolutionary relationships between species. For example, Willerslev et al. (2014) used aeDNA across the Arctic to understand changes in vegetation and therefore megafaunal diet over time, concluding that Arctic megafauna would have eaten both forbes and graminoid plants.

Until recently, aeDNA studies have used polymerase chain reaction or PCR to amplify for specific marker genes or regions to reduce the complexity of the dataset and to make it easier to identify species using metabarcoding while minimizing sequencing costs. However, in the case of aeDNA, amplification is not necessarily ideal because contaminating fragments will be amplified as well, and primers may bind differently to DNA from specific species or by nucleotide content (Bell et al., 2021). Furthermore, chemical modifications to ancient DNA such as crosslinks or oxidation are not reliably replicated by traditional PCR polymerases, so that modern contaminating fragments can be amplified at a higher rate than ancient templates (Ginolhac et al., 2011; Fulton and Shapiro, 2019). PCR and metabarcoding is still used often in aeDNA studies, but shotgun sequencing without the need for PCR for aeDNA is becoming more feasible as sequencing costs are dropping. Shotgun sequencing was first applied to aeDNA data in 2016 (Pedersen et al., 2016), which used animal and plant DNA to study the viability of an ice-free corridor between the Laurentian and Cordilleran ice sheets for human occupation during the Pleistocene-Holocene transition. When enough species in the ecosystem have closely-related and complete reference genomes, shotgun sequencing is more powerful than metabarcoding as it will recover a more diverse set of taxa (Parducci et al., 2019). The small fraction of the genome captured by metabarcoding often contains insufficient variable sites to reliably carry out population genetic methods or related downstream analyses going beyond detection of individual taxa, whereas shotgun sequencing can recover a much larger set of loci. On the other hand, shotgun sequencing can generate more false positives than metabarcoding due to the overall genetic similarity of many species in an ecosystem, whereas barcodes are specifically designed to be able to separate taxa - although in certain species with low divergence, barcodes can be identical. Overall, Bell et al. (2021) found that shotgun sequencing correlates more strongly than metabarcoding to the actual genetic makeup of the environment.

Ancient environmental DNA has been recovered from permafrost, ice, surface soils, the beds of lakes and oceans, and more (Pedersen et al., 2015). For example, Figure 1 shows the recovery of ancient environmental DNA from sediment profiles in a cave in Northern Mexico (which is

analyzed in Chapter 4). One of the most common sources of ancient environmental DNA is sediment cores, which are cylindrical samples showing layers of stratigraphy going from oldest at the bottom to youngest at the top. Two examples of recovering sediment cores from lakes in different environmental conditions are shown in Figure 2, and an actual sediment core is shown in Figure 3. Long before their use in aeDNA, sediment profiles and cores have been used by geologists and paleoecologists to study the changes in climate, pollen, mineral deposition or charcoal over time at a given site. Vertical sediment cores from lakes are often favourable, since pollen, charcoal and DNA particles will wash into the lake from the nearby region and accumulate on the lake bed. For DNA specifically, lake beds typically provide reasonably good preservation, because sediments are compacted and cold. Especially in the winter, lakes can be frozen over, providing a platform on which to stand during coring (e.g. see Figure 2(b)). Geologists and paleoecologists can count pollen particles from each subsection of the core and compare these to carbon dates, yielding a detailed reconstruction of the plant taxa abundance over time. However, pollen can derive from sources many kilometers away (Pasquet et al., 2008), so that reconstructions of plant taxa from pollen give a more regional record, rather than a local one. Because of this, plant history reconstructions from pollen records do not necessarily correlate well with the local history of the region. Furthermore, this procedure is incredibly time consuming and pollen is difficult to identify from sediment without proper expertise. On the other hand, eDNA provides a less dispersed measure of plant abundance, and importantly, includes information regarding non-plant taxa such as animals and microbes, which can be related to each other in correlation analyses to understand ecosystem interactions. For example, both von Hippel et al. (2022) and Talas et al. (2021) uncovered long-term fungus-plant interactions using metabarcoding data from lake sediment cores. Recently, marine sediment cores from the Antarctic ocean which are potentially more than a million years old have been used to study diatom transitions (Armbrecht et al., 2022).

Obtaining and analyzing ancient DNA from environmental sources combines challenges of both aDNA and eDNA fields. It can require extra contamination prevention protocols both in the field and a dedicated ancient DNA lab, such as bleaching sediment core tubes before obtaining sediment, in addition to the protocols described above. Control samples must also be taken from modern eDNA in the environment to account for contamination in the field. Ancient environmental DNA consists of short, damaged fragments in low copy number, and often a lot of sequencing is necessary to obtain sufficient information about the sample, which can be expensive. It contains DNA from multiple individuals and taxa, some of which will be extinct or unknown, and many of which will not have reference genomes, so assigning reads to individual taxa can become even harder than in the modern case (discussed further in Chapter 2). Using time-series samples for ancient envi-

Figure 1: Collaborators collecting sediment samples for aeDNA in Chiquihuite cave, Northern Mexico, 2019. Photo by Devlin Gandy.

ronmental DNA also introduces the new issue of leaching, where DNA can detach from its original depositional layer into older layers, confounding the time stratification of the sample (Capo et al., 2021). Overall, this means aeDNA can require specialized algorithms and analyses, which often extend existing methods to account for these special considerations. For example, metaDamage extends the ancient DNA mapDamage algorithm, which identifies and visualizes damage patterns, to multiple taxa (Everett et al., 2021; Ginolhac et al., 2011). In the next section, we discuss the fields of phylogenetics and population genetics, why ancient environmental DNA might not lend itself easily to traditional approaches, and how one might extend or alter methods to account for aeDNA-specific considerations.

(a) Mineral Lake, Oregon, USA

(b) Unnamed lake, Northeast Iceland

Figure 2: Collecting lake sediment cores from diverse environments requires different techniques. (a) Coring a lake in Oregon in July 2022 required that we used a platform built on top of boats with a hole in the middle, and with anchor points to keep the platform still. (b) To core this frozen lake in Iceland in March 2022, the ice was sufficiently thick that we could simply drill a hole and stand on the ice, so that no platform was needed. Photos taken by Yucheng Wang and myself, respectively.

Figure 3: A lake sediment core I collected from Gordon Lake, Oregon, USA, in July 2022. Photo taken by myself.

1.4 Population genetics and phylogenetics in the context of aeDNA

Though species detection is useful in applications such as biodiversity monitoring, the potential of what we can do with ancient environmental DNA is vast. Once we have separated out reads from aeDNA samples into species or population-specific groups, we would like to learn about the evolutionary history of these species, asking questions such as: How has the genetic diversity of a species changed over time? Do ancient populations closely resemble modern ones, and to which modern populations are they most similar? Can we detect admixture or interbreeding between populations? Can we infer the existence of extinct clades or species, and correlate the timing to environmental or other variables to determine the cause of their extinction? Can aeDNA help us reconstruct the past movement patterns of species? Answering these questions requires the use of phylogenetics, which studies evolution on a phylogenetic tree, and population genetics, the study of the change in allele frequencies over time. Here, I overview foundational concepts in these fields and discuss their applications and challenges with respect to aeDNA.

Allele frequencies change due to a suite of evolutionary forces, including genetic drift, mutation, migration and selection. The first of these, genetic drift, describes the change in allele frequencies due to inherent randomness in the number of alleles passed on to offspring in the next generation. Genetic drift alone will inevitably lead to the eventual loss or fixation of an allele, reducing overall genetic variation in a population in the long term. This loss of variation due to drift is countered by mutations, which introduce genetic variation into a population over time. A base substitution is a change at a single nucleotide in the DNA sequence and is the most common form of mutation. The state of having more than one allele at a loci is called a polymorphism, and so the resulting state in the genome caused by a base substitution is often called a single nucleotide polymorphism or SNP. All other mutations can be broadly classed as structural, sometimes called genomic rearrangements, including deletions, insertions, translocations, inversions or duplications. Because of the relative ease of modelling and reliably detecting SNPs compared to structural mutations, and their higher frequency compared to other types of mutations, many population genetic and phylogenetic methods will rely only on SNPs to infer evolutionary histories or relationships between populations. Despite the widespread existence of selection in genomes, population genetic and phylogenetic methods often assume that mutations are neutral and do not affect the reproductive success of an individual, because this assumption greatly reduces mathematical complexity and is approximately true in many cases (Jensen et al., 2018).

Perhaps the most commonly used rigorous mathematical framework in population genetics was introduced in the 1920s and 30s. The simplest version of a Wright-Fisher model tracks the change in allele frequencies at a single biallelic site, in a population with simplifying assumptions such

as panmixia, constant size and non-overlapping generations (Fisher, 1923; Wright, 1931). In this model, the number of copies of a mutant allele in a generation is determined by a binomial draw with a mean of the allele frequency in the last generation. The mean of this draw can be biased by various factors, such as recurrent mutation or selection, where the selection coefficient of the mutant allele can be calculated by the ratio between the fitness of the allele and its alternate, minus one. Considering the Wright-Fisher model backwards in time, and tracing the evolution of individuals at a single timepoint backwards in their genealogy until their most recent common ancestor, gives rise to coalescent theory. Wright-Fisher and related forward time models are powerful tools in population genetics, often used to make theoretical predictions or in simulation frameworks (e.g. Krukov et al. (2017); Sanctis et al. (2017); Haller and Messer (2019)). Modern phylogenetics often relies on coalescent theory as a prior when reconstructing evolutionary trees from individual sequences, and is therefore intricately linked to population genetics.

Recombination, and the resulting tendency for inheritance to occur in genetic blocks or haplotypes, complicates both population genetic and phylogenetic models. Genome-scale population genetic models often assume that loci are independent and so can be modelled more or less separately when in reality sites may be under linkage disequilibrium with each other so that their allele frequency changes are correlated. This is sometimes circumvented by removing sites within a certain distance of each other, but often simply ignored. This is justifiable in many organisms because mutation rates and recombination rates are of the same order, meaning that on average, neighbouring variable sites are on different haplotype blocks. Phylogenetic trees, on the other hand, may differ for neighbouring haplotypes, resulting in phylogenetic incongruence or gene-species discordance. In particular, in a phenomenon called incomplete lineage sorting, individuals in a species may fail to coalesce with each other before coalescing with individuals from another species. For example, though chimpanzees are on average more closely related to humans than either are to gorillas, gorillas are more similar to humans or chimpanzees on 30% of the genome than the latter two are to each other (Scally et al., 2012). Larger population sizes will increase the amount of incomplete lineage sorting, which obscures phylogenetic signal and makes cross-species coalescent inferences significantly more difficult. There are a few ways to increase the chance of inferring the true species tree. A common approach is to obtain sufficiently many independent sites in the genome that phylogenetic inferences between species can be reliably made despite incomplete lineage sorting. Another is to use algorithms based on ancestral recombination graphs, which explicitly model recombination events and allows different phylogenetic trees for each inferred haplotype (e.g. Speidel et al. (2019)). This option typically requires high quality data to infer recombination events, and is not yet suitable for ancient environmental DNA. Similarly, the

multi-species coalescent model, which has been mathematically described but is not yet widely used, extends the single-species coalescent model and can be used to explicitly model incomplete lineage sorting (Jiao et al., 2021). Finally, we can only use mitochondria or chloroplast genomes, which have very little recombination and can reasonably be treated as a single, non-recombining haplotype. These genomes are also uniparentally inherited and so have lower population sizes, meaning their phylogenetic trees are more likely to reflect species trees. However, analyses based on the mitochondrial or chloroplast genome only provide a single estimate of the species tree, and these genomes can also be introgressed leading to gene-species tree discordance with much of the nuclear DNA.

Phylogenetic reconstruction can be done using parsimony, distance-based or likelihood/Bayesian methods. Maximum parsimony based phylogenies, which means choosing the tree which minimizes the total number of internal changes, is especially popular when using morphological rather than genetic data, since carefully chosen morphological changes are likely to follow this assumption (Sansom et al., 2018). However, parsimony is not usually the first choice for molecular phylogenetics based on sequence data. Distance based methods cluster sequences together which have the least differences under some metric. For example, neighbour joining is essentially a bottom-up clustering method which results in an unrooted tree. Since they are more computationally efficient than likelihood methods but also generally more simplistic, distance based methods are often used in molecular phylogenetics as a "first pass" or when a highly accurate phylogenetics is not strictly necessary. For molecular phylogenetics, perhaps the most option is to use likelihood-based or Bayesian methods. Unlike parsimony, likelihood-based methods can account for the possibility of different substitution rates for different branches or substitution types, for multiple substitutions on a branch, and for various sources of uncertainty. In general, in phylogenies with more evolutionary change, parsimony-based approaches can be misleading and prone to long branch attraction, a phenomenon in which longer branches are placed erratically in the tree due to convergent evolution or the failure to account for multiple substitutions (Parks and Goldman, 2014; Felsenstein, 1978). Many standard software packages implement Bayesian phylogenetic reconstruction algorithms. MrBayes, released in 2001, widely popularized the use of Bayesian methods for phylogenetic reconstruction using MCMC (Huelsenbeck and Ronquist, 2001), and has been updated many times since. Another commonly used software, BEAST, allows a wide range of user-defined parameters such as specific demographic models (Suchard et al., 2018).

Likelihood-based phylogenetic reconstruction algorithms need to compute the likelihood of individual phylogenetic trees given input sequence data. To do this, they rely on a method called Felsenstein's pruning algorithm, first described by Felsenstein (1973). Felsenstein's pruning algo-

rithm is a specific implementation of belief propagation or message passing, an algorithm which infers the marginal distribution of unknown nodes in a graphical model given data about observed or known nodes. In general graphical models, it is approximate, but phylogenetic trees are always connected acyclic graphs, and on this subset of models, belief propagation is exact and relatively fast. Felsenstein's algorithm, better described in Chapter 3, relies on an underlying substitution model specifying the rates of change between nucleotides. Phylogenetic trees are called ultrametric if the distance from the root to each tip is equal.

An especially relevant application of phylogenetics to aeDNA data is molecular dating, or the estimation of divergence times or sequence ages using phylogenetic trees which have been calibrated with known generation lengths and mutation rates or by using the fossil record. Essentially, this boils down to phylogenetic tree reconstruction with external calibration to transform branch lengths into real time estimates like years. Molecular dating was first proposed in 1962 using a strict molecular clock, in which mutations on all branches and at all sites are assumed to occur at a constant rate (Sauquet, 2013). Since this assumption does not hold in general (e.g. see Arcones et al. (2021)), techniques have been updated since to use a relaxed molecular clock and rates inferred from sequence data, which can vary per site, per mutation type, per branch and over time (Drummond et al., 2006). Calibration should be done using known species-specific substitution rates where possible, such as those measured from existing populations (though substitution rates can and do change over longer timescales, see e.g. Arcones et al. (2021)). However, these are not generally known, and sometimes the point of molecular dating is to infer the rate of substitution, not the other way around. Node constraints can be given prior distributions when certain aspects of the topology or node ages are known beforehand. For example, the oldest known fossil from a species could be used as a prior to constrain the timing of its species divergence. Geological, morphological or past climate evidence can provide age constraints when they can be assumed to have shaped past speciations, such as by geographical isolation. For example, eudicots are often rooted with a maximum age constraint of 125 million years because tricolpate pollen, one of their defining characteristics as a group, is absent from the record before this point (Sauquet, 2013). Output age estimates from molecular dating will have posterior distributions and are normally reported as high posterior density intervals.

Molecular dating is useful in many ways. First of all, the inference of divergence times, or estimating the age of internal nodes in the phylogenetic tree, has greatly informed our understanding of evolutionary history and its relationship to past geological and climate changes. For example, Green et al. (2008) used the first Neanderthal mitochondrial sequence to obtain the first molecular date of the divergence of Neanderthals and modern humans. In fact, molecular dating is so

ubiquitously used in evolutionary studies that there is an entire database dedicated to documenting the timescale of the evolution of the tree of life, containing more than 4,000 studies (Kumar et al., 2022), which allows the user to query any pair of taxa and obtain an estimate of their divergence time. Inferring divergence times allows the estimation of time-dependent rates of speciation events, diversifications, and extinctions, which can be analyzed in comparison to each other or jointly with environmental factors. For example, molecular dating has been used extensively during the coronavirus pandemic to understand the emergence of new clades and to inform debates regarding the place and timing of the origin of SARS-CoV2 (Roberts et al., 2021). dos Reis et al. (2012) used molecular dating to confirm the fossil-record based conclusion that placental mammals radiated quickly following the Cretaceous-Paleogene extinction 65 million years ago. Gire et al. (2014) sequenced Ebola virus genomes from 78 patients in Africa to characterize intrahost and interhost mutations and transmission dynamics over a decade. Perhaps one of the most relevant applications of molecular dating for aeDNA is the ability to estimate the age of individual sequences in the context of a phylogenetic tree. This can act as an independent and additional form of validation for ancient DNA studies, especially those using material which is too old to reliably carbon date. To validate the date of the oldest DNA sequenced so far, for example, van der Valk et al. (2021) sequenced one million year old mammoth DNA from teeth, and confirmed consistency of their dates using molecular dating in BEAST.

Populations do not always evolve so cleanly as to fit a phylogenetic tree. Admixture, or the mixing of distinct genetic populations, affects the genetic makeup of one or more populations involved by sharing alleles across populations. Introgression refers to a type of one-way admixture where an external population donates genetic material to a source population, such as the introgression of Neanderthals into humans (Reilly et al., 2022). Admixture and introgression are extensive in humans, especially stemming from human migrations in the past, and can occur both in single pulse events such as large population movements, or in a continuous way such as when two neighbouring populations share genetic material over time (Reich et al., 2009). Both admixture and introgression will complicate the inference of phylogenetic trees, since any given phylogeny will not adequately fit the data. For a given set of populations and topology, we can attempt to evaluate the compatability of a phylogenetic tree model with no admixture by using aptly named statistical tests of treeness. Amongst the most commonly used treeness tests are F_2, F_3 and F_4 statistics, which measure allele frequency correlations between sets of populations, and correspond to shared genetic drift between paths on a tree (Reich et al., 2009; Patterson et al., 2012). For a given phylogeny, then, significant deviations from the expected values of these statistics indicates a poor fit. For example, a significantly negative F_3 statistic can suggest admixture, or a significantly

nonzero F_4 statistic suggests non-independence of the populations in consideration (Peter, 2016). When data fails tests of treeness, we can explicitly model admixture events via a more general model called an admixture graph, which adds extra edges between existing branches of a phylogenetic tree to account for admixture. Admixture edges are weighted by the proportion of the genetic material coming from each of the two source populations, and yield a new phylogenetic branch for the resulting admixed population. However, admixture graphs suffer from their requirement that admixture events occur at discrete times rather than in continuous waves, and are prone to overfitting (Maier et al., 2022). The latter can be controlled for by considering likelihood scores that penalize higher numbers of admixture edges.

For low quality data such as ancient environmental DNA, which is often damaged and only covers a fraction of the genome, traditional phylogenetic reconstruction methods, including admixture graphs, are not always appropriate. First of all, samples can contain DNA from multiple individuals in the environment and potentially from different related populations or species. These mixed samples are impossible to place at just one point in a phylogeny, because phylogenetic reconstruction algorithms assume that each sample is a single individual. Even without a mixed population, issues remain. Errors in variant calling due to deamination or population variation may be interpreted as private variants, resulting in sequences being incorrectly placed externally to their actual clade. Many phylogenetic algorithms require significant overlap of sequences and throw out missing sites, which are often in a high proportion in aeDNA. Clearly, there are many missing sites when data are low coverage, but even samples with a lot of sequence data can contain missing sites. For example, this could occur from the inability to reliably call variants due to inconsistencies from damage. Excessive missing data in individual sequences during phylogenetic reconstruction or molecular dating can exacerbate existing biases, such as by lessening the signal of multiple substitutions on long branches (Roure et al., 2012). Missingness in aeDNA data can also be a result of the upstream step in which we map reads against a reference database. Highly conserved regions of the genome will be shared among reference sequences, and so in these regions it can be impossible to assign aeDNA reads to individual species. This also means missingness in aeDNA samples following this mapping procedure is not random along the genome, but instead distributed preferentially in nonconserved regions, which could make private branches appear longer than they really are.

A high amount of missing data can also pose a problem to principal component analysis, a dimensionality reduction technique used often as an initial step to infer relationships between different groups or populations. Principal component analysis, or PCA, is a linear transformation of high dimensional data to a different coordinate basis, so that in the new basis the first dimension captures the most possible variance. Similarly, the second dimension represents an orthogonal di-

rection which captures the most possible of the remaining variance, and similarly for the following dimensions. The results are visualized in a reduced amount dimensions and can illuminate structure in the data. PCA is often used as an initial exploratory step because of its minimal need for assumptions. For genetic data, the input to PCA is generally many individuals or populations with thousands of polymorphic sites represented in high dimensional space by assigning a dimension to each locus and by assigning the individual or population in that dimension by its genotype (0, 1 or 2 in diploids) or population allele frequency. Often sites whose genotype is missing for any individuals are deleted prior to a PCA analysis (Yi and Latch, 2021), so aeDNA data with a high degree of missingness can present a problem to standard methods. One solution is to impute missing sites Alternatively if these sites are left as missing, this can bias results by dragging points with more missing data away from their real population groups and towards the origin in a phenomenon sometimes called "shrinkage" (Yi and Latch, 2021; François and Jay, 2020). PCA can also be biased by different sample dates (temporal bias) or by spatially correlated data, the latter yielding a so-called "horseshoe" artefact in which the 2D visualization appears to have a bowed shape (Skoglund et al., 2014; François and Jay, 2020). Some of these issues can be addressed by using a projection approach such as implemented in Price et al. (2006), in which the initial PCA is computed using only the modern or high-coverage samples, and the ancient samples are projected afterwards. Though projection ignores variation which is present in only the ancient samples, it is often preferable to the biases and artefacts which would otherwise be introduced.

In general, traditional phylogenetic and population genetic methods encounter new challenges when faced with fragmented, damaged and low-coverage data such as ancient environmental DNA. However, with the ability to shotgun sequence more total fragments, a better understanding of sources of error including degradation, contamination-minimizing protocols, more reference genomes and new computational techniques, we can now begin to overcome these issues, as shown in this present work.

1.5 Outline

The remainder of this thesis is organized as follows. The second chapter presents a simplified, mathematical framework to estimate error from population genetic and coalescent sources in supervised binning algorithms, which map environmental DNA reads against reference databases in order to assign reads to individual taxa. The third chapter describes a phylogenetic Bayesian placement algorithm which can overcome many challenges associated with ancient environmental DNA in phylogenetics, and applies this algorithm to an Arctic-wide dataset from the last 50,000 years, which includes mammoth and horse sequences. The fourth chapter describes a phylogenetic

and population analysis of two closely related species, the American black bear and the extinct giant short-faced bear, using aeDNA from 14-16,000 year old cave soil. Finally, the fifth chapter infers a molecular date for a birch tree aeDNA chloroplast sequence, which was determined by non-genetic geological and chemical dating techniques to be approximately 2 million years old. The thesis concludes with a short conclusion chapter.

Chapters 2-5 contain material from published papers. In each case, I give the relevant publications at the beginning of the chapter, and only include work that I did myself unless explicitly stated otherwise.

2 A Theoretical Analysis of Taxonomic Binning Accuracy

This chapter has been published: De Sanctis, B., Money, D., Pedersen, M. W., and Durbin, R. (2022). A theoretical analysis of taxonomic binning accuracy. *Molecular Ecology Resources*, 22(6):2208–2219.

Many metagenomic and environmental DNA studies require the taxonomic assignment of individual reads or sequences by aligning reads to a reference database, known as taxonomic binning. When a read aligns to more than one reference sequence, it is often classified based on sequence similarity. This step can assign reads to incorrect taxa, at a rate which depends both on the assignment algorithm, and on underlying population genetic and database parameters. In particular, as we move towards using environmental DNA to study eukaryotic taxa subject to regular recombination, we must take into account issues concerning gene tree discordance. Though accuracy is often compared across algorithms using a fixed dataset, the relative impact of these population genetic and database parameters on accuracy has not yet been quantified. Here, we develop both a theoretical and simulation framework in the simplified case of two reference species, and compute binning accuracy over a wide range of parameters, including sequence length, species-query divergence time, divergence times of the reference species, reference database completeness, sample age, and effective population size. We consider two assignment methods, and contextualize our results using parameters from a recent ancient environmental DNA study, comparing them to the commonly used discriminative k-mer based method Clark (Pedersen et al., 2021; Ounit et al., 2015). Our results quantify the degradation in assignment accuracy as the samples diverge from their closest reference sequence, and with incompleteness of reference sequences. We also provide a framework in which others can compute expected accuracy for their particular method or parameter set. Code is available at https://github.com/bdesanctis/binning-accuracy.

2.1 Introduction

Environmental DNA analyses require the assignment of individuals reads to reference taxa, which is known as supervised or taxonomic binning. This process can be confounded by local gene tree variation present in recombining systems, as is the case for most eukaryotic taxa. Since some reads will align to more than one reference sequence, these methods necessarily include a decision step on whether to assign these reads, and how. Often this is done based on sequence similarity. For example, one could assign the query read to its "closest" reference sequence by choosing the assignment that minimizes mismatches between the query and the reference taxon, which we call the "least-mismatch" method (Prüfer et al., 2010; de Filippo et al., 2018; Kircher, 2011). A more conservative approach is to only assign the query to a reference taxon when it aligns to a reference sequence with no mismatches, which we call the "exact-match" method (Key et al., 2017; Lammers et al., 2021; Pedersen et al., 2016). However, this latter approach will fail to assign query reads that differ from the reference sequence, rendering downstream population genetic analyses that rely on these differences, such as admixture or demography, ineffective (Kircher, 2012).

Many dedicated algorithms for taxonomic binning exist. MALT or MEGAN (Herbig et al., 2016) competitively maps to a user-defined database then employs a lowest common ancestor algorithm. HOPS uses modified alignment parameters to account for damage in ancient DNA (Hübler et al., 2019). SPARSE aligns to clusters of reference genomes (Zhou et al., 2018). Pathoscope assigns microbial reads and removes reads that are similar to a set of filter genomes (Hong et al., 2014). Clark and Kraken use a discriminative k-mer based approach to assign reads to a reference database (Ounit et al., 2015; Wood et al., 2019). BLAST is also used in this context, even though it was not originally designed for the purpose of binning (Altschul et al., 1990). However, since the premise of taxonomic binning is straightforward, many studies forgo specialized software and align their query sequences to a reference database themselves, e.g. using bwa (Li and Durbin, 2009) or bowtie2 (Langmead and Salzberg, 2012), then assign query reads to species using least-mismatch or exact-match as described above or something similar (Warinner et al., 2017; Prüfer et al., 2010; Feuerborn et al., 2020; Key et al., 2017; de Filippo et al., 2018; Anari, 2020). Common choices of reference database are NCBI Genbank (Clark et al., 2015), Refseq (O'Leary et al., 2015), Ensembl (Howe et al., 2020), or a curated set of reference sequences built to suit the metagenomic dataset in consideration.

In any of these methods, a decision may leave some query reads unassigned, such as those that align with no mismatches to multiple reference taxa. Furthermore, in many scenarios population variation may mean that the reference from a more distant taxon is actually more similar in the region of the query read, resulting in an erroneous assignment. The error rate resulting from this is

influenced by a number of parameters, including the length of the query sequences, the divergence between the query species and its closest reference species (Prüfer et al., 2010), the divergence between related reference species (Brown et al., 2015), and coverage or completeness of the closest reference species in the database (Warinner et al., 2017). Divergence between the query and the reference sequence will decrease accuracy as reads will be less likely to match the reference sequence of the correct species. Using reference species that are too close to each other can actually lead to a negative effect on binning accuracy (Brown et al., 2015), although this can be overcome by mapping to molecular operational taxonomic units (MOTUs), where a unit is defined based on some fixed sequence similarity or clustering algorithm, or by using a lowest common ancestor approach when reads are assigned to multiple species. An incomplete or low-coverage reference sequence that does not contain every position of the genomes will cause reads to be assigned to the closest related species instead, leading to incorrect assignments (Warinner et al., 2017).

Though it is widely understood that parameters concerning the read, database or species in question can improve or reduce the accuracy of this binning step, recommendations and practices for how to cope with this differ among the literature. For instance, filters or criteria in existing studies can be based on sequence similarity (e.g. 85% in (Krause-Kyora et al., 2018), 95% in (Willerslev et al., 2014) and (Haile et al., 2009), 100% (Pedersen et al., 2016)), alignment score, total percentage of identifiable sequences assigned to that species (e.g. 1% (Slon et al., 2017)), read length, divergence between reference species (e.g. 3% in (Brown et al., 2015)) and more.

To our knowledge, the exact relationships of these parameters to binning accuracy have not been quantified, either in relative or absolute terms (although see (Nielsen and Matz, 2006) for a related study). Because of this, it may be difficult to know which filtering or database construction steps or methods to use and prioritize in practice. Here, we use a two-species coalescent model to quantify the effects of relevant database, population genetic, or read parameters on the accuracy of the binning step. We consider a simplified case for which analytical results can be obtained under a coalescent model (Fig 4). We consider three species or populations: one for the "query" species from which the query sequence is sampled, one for the "true" closest species represented in the reference database, and one for the "false" species which is the next closest in the reference database. For simplicity we use the term "species" here in all three cases, but each may equally well be a subspecies, population or other genetically mixing taxonomic group.

Our model includes parameters not usually mentioned in this context, such as the age of the sample, which can impact accuracy when comparing to a present day reference sequence. In total, we consider how sequence length, species-query divergence time, reference species divergence time, reference species completeness, sample age, and effective population size impact the accu-

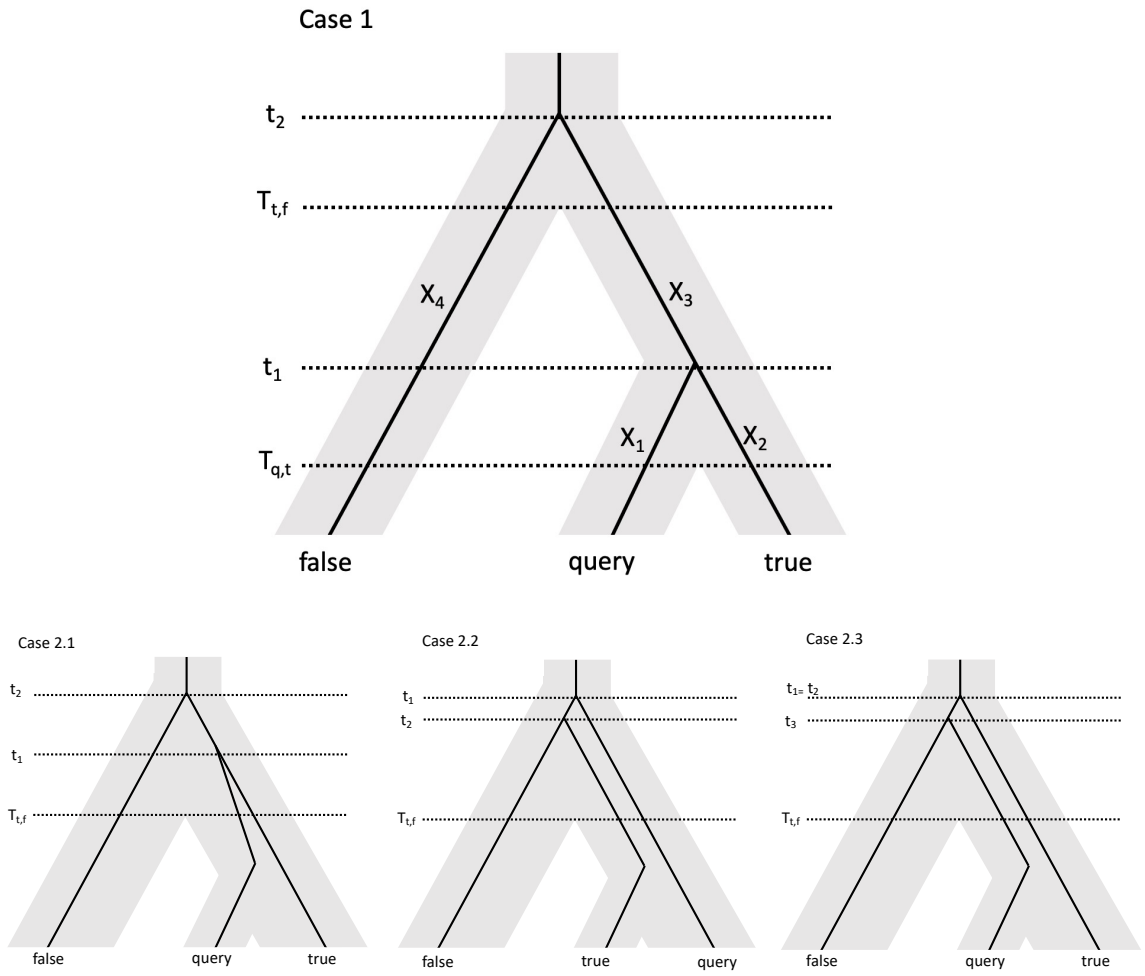


Figure 4: An illustration of the four possible scenarios. Here, $T_{t,f}$ represents the divergence time between the true and false species, and $T_{q,t}$ represents the divergence time between the query and true species, the latter of which is irrelevant in the bottom three cases. Note the change in order of the branches in case 2.2 and 2.3. See Methods for details.

racy of the taxonomic identification of a single sequence. Although we use a two-species database and consider only two assignment methods for simplicity and exactness, we believe that the conclusions from our model concerning the relative impact of parameters on accuracy can be applied more generally to inform an understanding of error rates in analyses using larger databases or other binning algorithms. In particular, because the errors arise from genuine overlap between the patterns of similarity of the query to true or false taxa, we expect the relative impact on the error rate of population genetic parameters that affect this overlap to be similar regardless of the algorithm or reference database size.

This problem of assigning individual sequences to taxa is not a problem isolated to the field of environmental DNA, but also appears in ancient DNA studies, where a large fraction of reads collected from a fossil can originate from microbial contamination (de Filippo et al., 2018). However, in ancient DNA studies, even with a high level of expected binning accuracy, additional criteria need to be met to ensure the authenticity of the reads. This includes, for example, comparing edit distance distributions to related species and confirming signs of ancient DNA damage such as deamination, and is covered in depth elsewhere (Orlando et al., 2021; Renaud et al., 2019).

2.2 Materials and Methods

Our analysis will start with the case where the query and true species are the same, that is where the query sample derives from the same panmictic population from which the reference sequence for the true species was obtained. We then go on to extend to the case where there was some divergence between the populations from which the query and the true reference sequence were sampled, which is typical in practice. We also first consider the case when all individuals are from the present day and the two references are complete, but expand on all of these assumptions in the following sections. We note that this approach assumes a single reference sequence for each species in the reference database - there are more complex scenarios that use multiple reference sequences within a single species or other taxonomic unit.

From here on, we refer to the query sequence as q, the true sequence as t, and the false sequence as f. We show calculations for the least mismatch method, where the correct assignment probability is the probability that the number of mutations between the query and the false sequence is greater than the number of mutations between the query and the true sequence. Denote the number of mutations between the query and the true sequence as the random variable K_t, and the number of mutations between the query and the false sequence as the random variable K_f. Notably, these two variables are not independent. We then have in the least-mismatch method $P(K_t < K_f)$ as

the probability of correct assignment, $P(K_t > K_f)$ as the probability of incorrect assignment, and $P(K_t = K_f)$ as the probability of no assignment. First, we will calculate the probability of correct assignment.

Denote the divergence time of the true and false species as $T_{t,f}$, and the divergence time of the true and query species as $T_{q,t}$, both in generations. We first assume that $T_{q,t} = 0$ and generalize this below. There are two cases to consider: either the query coalesces with the true sequence before time $T_{t,f}$, or after. The latter is possible in the case of incomplete lineage sorting, and often occurs with negligible probability unless the two reference sequences are closely related. An illustration of the possible coalescent scenarios is shown in Figure 4. For the remainder of this section, we use $T = T_{t,f}$ for ease of reading.

Case 1. The query coalesces with the true sequence before time T. This happens with probability $1 - e^{-T/2N}$, where N is the effective population size of the population q and t are drawn from. In this case, let $t_1 < T$ be the coalescent time for q and t, and let $t_2 > T$ be the coalescent time for q and f. Let X_1 and X_2 be the number of mutations on the branches from q and t respectively to the common ancestor of q and t, X_3 the number of mutations on the branch between the common ancestor of q and t to the root (the most recent common ancestor of q, t and f), and X_4 be the number of mutations on the branch between f and the root. This scenario is shown in the top of Figure 4, with $T_{q,t} = 0$ (the generalization for $T_{q,t} \neq 0$ is derived below). We want to compute

$$\begin{aligned} P(K_t < K_f) &= P(X_1 + X_2 < X_1 + X_3 + X_4) = p(X_2 < X_3 + X_4) \\ &= P(\text{Pois}(\mu t_1) < \text{Pois}(\mu(2t_2 - t_1))) \end{aligned}$$

given that $t_1 < T \leq t_2$, since we expect that the number of mutations on each branch is Poisson distributed. Let $A = \text{Pois}(\mu(2t_2 - t_1))$ and $B = \text{Pois}(\mu t_1)$ where μ is the mutation rate of the sequence per generation, the product of the per base mutation rate and the match length. Then, given t_1 and t_2, we want

$$\begin{aligned} P(B < A) = P(A \geq B) &= \sum_{k=0}^{\infty} P(A > B; B = k) P(B = k) = \sum_{k=0}^{\infty} P(A > k) P(B = k) \\ &= \sum_{k=0}^{\infty} \left(\sum_{l=k+1}^{\infty} \frac{\lambda_A^l e^{-\lambda_A}}{l!} \right) \frac{\lambda_B^k e^{-\lambda_B}}{k!} = \sum_{k=0}^{\infty} \left(\sum_{l=k+1}^{\infty} \frac{(\mu(2t_2 - t_1))^l e^{-\mu(2t_2 - t_1)}}{l!} \right) \frac{(\mu t_1)^k e^{-\mu t_1}}{k!} \\ &= \sum_{k=0}^{\infty} \sum_{l=k+1}^{\infty} \frac{(\mu(2t_2 - t_1))^l (\mu t_1)^k e^{-\mu(2t_2)}}{l!\, k!} \end{aligned}$$

Since t_1 and t_2 are unknown, we take the expectation over t_1 and t_2 given $t_1 < T \leq t_2$. That

means, in the case that the query coalesces with the true sequence before time T, the probability of correct assignment as a function of T, μ and N is

$$P_{CA}(T, \mu, N \,;\, t_1 < T < t_2)$$
$$= \int_T^\infty \int_0^T \frac{e^{-t_1/2N}}{2N} \frac{e^{-t_2/2N}}{2N} \left(\sum_{k=0}^{\infty} \sum_{l=k+1}^{\infty} \frac{(\mu(2t_2 - t_1))^l (\mu t_1)^k e^{-\mu(2t_2)}}{l!\; k!} \right) dt_1 dt_2 \Big/ \left(e^{-T/2N} \left(1 - e^{-T/2N} \right) \right)$$

where the first two exponentials are for the distributions of t_1 and t_2 respectively, and where it is assumed that the effective sizes of the population consisting of q and t before time T and the ancestral population to q, t and f after time T are equal and of value N. The denominator is obtained by integrating the numerator without the double summation inside and acts as a normalizing constant.

Case 2. The query does not coalesce with the true sequence before time T. This happens with probability $e^{-T/2N}$. In this case, we essentially have 3 sequences at time T which have not coalesced with each other, and therefore three subcases depending on which two coalesce first, each of which will happen with probability $1/3$.

Case 2.1. $T \leq t_1 < t_2$, so that the query and true sequence coalesce first (but after time T). In this case, we can make a similar argument to Case 1, and simply have to change the denominator to the relevant domain, and the limits on the integral, to get

$$P_{CA}(T, \mu, N \,;\, T \leq t_1 < t_2) =$$
$$= \int_T^\infty \int_{t_1}^\infty \frac{e^{-t_1/2N}}{2N} \frac{e^{-t_2/2N}}{2N} \left(\sum_{k=0}^{\infty} \sum_{l=k+1}^{\infty} \frac{(\mu(2t_2 - t_1))^l (\mu t_1)^k e^{-\mu(2t_2)}}{l!\; k!} \right) dt_2 dt_1 \Big/ \left(\frac{1}{2} e^{-T/N} \right)$$

Again, the denominator here is obtained by integrating the numerator without the double summation inside.

Case 2.2. $T \leq t_2 < t_1$, so that the true and false sequence coalesce first. In this case, the phylogenetic distance between the query and the true sequence is the same as the phylogenetic distance between the query and the false sequence. Since the mutations on the query branch will be the same in both cases, we will have $P(K_t < K_f)$ simply when the number of mutations on the true branch is less than the number of mutations on the false branch. Both of these branches have length t_2, so we would like the probability that one Poisson process with mean μt_2 is bigger than another which is identically and independently distributed. Given t_2, this will be

$$\sum_{k=0}^{\infty} \left(\sum_{l=k+1}^{\infty} \frac{(\mu t_2)^l e^{-\mu t_2}}{l!} \right) \frac{(\mu t_2)^k e^{-\mu t_2}}{k!} = \sum_{k=0} \sum_{l=k+1} \frac{(\mu t_2)^{l+k} e^{-2\mu t_2}}{l!\; k!}$$

And now, as usual, we integrate over the relevant domain to get

$$P_{CA}(T, \mu, N \,;\, T \leq t_2 < t_1) = \int_T^\infty \int_{t_2}^\infty \frac{e^{-t_1/2N}}{2N} \frac{e^{-t_2/2N}}{2N} \sum_{k=0}^{} \sum_{l=k+1} \frac{(\mu t_2)^{l+k} e^{-2\mu t_2}}{l!\; k!} dt_1 dt_2 \Big/ \left(\frac{1}{2} e^{-T/N}\right)$$

Case 2.3. $t_2 = t_1 > T$, so that the query and the false sequence coalesce first. This is possible due to incomplete lineage sorting, but only if it happened at a point older than T. Let t_3 be the time at which the query and the false sequence coalesce. We want the probability that a Poisson process with mean $\mu(2t_1 - t_3)$ is less than one with mean μt_3, which for a given t_1 and t_3 is

$$\sum_{k=0}^{\infty} \left(\sum_{l=k+1}^{\infty} \frac{(\mu t_3)^l e^{-\mu t_3}}{l!} \right) \frac{(\mu(2t_1 - t_3))^k e^{-\mu(2t_1 - t_3)}}{k!} = \sum_{k=0}^{\infty} \sum_{l=k+1}^{\infty} \frac{(\mu t_3)^l (\mu(2t_1 - t_3))^k e^{-\mu 2t_1}}{l!\; k!}$$

Integrating over the relevant domain gives

$$P_{CA}(T, \mu, N \,;\, t_2 = t_1 < T) = \int_T^\infty \int_{t_3}^\infty \frac{e^{-t_1/2N}}{2N} \frac{e^{-t_3/2N}}{2N} \sum_{k=0}^{\infty} \sum_{l=k+1}^{\infty} \frac{(\mu t_3)^l (\mu(2t_1 - t_3))^k e^{-\mu 2t_1}}{l!\; k!} dt_1 dt_3 \Big/ \left(\frac{1}{2} e^{-T/N}\right)$$

In summary, the probability of correctly assigning the query sequence to the true sequence is

$$\begin{aligned} P_{CA}(T, \mu, N) = (1 - e^{-T/2N}) P_{CA}(T, \mu, N \,;\, t_1 < T < t_2) + \\ \frac{e^{-T/2N}}{3} \big(P_{CA}(T, \mu, N \,;\, T \leq t_1 < t_2) + P_{CA}(T, \mu, N \,;\, T \leq t_2 < t_1) \\ + P_{CA}(T, \mu, N \,;\, t_2 = t_1 < T) \big) \end{aligned}$$

However, in many cases $T/2N \gg 1$ so that the latter three terms are negligible and we can obtain a sufficiently good approximation by only calculating the first term.

Similar calculations give the results for the "exact-match" method, the probability of incorrect assignment, and the probability of not making an assignment. The latter occurs when a query matches both reference sequences equally well in the "least-mismatch" method, or when it matches both equally well or does not match either with zero mismatches in the "exact-match" case. We compute the probabilities of correct, incorrect, and no assignment, while varying the input parameters and assignment method, as shown in Figure 5 and 6. For very high population sizes such as in Supplementary Figure 1, terms in the integrand can become small enough to be prone to errors, and

so we instead scaled time appropriately to compute the equivalent probabilities with smaller population sizes (ie. divided the effective population size, multiplied the mutation rate, and divided all divergence times by a fixed scaling factor). Code is available at https://github.com/bdesanctis/binning-accuracy.

2.2.1 Extensions to the Model

Next we can generalize this to include ages of each of the sequences if they are not contemporaneous, and to include a nonzero divergence time between the query and true species. The latter is difficult to escape at a small scale, because in most applications the query will not be expected to perfectly derive from the true reference species (in the coalescent sense we are using here), but as we shall see, a small level of divergence will not significantly impact the assignment probability. However, this can also present itself on a larger scale when the query species is not in the reference panel at all, but assignments are made to the closest relevant species. This applies especially when the query represents a new, extinct, or previously unsequenced species.

Let $T_{q,t}$ denote the coalescence time between the true and query species or populations. For clarity, we will rename the coalescence time of t and f species, previously called T in the above section, to $T_{t,f}$. We then require $T_{q,t} < T_{t,f}$. Times which are variables will remain denoted by lowercase ts, with appropriate subscripts.

Denote the age of each of the individuals as A_q, A_t and A_f, measured in generations. There are constraints on these ages with regard to the coalescence times for consistency. That is, we require $A_q, A_t < T_{q,t}$ and $A_f, A_t < T_{t,f}$.

Incorporating these two generalizations into each case is fairly straightforward. In case 2, $T_{q,t}$ is irrelevant, and so generalizing these is a simple matter of building in ages. This gives

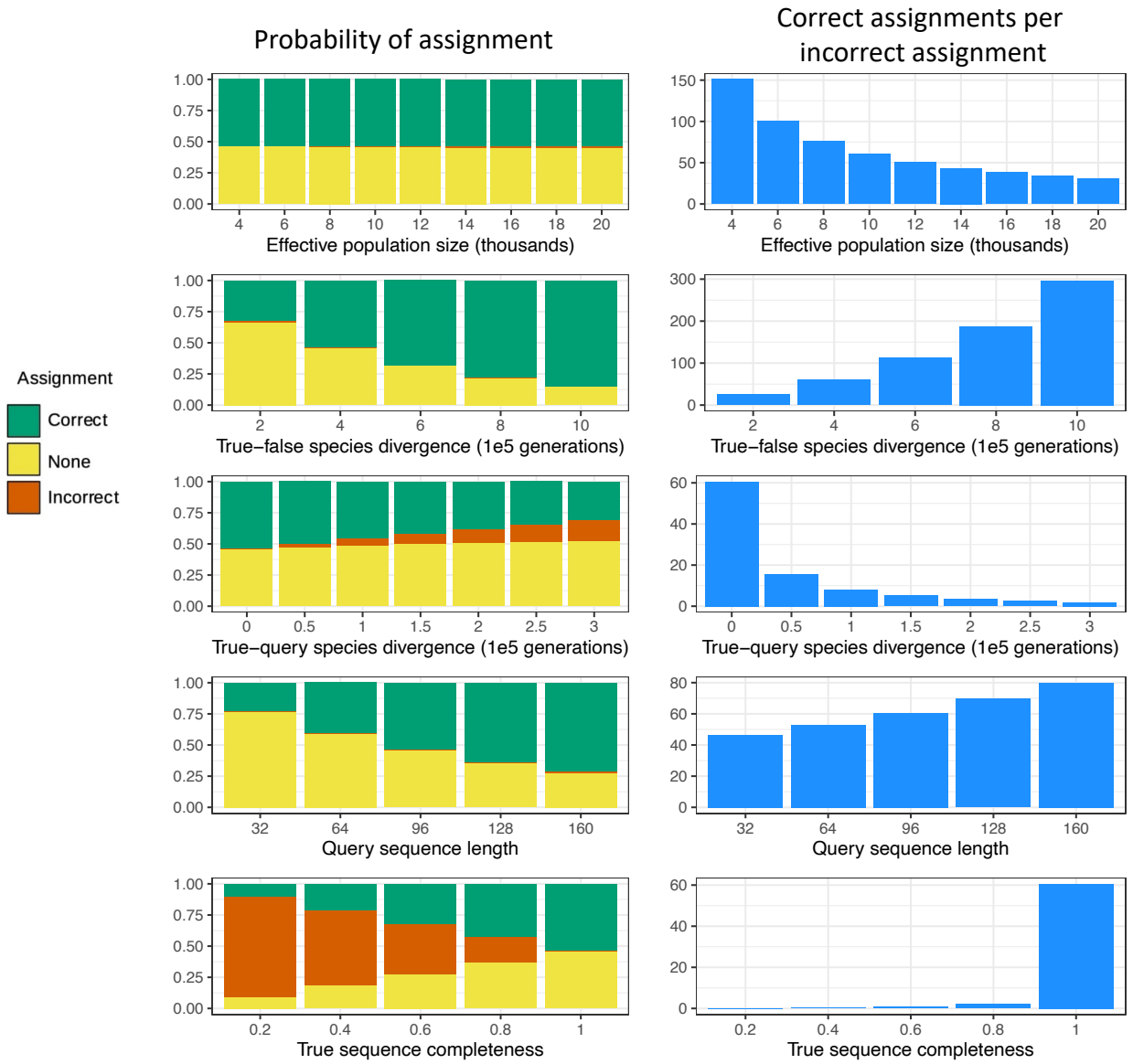


Figure 5: Left: Probability of assigning the query sequence correctly (to the true reference sequence) in green, of assigning the query sequence incorrectly (to the false reference sequence) in red, and of making no assignment in yellow, using the least mismatch method. Right: The expected number of correct assignments (to the true reference sequence) made for every one incorrect assignment (to the false reference sequence). In both, each row varies a different parameter while keeping the others constant.

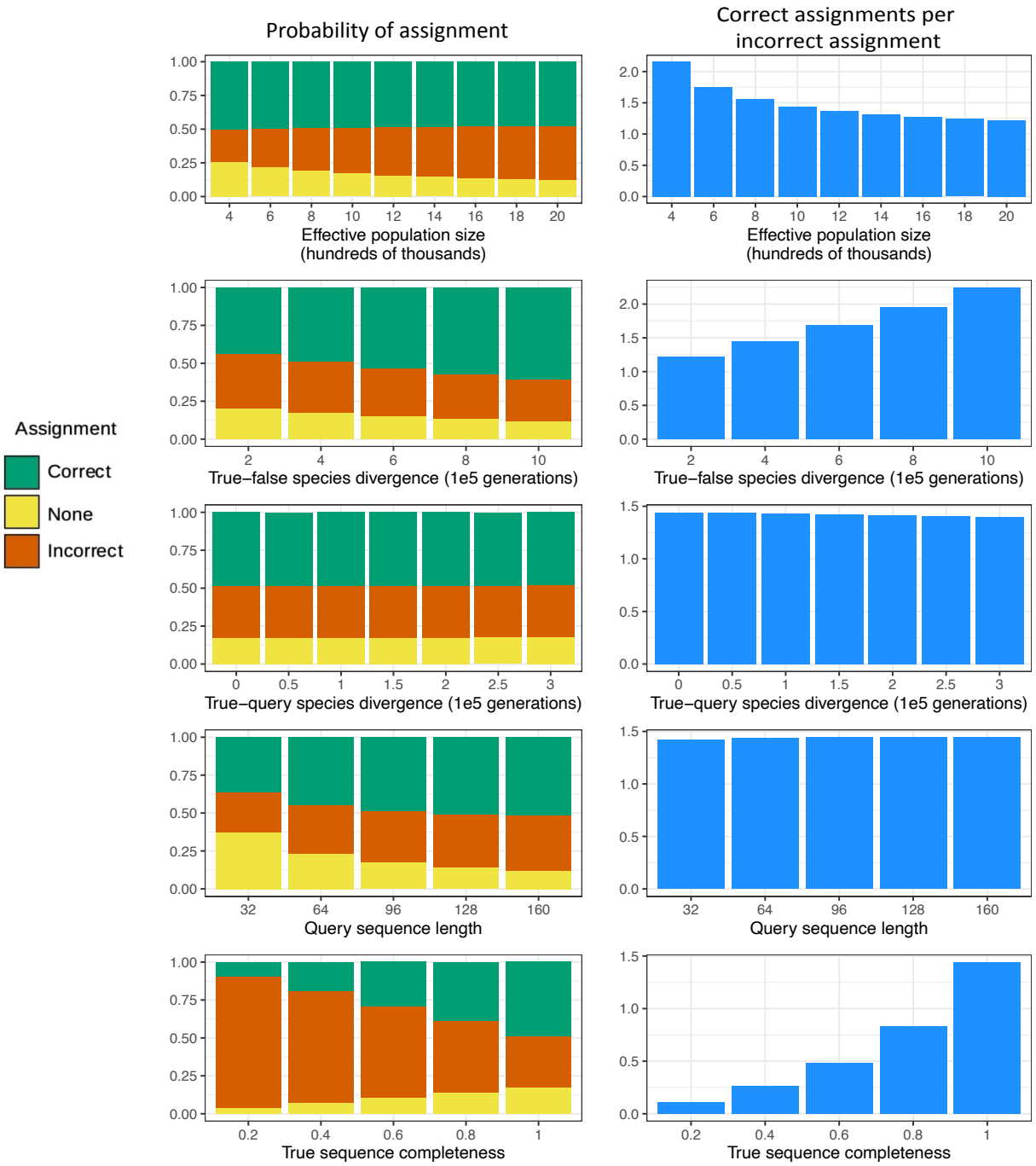


Figure 6: A replicate of Figure 5, but with effective population sizes two orders of magnitude larger. Left: Probability of assigning the query sequence correctly (to the true reference sequence) in green, of assigning the query sequence incorrectly (to the false reference sequence) in red, and of making no assignment in yellow, using the least mismatch method. Right: The expected number of correct assignments (to the true reference sequence) made for every one incorrect assignment (to the false reference sequence). In both, each row varies a different parameter while keeping the others constant.

$$P_{CA}(T_{t,f}, T_{q,t}, \mu, N, A_q, A_t, A_f \,;\, t_1 < T_{t,f} < t_2) =$$
$$\int_{T_{t,f}}^{\infty}\int_{T_{q,t}}^{T_{t,f}} \frac{e^{-t_1/2N}}{2N}\frac{e^{-t_2/2N}}{2N}\left(\sum_{k=0}^{\infty}\sum_{l=k+1}^{\infty}\frac{\mu^{l+k}(2t_2-t_1-A_f)^l(t_1-A_t)^k e^{-\mu(2t_2-A_f-A_t)}}{l!\;k!}\right)dt_1dt_2 \Big/$$
$$\left(e^{-T_{t,f}/2N}\left(e^{-(T_{q,t}/2N}-e^{-T_{t,f}/2N}\right)\right)$$
$$P_{CA}(T_{t,f}, T_{q,t}, \mu, N, A_q, A_t, A_f \,;\, T_{t,f} \le t_1 < t_2) =$$
$$\int_{T_{t,f}}^{\infty}\int_{t_1}^{\infty} \frac{e^{-t_1/2N}}{2N}\frac{e^{-t_2/2N}}{2N}\left(\sum_{k=0}^{\infty}\sum_{l=k+1}^{\infty}\frac{\mu^{l+k}(2t_2-t_1-A_f)^l(t_1-A_t)^k e^{-\mu(2t_2-A_f-A_t)}}{l!\;k!}\right)dt_2dt_1 \Big/$$
$$\left(\frac{1}{2}e^{-T_{t,f}/N}\right)$$
$$P_{CA}(T_{t,f}, T_{q,t}, \mu, N, A_q, A_t, A_f \,;\, T_{t,f} \le t_2 < t_1) =$$
$$\int_{T_{t,f}}^{\infty}\int_{t_2}^{\infty} \frac{e^{-t_1/2N}}{2N}\frac{e^{-t_2/2N}}{2N}\sum_{k=0}^{\infty}\sum_{l=k+1}^{\infty}\frac{\mu^{l+k}(t_2-A_f)^l(t_2-A_t)^k e^{-\mu(2t_2-A_f-A_t)}}{l!\;k!}dt_1dt_2 \Big/ \left(\frac{1}{2}e^{-T_{t,f}/N}\right)$$
$$P_{CA}(T_{t,f}, T_{q,t}, \mu, N, A_q, A_t, A_f \,;\, t_2 = t_1 < T_{t,f})$$
$$\int_{T_{t,f}}^{\infty}\int_{t_3}^{\infty} \frac{e^{-t_1/2N}}{2N}\frac{e^{-t_3/2N}}{2N}\sum_{k=0}^{\infty}\sum_{l=k+1}^{\infty}\frac{\mu^{l+k}(t_3-A_f)^l(2t_1-t_3-A_t)^k e^{-\mu(2t_1-A_t-A_f)}}{l!\;k!}dt_1dt_3 \Big/$$
$$\left(\frac{1}{2}e^{-T_{t,f}/N}\right)$$

The probability of each of these cases differs with ancient samples as well. In particular, the probability of the true and query individuals coalescing before $T_{t,f}$ is

$$\exp(-\min((T_{t,f}-A_q, T_{t,f}-A_t)/2N))$$

and the other probabilities are modified similarly.

2.2.2 Modelling Incomplete Reference Sequences

So far, we have assumed completeness of both reference genomes, where we can define completeness as the fraction of covered bases. Note that the fraction of covered bases ranges from 0 to 1 and is not equivalent to coverage. In particular, we are not considering base quality or depth here, only whether or not the bases are represented in the reference sequence. This concept of completeness is sometimes referred to as breadth of coverage.

In the case of an incomplete reference sequence, there will be regions of the reference genomes to which the query will not align due to its lack of representation in the reference, which will happen with probability directly proportional to its completeness. Let C_t and C_f denote the completeness of the true and false sequences respectively. We have three relevant cases:

- The query read aligns to both true and false sequences, with probability $C_t C_f / (1 - (1 - C_t)(1 - C_f))$
- The query read aligns to the true sequence but not the false sequence, with probability $C_t(1 - C_f)/(1 - (1 - C_t)(1 - C_f))$
- The query read aligns to the false sequence but not the true sequence, with probability $(1 - C_t)C_f/(1 - (1 - C_t)(1 - C_f))$

where the denominator is there in every case to represent the condition that the query does align to at least one of the two references. In the first case, the correct assignment probability is that presented above as $P_{CA}(T, \mu, N)$. In the second, the correct assignment probability is 1, and in the third it is 0. Therefore, accounting for reference completeness, the probability of assigning the query sequence to the true sequence is

$$\frac{C_t C_f P_{CA}(T, \mu, n) + C_t(1 - C_f)}{1 - (1 - C_t)(1 - C_f)} \tag{1}$$

2.2.3 Simulations

Next we wrote a simulation script using Python with msprime (Kelleher et al., 2016) as an engine. The script takes as input the genome length, effective population size, number of reads, read length, mutation rate, true-false divergence time, true-query divergence time, generation length, ages of the samples, and recombination rate. We build a phylogenetic tree using the input parameters, pass the relevant parameters and tree to msprime to create variable sites between the three sequences, randomly simulate nucleotides to fill in the remaining non-variable sites, and obtain diploid sequences for each of the true, query and false sequences. We randomly choose one of the two strands to represent the true and false sequences, and sample reads with equal probability from both strands of the query sequence. We can then either output these sequences and reads as fasta files, or directly compare the number of mismatches with the query reads to the true and false sequences to obtain assignment probabilities under each method. We note that msprime throws an error if one attempt to build a tree with a length zero branch, so we used $T_{q,t} = 0.0001$ in the simulations to

compare with the theoretical results for $T_{q,t} = 0$. To validate our theoretical predictions, we simulated all values in Figure 5, as shown in Supplementary Table 1. The simulation script is available at https://github.com/bdesanctis/binning-accuracy.

Next, we wanted to compare the least mismatch and exact match methods to a commonly used existing taxonomic binning method, and to do this based on a context previously established by an empirical study (Pedersen et al., 2021). For the former, we used a method based on discriminative k-mers, called Clark (Ounit et al., 2015), and ran it with default settings on the simulated fasta sequences and reads. The parameter set used was that from a recent ancient environmental DNA study which contained two Ursid species, American black bear *Ursus americanus* and the extinct giant short-faced bear *Arctodus simus*, which diverged approximately at $T_{t,f} = 13.4$mya (Pedersen et al., 2021). In the study, they used a spectacled bear *Tremarctos ornatus* reference sequence to assign giant short-faced bear reads. This means the true and the query sequence were from different species, diverged approximately $T_{q,t} = 5$mya (Pedersen et al., 2021). As in the original study, we also used an ursid generation length of 6 years, an effective population size of $10,000$, a mutation rate of $0.6e-8$ per site per generation, a query age of 14kya, and an age of 0 for the true and false sequences. We further used a recombination rate of $1e-8$ per site, and read lengths of 40 and 100bp. To obtain accuracy estimates, we simulated $10,000$ reads and a genome length of 10 million, and mapped the query reads back to the true and false sequences using bwa aln (Li and Durbin, 2009) with default parameters, then counted the mismatches between each read and the reference sequences using a custom script and samtools (Li et al., 2009a). In the least mismatch method, if a read did not align to one of the reference sequences at all, we assigned it to the other.

Lastly, we wanted to test the effect of deamination, a damage process that occurs in ancient DNA and appears as C to T or G to A transition SNPs in the reads. Effectively, this places "mutations" on the query branch (see Figure 4) and therefore, in theory, it should decrease both correct and incorrect assignments of the exact match method, and not significantly impact the least mismatch method. We used gargammel (Renaud et al., 2017) to add deamination to our simulated query reads, using a misincorporation matrix from the original dataset from the sample library "UE1210 Mex 18 Lib4" (Pedersen et al., 2021). We then re-ran the above analysis on the deaminated reads.

2.3 Results and discussion

2.3.1 Theoretical results

We first calculate the theoretical probability of a single query sequence being assigned to either the correct reference sequence, the incorrect reference sequence, or neither sequence under a coalescent

model. In particular, we model the situation where a single read or query sequence aligns to two reference sequences, and (a) the query is assigned to the reference sequence to which it has the least number of mismatches (referred to as "least mismatch"), or (b) the query is assigned to a reference sequence if they have no mismatches, and this is not the case for any other reference sequence (referred to as "exact match"). If the query sequence has the same number of mismatches as both reference sequences in either case, or does not exactly match any reference sequence in the exact match case, it is not assigned to either.

The relevant theory, which is detailed in the Methods section, relies on a few key parameters. First, we need to differentiate between the coalescence time of the true, false and query sequences, and the divergence times of their respective species or populations (Fig. 4). We use the terms "true-query species divergence time" and "true-false species divergence time" to refer to those of the latter class, as the exact coalescence times of the true, false and query sequences will be unknown in practice (see Methods and Fig. 4 for details). We also incorporate the length of the query sequence, the mutation rate, the effective population size (which is assumed to be constant), and the completeness of the reference sequences, the latter of which is defined as the proportion of sites represented in the reference genome. Further generalizations which incorporate the age of each of the sequences are derived in the Methods section. We use a fixed set of parameters and modify one at a time over a range to determine the relative impact of different parameters on assignment accuracy.

In Figure 5, we used a baseline parameter set consisting of the following parameters: the query sequence length $k = 96$, a mutation rate of $\mu = k \cdot 1e - 8$ per generation, an effective population size of $N_e = 10,000$, a query-true species divergence time of zero generations (i.e. assuming they come from the same population), a true-false species divergence time of $400,000$ generations, and a true sequence completeness and a false sequence completeness of 1. On each row of Figure 5, we use this baseline parameter set and modify one parameter at a time. On the left of Figure 5, we plot the probability of correctly assigning the query sequence to the true sequence in green, the probability of incorrectly assigning it to the false sequence in red, and the probability of not assigning it to either sequence in yellow. Since many of the incorrect assignment probabilities are too small to visually compare in this way, we also show, on the right, the number of expected correct assignments per one incorrect assignment. All results shown in Figure 5 are for the least mismatch assignment method. The exact-match assignment method gives similar results in general, and the differences between the two are small for the parameters shown in Figure 5, with exact values given in Supplementary Table 1. Our baseline parameter set leads to an expected 60.6 correct assignments for every incorrect assignment using the least-mismatch method.

We validated these theoretical results using simulations as described in Methods (see Supplementary Table 1), where we also give results for the exact-match assignment method. For each parameter combination, we simulated $10,000$ query reads from a sequence of length 10 million with a recombination rate of $1e-8$, so that each query read effectively acted as an independent replicate. To check that our simulated values matched our theoretical values, we performed two-sided binomial tests for each parameter combination, with p-values shown in Supplementary Table 1. Only five were significant at a level of $p = 0.05$, which is consistent with random chance since we checked a total of 128 different parameter combinations.

As seen in the first row of Figure 5, a smaller effective population size will lead to a more recent coalescence of the query and true sequence, so that there will be fewer expected differences between the two sequences and we have a higher chance of correctly assigning the query based on these differences. We note that selection can reduce the effective population size locally in the genome, with similar consequences (Liu and Mittler, 2008). We would also expect a higher correct assignment probability if the true and false species diverged a long time ago, as shown in the second row, in which case there will be more differences between the true and false sequences, and therefore more between the query and false sequences. Adaptive selection can also locally increase the number of differences. However, though both these factors impact the number of expected correct assignments per incorrect assignments, the effective population size barely impacts how many total sequences one expects to assign, whereas a higher true-false divergence time or local adaptive selection will lead to more total assigned sequences.

If we assume the query and the true sequence come from different populations or even different species (Fig 5 row 3), the probability of correct assignment can drop dramatically. Even when these two populations diverged $50,000$ generations ago, compared to the $400,000$ generation divergence of the true and false species, we would expect less than 20 correct assignments per incorrect assignment. A higher query-true divergence time amplifies the problem quickly, emphasizing the importance of using reference sequences which are as close as possible to the expected query species. In practice, we might expect a small but nonzero true-query divergence time even when using a reference sequence assumed to be very similar to the query, since in most cases one cannot assume that the query sequence and the reference sequence are from the same population. Furthermore, when using an ancient query sequence, we would expect the populations to have diverged some time ago, leading to a higher query-true divergence time and higher expected error rates. Another relevant case is when an ancient query sequence belongs to an extinct species for which there is no reference genome, and therefore the sequence of a nearby species must be used in the reference set, leading to a high true-query divergence time and consequently increasing error

rates. The age of a sample affects assignment accuracy in the exact same way as the query-true divergence time, and so is not shown here, though the relevant equations are given in the Methods.

The length of the query sequence has some impact on the assignment accuracy, where higher sequence lengths correspond to higher accuracies. Often a 30bp minimum sequence length is emphasized (Schubert et al., 2012), and while this is helpful, other parameters have a far greater effect on the assignment accuracy.

By far, the parameter that has the most impact on the assignment accuracy in this parameter range is the completeness of the true sequence, where completeness is defined as the fraction of represented sites in the reference genome. As can be seen in the last row of Figure 2, failing to have even a fifth of the sites in the reference genome covered will lead to a significant assignment error. This is a natural consequence of Equation 1. Perhaps the most important consideration to maximize binning accuracy is therefore ensuring that one uses high quality and complete reference genomes in the reference database. However, when this is not possible, a practical fix could be to remove from consideration those genomic regions, and associated query reads, which are not represented in all of the reference genomes where one would expect them to be. In particular, the query reads most susceptible to incorrect assignment due to insufficient reference genome completeness are those which map to a single region in one of the reference genomes, but fail to map to the other due to the lack of representation of that region in that reference genome. In practice, identifying these regions would probably require aligning the reference sequences.

Having multiple nearby reference sequences is not modelled here, but if they spanned the population to which the query belongs or has recently diverged from, this would also likely reduce assignment errors by increasing the chance of the query coalescing more quickly with one of these sequences. We are also not considering the case where we have high read coverage, where it may be more appropriate to undertake metagenomic assembly and assign contigs (Yang et al., 2021).

Microbial populations can have much larger effective population sizes than those shown in Figure 5. Because of this, in Figure 6, we also show theoretical results for effective population sizes two orders of magnitude larger than those in Figure 5, in the range of $N_e = 1,000,000$, with all other parameters as in Figure 5. The general patterns described above persist, but there is a much higher incorrect assignment probability, and the ratio of correct to incorrect assignments expected is often close to 1. This means that, with very high effective population sizes, individual read or sequence assignments of this type can be extremely unreliable, and one may have a high proportion of their reads incorrectly assigned.

Higher population sizes result in decreased accuracy because individuals in a population coalesce with each other at a rate inversely proportional to the population size, leading to more incom-

plete lineage sorting in larger populations. As the population size increases, then, the true, query and false sequences will be less likely to coalesce before their species divergence times T_{tf} and T_{qt}, after which there is equal probability of all three sequences coalescing. In the limit, with fixed divergence times and increasing population size, this leads to an equal probability of false or true assignment. In reality, however, such large populations (relative to species divergence times) are usually associated with bacteria or viruses, for which these types of assignment methods are not typically used.

Since there are scaling laws in coalescent theory, parameters such as effective population size, divergence time and mutation rate are only relevant in their relationships to other parameters. Any given parameter combination of effective population size N, mutation rate μ and divergence time T (which are the main determinants of error here) will be equivalent to that with an effective population size N/k, $k\mu$ and T/k. Indeed, as stated in the methods, this scaling is how the results for Figure 6 were calculated here, given the numerical instability of the integrals with high population sizes. This means that Figures 5 and 6 can be interpreted equally well in alternative regimes with proportionally lower population size and divergence times, and higher mutation rates, or the other way around.

2.3.2 Simulation results

We next wanted to (a) show how our two assignment methods compared to a commonly used binning software, (b) present results in a context previously established by an empirical study and (c) study how deamination, an ancient DNA damage process, might affect accuracy. We therefore compared the least-mismatch and the exact-match methods to Clark (Ounit et al., 2015), a discriminative k-mer method often used for taxonomic binning, on a parameter set motivated by a recent ancient environmental DNA study (Pedersen et al., 2021). This study contained reads from two separate bear species: the American black bear *Ursus americanus* and the extinct giant short-faced bear *Arctodus simus*. The closest living relative to the extinct *Arctodus* is the spectacled bear *Tremarctos ornatus*. Here, we simulated *Arctodus simus* query reads, and consider the accuracy when attempting to assign these reads to the closest "true" reference sequence *Tremarctos ornatus*, with the "false" reference sequence as *Ursus americanus*. The *Ursus-Arctodus* divergence time is approximately $T_{t,f} = 13.4$ million years ago, the *Arctodus-Tremarctos* divergence time is approximately $T_{t,f} = 5$ million years ago (Pedersen et al., 2021), and other relevant parameters are given in the Methods section. We simulated query reads from *Arctodus simus* using read lengths of $k = 40$ and $k = 100$, both with and without deamination, a type of ancient DNA damage.

Results are shown in Figure 7. First of all, as expected, exact match is a more conservative

method than least mismatch. This is especially true for a longer query sequence length, where the exact match method has a very low rate of incorrect assignments, but also assigns few total query sequences. This is because, when keeping other population genetic parameters constant, longer reads will be more likely to have mismatches with the reference sequence. Clark is an intermediate method between least mismatch and exact match in terms of both the total number of reads assigned and the proportion of reads which were assigned correctly. This is somewhat expected, since least mismatch effectively assigns reads wherever possible, and exact match is highly conservative in its assignments. However, when the read length is increased to 100, Clark has a higher rate of incorrect assignments than either exact match or least mismatch. We believe this is a consequence of Clark making decisions based on shorter k-mers than the whole sequence (we used the default k-mer size of 31, but the maximum k-mer length in Clark is 32). We expect this to be a general phenomenon of k-mer based methods, including others such as Kraken (Wood et al., 2019), in scenarios with longer query reads that are expected to have fairly close reference sequences. We therefore do not recommend using k-mer based methods in these situations. However, for very long query sequences such as scaffolds or chromosomes, those assembled with query reads, or when a nearby reference sequence does not exist, the full query may fail to align to any reference sequence and k-mer methods may be superior.

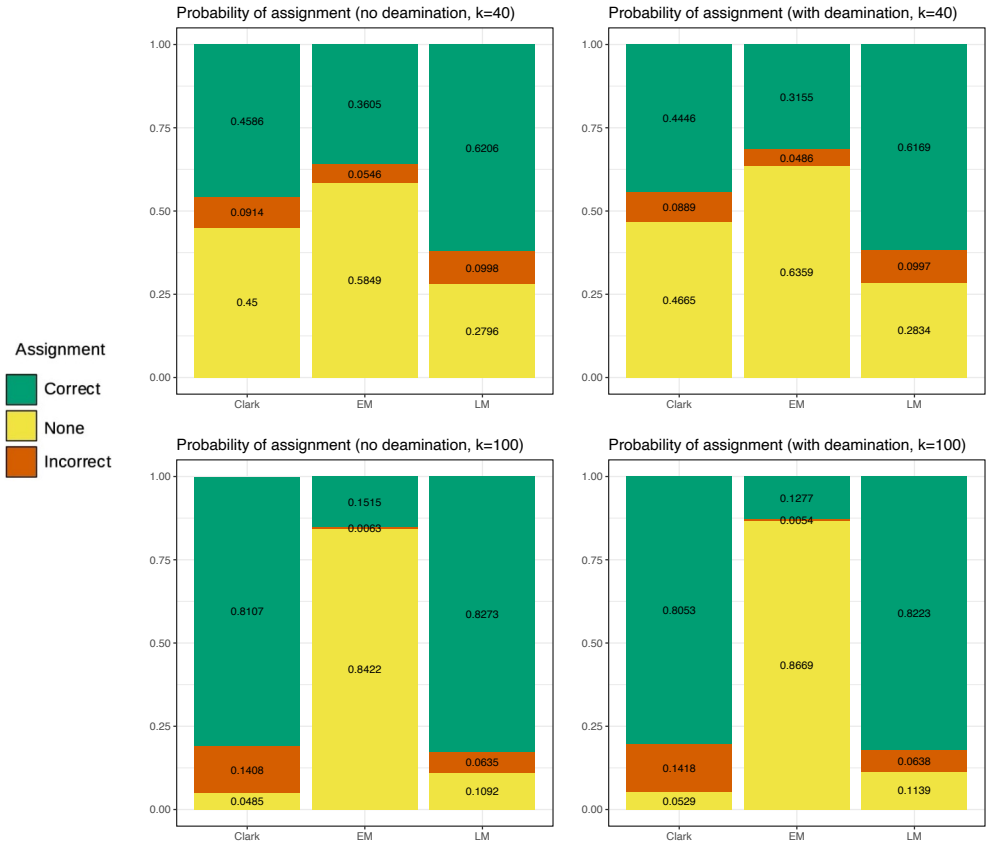


Figure 7: Simulated binning accuracy results using a parameter set motivated by a recent empirical study on ancient environmental DNA of black bears and giant short-faced bears (Pedersen et al., 2021). Results including deamination are shown on the right. Here Clark refers to a discriminative k-mer based method (Ounit et al., 2015), EM is "exact-match" and LM is "least-mismatch". Parameters used are given in the Methods.

We also simulated deamination, a damage process affecting ancient DNA which introduces extra SNPs to the query reads. In theory, this will increase the number of mismatches between the query read and both the true and false sequences by the same number. Therefore, we expect that deamination should not impact the least mismatch method, but should reduce total assignments in both exact match and Clark. This is indeed what we see in Figure 7.

2.4 Conclusion

We hope that the theoretical and simulation framework presented here will add to an improved understanding of the uncertainties in the use of reference databases and binning methods when carrying out taxonomic assignment with both ancient and present day metagenomic DNA. While many of the factors we discuss have already been acknowledged to be an issue in the community, there has previously been a lack of quantitative investigation of relationship of these parameters to the accuracy of taxonomic binning methods. We have also introduced a simulation framework to compare the least-mismatch and exact-match methods to a commonly used k-mer method Clark (Ounit et al., 2015) using parameters from a recent study to contextualize our results (Pedersen et al., 2021). We have made both the theoretical and simulation code publicly available, so that researchers may analyze and compare the performance of their own binning algorithms beyond fixed datasets or on specific parameters relating to their own studies. A framework in which to understand these issues is especially important as the fields of environmental and metagenomic DNA move into individual and population genetic analysis.

Table 1: Comparison of theoretical and simulated results for values in Figures 5 and 6.

Assignment	Method	N	Ttf	k	Tqt	Theory	Simulated	p-value
correct	least-mismatch	4000	4e+05	96	0	0.5368	5407	0.4401
incorrect	least-mismatch	4000	4e+05	96	0	0.0036	44	0.1531
no	least-mismatch	4000	4e+05	96	0	0.4596	4549	0.3456
correct	exact-match	4000	4e+05	96	0	0.5314	5353	0.4404
incorrect	exact-match	4000	4e+05	96	0	0.0035	42	0.2334
no	exact-match	4000	4e+05	96	0	0.4651	4605	0.3617
correct	least-mismatch	6000	4e+05	96	0	0.5372	5347	0.6232
incorrect	least-mismatch	6000	4e+05	96	0	0.0053	49	0.6303
no	least-mismatch	6000	4e+05	96	0	0.4575	4604	0.5673
correct	exact-match	6000	4e+05	96	0	0.5292	5261	0.5412
incorrect	exact-match	6000	4e+05	96	0	0.0052	48	0.6754
no	exact-match	6000	4e+05	96	0	0.4657	4691	0.4955
correct	least-mismatch	8000	4e+05	96	0	0.5375	5406	0.5341
incorrect	least-mismatch	8000	4e+05	96	0	0.0071	70	0.9525
no	least-mismatch	8000	4e+05	96	0	0.4554	4524	0.5469
correct	exact-match	8000	4e+05	96	0	0.5269	5311	0.4002
incorrect	exact-match	8000	4e+05	96	0	0.0068	70	0.8076
no	exact-match	8000	4e+05	96	0	0.4663	4619	0.3832
correct	least-mismatch	10000	4e+05	96	0	0.5377	5363	0.7789
incorrect	least-mismatch	10000	4e+05	96	0	0.0089	69	0.0373
no	least-mismatch	10000	4e+05	96	0	0.4534	4568	0.501
correct	exact-match	10000	4e+05	96	0	0.5247	5246	0.992
incorrect	exact-match	10000	4e+05	96	0	0.0084	67	0.0625
no	exact-match	10000	4e+05	96	0	0.4669	4687	0.7258
correct	least-mismatch	12000	4e+05	96	0	0.538	5369	0.8332
incorrect	least-mismatch	12000	4e+05	96	0	0.0106	105	0.9611
no	least-mismatch	12000	4e+05	96	0	0.4514	4526	0.8095
correct	exact-match	12000	4e+05	96	0	0.5224	5239	0.7716
incorrect	exact-match	12000	4e+05	96	0	0.01	101	0.8801
no	exact-match	12000	4e+05	96	0	0.4676	4660	0.7561
correct	least-mismatch	14000	4e+05	96	0	0.5382	5433	0.3063
incorrect	least-mismatch	14000	4e+05	96	0	0.0124	124	1

no	least-mismatch	14000	4e+05	96	0	0.4494	4443	0.3053
correct	exact-match	14000	4e+05	96	0	0.5202	5259	0.2581
incorrect	exact-match	14000	4e+05	96	0	0.0115	119	0.7077
no	exact-match	14000	4e+05	96	0	0.4682	4622	0.2292
correct	least-mismatch	16000	4e+05	96	0	0.5384	5357	0.595
incorrect	least-mismatch	16000	4e+05	96	0	0.0142	132	0.4461
no	least-mismatch	16000	4e+05	96	0	0.4475	4511	0.4691
correct	exact-match	16000	4e+05	96	0	0.5181	5166	0.7717
incorrect	exact-match	16000	4e+05	96	0	0.013	118	0.3101
no	exact-match	16000	4e+05	96	0	0.4689	4716	0.5954
correct	least-mismatch	18000	4e+05	96	0	0.5385	5322	0.2063
incorrect	least-mismatch	18000	4e+05	96	0	0.0159	148	0.4014
no	least-mismatch	18000	4e+05	96	0	0.4455	4530	0.1339
correct	exact-match	18000	4e+05	96	0	0.5159	5110	0.3269
incorrect	exact-match	18000	4e+05	96	0	0.0145	134	0.3796
no	exact-match	18000	4e+05	96	0	0.4696	4756	0.2293
correct	least-mismatch	20000	4e+05	96	0	0.5387	5371	0.7559
incorrect	least-mismatch	20000	4e+05	96	0	0.0177	183	0.6218
no	least-mismatch	20000	4e+05	96	0	0.4436	4446	0.8484
correct	exact-match	20000	4e+05	96	0	0.5138	5153	0.7641
incorrect	exact-match	20000	4e+05	96	0	0.0159	161	0.873
no	exact-match	20000	4e+05	96	0	0.4703	4686	0.741
correct	least-mismatch	10000	2e+05	96	0	0.3262	3264	0.9745
incorrect	least-mismatch	10000	2e+05	96	0	0.0129	114	0.1842
no	least-mismatch	10000	2e+05	96	0	0.6608	6622	0.7675
correct	exact-match	10000	2e+05	96	0	0.3194	3210	0.7396
incorrect	exact-match	10000	2e+05	96	0	0.0124	111	0.2771
no	exact-match	10000	2e+05	96	0	0.6682	6679	0.9492
correct	least-mismatch	10000	6e+05	96	0	0.6828	6917	0.0558
incorrect	least-mismatch	10000	6e+05	96	0	0.0061	53	0.3351
no	least-mismatch	10000	6e+05	96	0	0.3111	3030	0.0802
correct	exact-match	10000	6e+05	96	0	0.6644	6748	0.0284
incorrect	exact-match	10000	6e+05	96	0	0.0057	50	0.3885
no	exact-match	10000	6e+05	96	0	0.3298	3202	0.0412

correct	least-mismatch	10000	8e+05	96	0	0.7823	7904	0.0511
incorrect	least-mismatch	10000	8e+05	96	0	0.0042	32	0.1403
no	least-mismatch	10000	8e+05	96	0	0.2135	2064	0.0853
correct	exact-match	10000	8e+05	96	0	0.7596	7698	0.0175
incorrect	exact-match	10000	8e+05	96	0	0.0039	31	0.2282
no	exact-match	10000	8e+05	96	0	0.2365	2271	0.0278
correct	least-mismatch	10000	1e+06	96	0	0.8507	8547	0.2618
incorrect	least-mismatch	10000	1e+06	96	0	0.0029	25	0.5741
no	least-mismatch	10000	1e+06	96	0	0.1465	1428	0.302
correct	exact-match	10000	1e+06	96	0	0.8245	8284	0.3115
incorrect	exact-match	10000	1e+06	96	0	0.0027	25	0.846
no	exact-match	10000	1e+06	96	0	0.1728	1691	0.3278
correct	least-mismatch	10000	4e+05	32	0	0.2293	2276	0.6947
incorrect	least-mismatch	10000	4e+05	32	0	0.0049	41	0.2535
no	least-mismatch	10000	4e+05	32	0	0.7656	7683	0.5316
correct	exact-match	10000	4e+05	32	0	0.2279	2260	0.6678
incorrect	exact-match	10000	4e+05	32	0	0.0049	41	0.3135
no	exact-match	10000	4e+05	32	0	0.7673	7699	0.5384
correct	least-mismatch	10000	4e+05	64	0	0.4042	3974	0.169
incorrect	least-mismatch	10000	4e+05	64	0	0.0076	80	0.6455
no	least-mismatch	10000	4e+05	64	0	0.5882	5946	0.1935
correct	exact-match	10000	4e+05	64	0	0.3981	3919	0.209
incorrect	exact-match	10000	4e+05	64	0	0.0074	77	0.6827
no	exact-match	10000	4e+05	64	0	0.5945	6004	0.2335
correct	least-mismatch	10000	4e+05	128	0	0.6401	6357	0.3648
incorrect	least-mismatch	10000	4e+05	128	0	0.0092	85	0.496
no	least-mismatch	10000	4e+05	128	0	0.3507	3558	0.2899
correct	exact-match	10000	4e+05	128	0	0.6182	6143	0.4281
incorrect	exact-match	10000	4e+05	128	0	0.0085	76	0.3548
no	exact-match	10000	4e+05	128	0	0.3733	3781	0.3261
correct	least-mismatch	10000	4e+05	160	0	0.7188	7200	0.7981
incorrect	least-mismatch	10000	4e+05	160	0	0.009	91	0.8737
no	least-mismatch	10000	4e+05	160	0	0.2722	2709	0.7788
correct	exact-match	10000	4e+05	160	0	0.6866	6879	0.7959

incorrect	exact-match	10000	4e+05	160	0	0.0081	83	0.8233
no	exact-match	10000	4e+05	160	0	0.3053	3038	0.7611
incorrect	exact-match	10000	4e+05	96	1e+05	0.0485	475	0.6417
no	exact-match	10000	4e+05	96	1e+05	0.5549	5614	0.1943
correct	least-mismatch	10000	4e+05	96	150000	0.4219	4156	0.2021
incorrect	least-mismatch	10000	4e+05	96	150000	0.0834	838	0.885
no	least-mismatch	10000	4e+05	96	150000	0.4947	5006	0.238
correct	exact-match	10000	4e+05	96	150000	0.3425	3395	0.5411
incorrect	exact-match	10000	4e+05	96	150000	0.0672	677	0.8417
no	exact-match	10000	4e+05	96	150000	0.5903	5928	0.6184
correct	least-mismatch	10000	4e+05	96	2e+05	0.3839	3813	0.6001
incorrect	least-mismatch	10000	4e+05	96	2e+05	0.1117	1113	0.9241
no	least-mismatch	10000	4e+05	96	2e+05	0.5044	5074	0.5552
correct	exact-match	10000	4e+05	96	2e+05	0.2942	2895	0.3127
incorrect	exact-match	10000	4e+05	96	2e+05	0.085	848	0.9571
no	exact-match	10000	4e+05	96	2e+05	0.6208	6257	0.3225
correct	least-mismatch	10000	4e+05	96	250000	0.3465	3444	0.6666
incorrect	least-mismatch	10000	4e+05	96	250000	0.1414	1391	0.5185
no	least-mismatch	10000	4e+05	96	250000	0.5121	5165	0.3787
correct	exact-match	10000	4e+05	96	250000	0.2511	2503	0.8536
incorrect	exact-match	10000	4e+05	96	250000	0.1019	1017	0.9605
no	exact-match	10000	4e+05	96	250000	0.6469	6480	0.8343
correct	least-mismatch	10000	4e+05	96	3e+05	0.3102	3114	0.787
incorrect	least-mismatch	10000	4e+05	96	3e+05	0.1723	1758	0.354
no	least-mismatch	10000	4e+05	96	3e+05	0.5176	5128	0.3418
correct	exact-match	10000	4e+05	96	3e+05	0.2131	2151	0.6253
incorrect	exact-match	10000	4e+05	96	3e+05	0.1179	1203	0.4568
no	exact-match	10000	4e+05	96	3e+05	0.669	6646	0.3553

3 Phylogenetic placement of Arctic mammoth and horse from ancient environmental DNA

This chapter applies a novel algorithm called pathPhynder to phylogenetically place aeDNA reads from an Arctic-wide dataset spanning the last 50,000 years.

The algorithm itself is published in:
Martiniano, R., De Sanctis, B., Hallast, P., and Durbin, R. (2022). Placing ancient DNA sequences into reference phylogenies. *Molecular Biology and Evolution*, 39(2).

The phylogenetic placement of the mammoth and horse samples is published as part of:
Wang, Y., Pedersen, M. W., Alsos, I. G., De Sanctis, B., Racimo, F., Prohaska, A., Coissac, E., Owens, H. L., Merkel, M. K. F., Fernandez-Guerra, A., Rouillard, A., Lammers, Y., Alberti, A., Denoeud, F., Money, D., Ruter, A. H., McColl, H., Larsen, N. K., Cherezova, A. A., Edwards, M. E., Fedorov, G. B., Haile, J., Orlando, L., Vinner, L., Korneliussen, T. S., Beilman, D. W., Bjørk, A. A., Cao, J., Dockter, C., Esdale, J., Gusarova, G., Kjeldsen, K. K., Mangerud, J., Rasic, J. T., Skadhauge, B., Svendsen, J. I., Tikhonov, A., Wincker, P., Xing, Y., Zhang, Y., Froese, D. G., Rahbek, C., Bravo, D. N., Holden, P. B., Edwards, N. R., Durbin, R., Meltzer, D. J., Kjær, K. H., Möller, P., and Willerslev, E. (2021). Late Quaternary dynamics of Arctic biota from ancient environmental genomics. *Nature*, 600(7887):86–92.

In the last decade, there have been a number of genetic studies on the extinct woolly mammoths *Mammuthus primigenius* and horses *Equus caballus, lambei* and *scotti* from fossils, leading to over 100 published full or partial mitochondrial genomes in each case. Here I studied ancient environmental DNA from 1671 permafrost and lake sediment samples from across the Arctic circle from the last 50,000 years. Nearly half of these samples contain woolly mammoth DNA, and approximately a fifth of these contain reads that map to the woolly mammoth mitochondrial genome. Additionally, more than a hundred of these samples contain horse DNA mapping to a reference panel of hundreds of horse genomes. Here I use pathPhynder, a novel phylogenetic placement software for ancient DNA, to place these aeDNA mitochondrial reads on the existing mammoth and horse mitochondrial trees. In the case of the mammoths, the majority of cases achieve at least a clade-level placement, with some achieving a surprising level of depth. In the case of the horses, only a small number of reads achieve reliable placement, and serve to thicken out a previously known ancient Yukon clade. In the main our assignments are consistent with previous findings from skeletal material, although there are some suggestions of previously unseen subclades in the case of the mammoths. Not only does this study illuminate the population dynamics and evolution of woolly mammoths and horses, but it also serves as a proof of concept that even fragmented, damaged, and low-quality ancient environmental DNA reads can often be placed into a species-specific phylogenetic tree.

3.1 Introduction

This study provides a phylogenetic placement analysis using aeDNA isolated and shotgun sequenced from 1671 lake and permafrost samples from 73 different sites across the last 50,000 years in the Arctic. These samples were filtered and mapped to the NCBI database using the ngsLCA lowest common ancestor algorithm (Wang et al., 2022), which assigns reads to individual taxa where possible. Further details of the sampling procedure, sequencing, and mapping can be found in (Wang et al., 2021). In this chapter, I apply a novel phylogenetic placement algorithm to place mitochondrial reads from this dataset which were assigned to *Mammuthus* or mammoth and *Equus* or horse into phylogenies constructed from previously published high-quality mitochondrial reference genomes. First in this section, I will briefly overview what is known about the evolutionary history of mammoths and horses, then describe the phylogenetic placement algorithm pathPhynder.

3.1.1 Evolutionary history of mammoths

Elephantids first evolved around 10 million years ago, with mammoths diverging from other elephantid species approximately 6 million years ago (Palkopoulou et al., 2018). Like elephants in general, mammoths first evolved in Africa, moved north into Asia and Europe, and then made their way into North America using the Bering Land Bridge. The main mammoth species are *Mammuthus meridionalis* or Southern mammoths, *M. trogontherii* or Steppe mammoths, *M. columbi* or Columbian mammoths, and finally *M. primigenius* or woolly mammoths (additionally, sometimes the North American woolly mammoths are separated from woolly mammoths in the literature and called *M. americanus*). Up until recently, it was thought that these species more or less gave rise to each other, in a linear fashion in the order presented above, with the transition to woolly mammoth occurring approximately 700,000 years ago. The evidence for this was largely fossil and morphology based. However, this view has been challenged by progress in the field of mammoth genetics, such as in Chang et al. (2017). The situation now appears more complicated, with possible long-term coexistence of different mammoth species and multiple hybridization events (Roca et al., 2015). It has also been suggested that earlier mammoth species such as the Steppe mammoth migrated to North America and evolved there, resulting in a separate species - though this now seems unlikely, given recent phylogenetic studies (Lister and Sher, 2015).

Within woolly mammoths, there appear to have been five mitochondrial haplogroups (labelled A to E) forming three major clades (labelled 1 to 3). In a potentially confusing labelling, clade 3 is equivalent to haplogroup B, and clade 2 is equivalent to haplogroup A. However, clade 1 contains haplogroups C, D and E, although the latter two are often grouped together and labelled

as 1DE. Both clades 2 and 3 are known to be older, dying out around 30,000 years ago, while clade 1 contains the last persisting mammoth populations. Clade 1C is often called the American clade (and, as previously mentioned, sometimes called *Mammuthus Americanus*), because of their presence in the Americas. These clades diverged around 1 million years ago, with the MRCA of clade 1 estimated to be around 500 ka (Chang et al., 2017).

Woolly mammoth remains, or evidence of woolly mammoths in eDNA, have not been found more recently than approximately 11,000 years ago except on Wrangel Island, an island in the Arctic Ocean off Eastern Siberia, where mammoth remains have been found dating up to as recently as 4,000 years ago. Wrangel Island was isolated from mainland Siberia 10,000 years ago once the sea level rose during the transition from the Pleistocene to the Holocene. Because of this, the mammoths on the island would have been genetically isolated around this time. Indeed, evidence for an apparent founder effect and population bottleneck is shown in Pečnerová et al. (2017). There has been no evidence for mammoths outside of Wrangel Island after 11,000 years ago, leading to descriptions of the island in the literature such as "the refugium of the last surviving population of the species" (Pečnerová et al., 2017). However, in stark contradiction to these claims, the current study finds strong evidence for mainland mammoths existing more recently, dating as recently as 7,000 years ago.

The first mammoth mitochondrion was sequenced in 2006, and at the time was the oldest mitochondrial sequence of any species to date (Rogaev et al., 2006). Since then, over 100 partial or complete mitochondrial genomes for woolly mammoths have been published, all from mammoth remains. However, only a few whole genomes have been sequenced. Because of this, the mammoth mitochondrial tree is significantly more complete and well understood than the nuclear tree.

3.1.2 Evolutionary history of horses

The *Equus* genus, which includes horses, zebras and donkeys, diverged around 6-8 million years, with horses in particular diverging around 5 million years ago (Orlando et al., 2013a). In the horse clade, the only living species are the modern domestic horse, *Equus ferus caballus*, and Przewalski's horse, *Equus przewalski*, although multiple other species have gone recently extinct. For example, both *Equus lambei* and *Equus scotti* specimens have been found from 20-30kya. Not much is known about either of these species, though we do have a few mitochondrial genomes of each. *Equus ferus ferus*, also called the tarpan, just recently went extinct in 1909. In general, the taxonomy of these horse subspecies is still unclear, and categorizations and naming conventions can differ throughout the literature. Often the term "caballoid" is used to refer to the horse clade in general.

Along with *caballus*, the only other living *Equus* species is Przewalski's horse, which lives in central Asia and diverged from other horses approximately 45,000 years ago (MacHugh et al., 2017). It should be noted that this horse is sometimes referred to as its own species, *Equus przewalski*, as a subspecies of *ferus*, as *Equus ferus przewalski*, or as a subgroup of *caballus*. Up until recently, it was thought that the extant Przewalski's horse was the only surviving wild horse (Gaunitz et al., 2018). However, a recent genetic study suggested that *przewalski* descended from horses that were domesticated some 5,500 years ago in the Botai region of Kazahkstan (Gaunitz et al., 2018). This would imply that there are no living horses which have not undergone some degree of domestication.

Domestication of the modern horse, *Equus ferus caballus*, likely began around 4000-5000 years ago in Eurasia, and became widespread by 3000 years ago (Anthony, 2010). This led to an increase in gene flow and a significant decline in genetic diversity. Though the last undomesticated horse only lived a century ago (see *Equus ferus ferus* or tarpan), horses in North America disappeared around 10,000 years ago, likely due to the warming temperatures (Librado et al., 2017). Domesticated horses were then reintroduced to North America around 500 years ago by European settlers.

The horse mitochondrial phylogeny has at least 18 different haplogroups, and is not generally structured with respect to geography or age as in the case of the mammoth phylogeny (Achilli et al., 2012). For this reason, we will not find it meaningful to refer to these haplogroups in this study. In fact, in the mitochondrial phylogeny, *Equus caballus, lambei, scotti* and *przewalski* are not represented by distinct clades. It has been suggested that although both *lambei* and *scotti* appear to be morphologically distinct from *caballus*, they may not be not genetically well defined species. In fact, it remains unclear whether any or all of these four are different species. For this reason, we will use the term "horse" to refer to any of these four species.

Until recently, the oldest full genome of any species was of a horse, dating to between 780,000 and 560,000 years ago from a fossilized foot bone at a site called Thistle Creek in Yukon, Canada (Orlando et al., 2013a). This sample, along with a few others from the same region but with a wide range of ages, represents the most diverged outgroup in the horse mitochondrial phylogeny.

3.1.3 Phylogenetic placement with pathPhynder

Even when a single phylogeny can be assumed to exist, for example on a non-recombining haplotype such as the mitochondrion or chloroplast genome, reconstructing an entire phylogeny and integrating ancient environmental DNA is often unrealistic or inaccurate because of high missingness resulting in too few overlapping SNPs (e.g. see Lemmon et al. (2009)). Instead, we can use

phylogenetic placement, or the assignment of a query sample to a branch on a fixed reference phylogeny, such as implemented in pathPhynder (Martiniano et al., 2022). The essential idea behind phylogenetic placement is to first assign variant sites to the branches on the reference on which the mutation creating them is inferred to have occurred, overlap the ancient variants in a query sample with the modern ones, and use this information to trace a path for the query sample through the phylogenetic tree. A toy example is shown in Figure 8, where Figure 8C shows the best path through the tree as determined by the overlapping SNPs in the query sample.

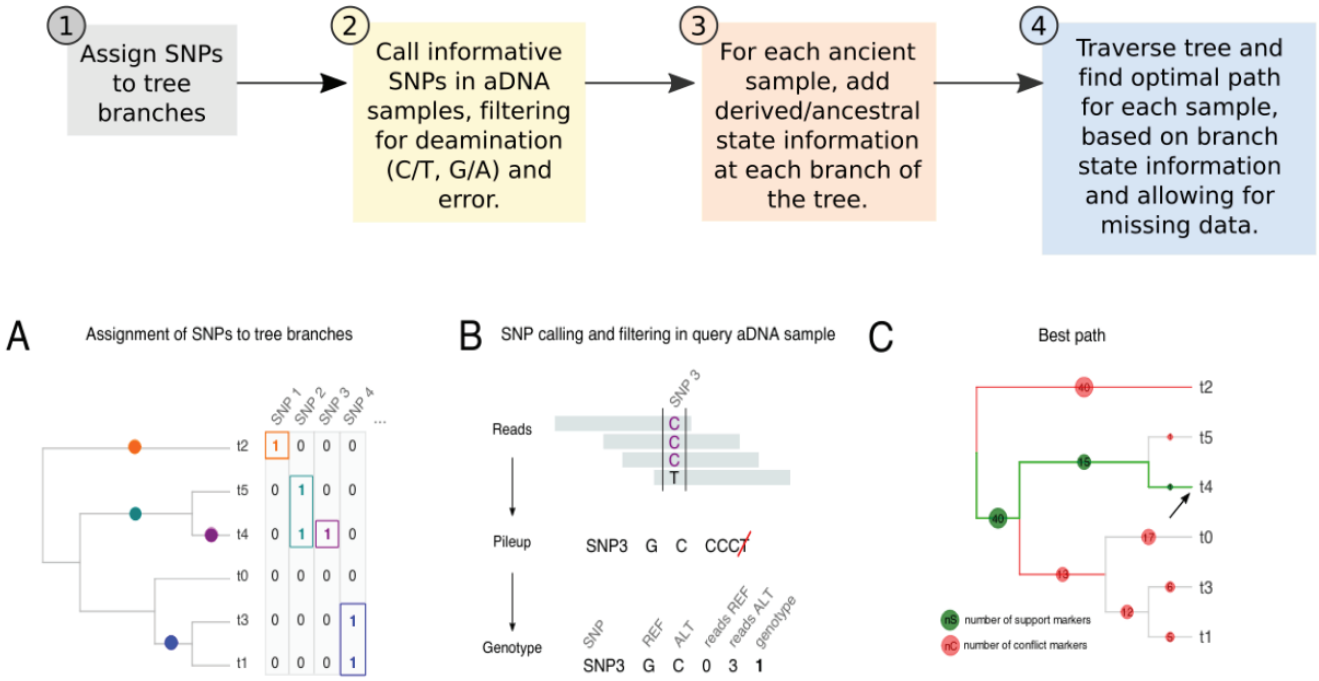


Figure 8: Top: Steps in the pathPhynder algorithm. Bottom: An illustration of a toy example. A: Biallelic SNPs are assigned to branches in a fixed reference phylogeny based on which samples they are present in. B: Processing the query sample on the variant site called SNP 3. In this example, a deaminated C-to-T is filtered out. The query sample is called as ALT on SNP3, adding a supporting marker to the branch with the purple marker above t4 in A. If the query sample was called as REF on SNP3, this would have added a conflicting marker on that same branch. C: Traversing the tree using all of the supporting and conflicting SNPs from the query sample gives a "best path" in green. Figure adapted from (Martiniano et al., 2022).

This approach has several limitations. First, it fails to account for any private variation in the query sample. In theory, private mutations could be used to estimate the private branch length belonging to the query, but this is not necessarily very accurate with ancient environmental DNA due to high missingness and bias in which regions of the genome are represented (see Chapter

1). Second, it can only account for biallelic SNPs which are present in the reference panel with no missingness and which conform to the tree structure. Importantly, phylogenetic placement by tracing the best path does not account for double mutations or yield a posterior probability, as a likelihood framework would. This likelihood calculation is also implemented in pathPhynder, usually gives the same or a similar result as the best path option, and is much faster. Likelihood-based phylogenetic placement is also implemented in Matsen et al. (2010), for example. However, for the data in this chapter and many ancient DNA applications, it is beneficial to know exactly which SNPs led to the phylogenetic placement, which is not so easily visualized under a standard likelihood framework. This is so that, especially when placements are based on one or a few SNPs, the loci can be examined for possible deamination or other errors, and so that one can explicitly report how many and which SNPs led us to the placement, giving a straightforward measure of reliability. Explicitly visualizing the number of SNPs can also give a rough idea of where on its placement branch the query sample diverged, as determined by the ratio of supporting to conflicting to supporting SNPs. Furthermore, environmental samples may include reads from a mixture of populations which are not nearby each other in the reference tree, obscuring phylogenetic signal. In this case, a likelihood approach would simply result in an unclear placement with low confidence, but visualization of the supporting and conflicting SNPs on each branch can help explicitly identify these separate paths and indicate a mixed sample, which might then be separated (for example, see Figure 34 in Chapter 5 or Figure 27a in Chapter 4). For these reasons, we rely on the "best path" option in pathPhynder in this chapter.

3.2 Materials and Methods

3.2.1 Previously published reference genomes

Mammoths. I used 78 previously published whole or partial mitochondrial genome .fasta files from NCBI, representing most of the published mitochondrial *Mammuthus primigenius* genomes. 65 of these genomes were originally proposed by my collaborator at the start of the project for their high quality and diversity, to which I added an extra 13 genomes to fill out clades 2 and 3 as much as possible. The list of GenBank IDs for these 78 samples, along with their latitude, longitude, age and reference (if available) is given in Table 2. This data was compiled by searching through the original publications in each instance and manually entering each entry. The publications containing the original data are (Chang et al., 2017; Pečnerová et al., 2017; Palkopoulou et al., 2013; Gillette and Madsen, 1993; Debruyne et al., 2008; Enk et al., 2016; Krause et al., 2005; Poinar et al., 2006; Kornienko et al., 2018; Palkopoulou et al., 2015). The permafrost samples were originally mapped against two mammoth whole reference genomes (NCBI Ids EU153446.1 and DQ316067.1), and all 78 genomes were used to construct a reference phylogeny in which to place the permafrost samples.

Table 2: References for 78 previously published mitochondrial mammoth genomes.

GenBank ID	Latitude	Longitude	Locality Reference	Age	Age Reference
EU153446	69.8	169	Palkapoulou et al. 2013	13995	Gilbert et al. 2008
DQ316067	68.17	165.93	Palkapoulou et al. 2013	32850	Rogaev et al. 2006
EU153445	72.5	127.5	Palkapoulou et al. 2013	35800	Gilbert et al. 2008
EU153453	69.79	157.7	Palkapoulou et al. 2013	>55200	Gilbert et al. 2008
EU153450	73.64	142.89	Palkapoulou et al. 2013	>58000	Gilbert et al. 2008
EU153451	73.21	143.6	Palkapoulou et al. 2013	>63500	Gilbert et al. 2008
EU153452	73.64	142.67	Palkapoulou et al. 2013	50200	Gilbert et al. 2008
EU153458	62.67	142.93	Palkapoulou et al. 2013	46900	Gilbert et al. 2008
EU153454	68.6	147.06	Palkapoulou et al. 2013	24740	Gilbert et al. 2008
EU153448	71.87	140.58	Palkapoulou et al. 2013	18560	Gilbert et al. 2008
EU153444	-	-	-	-	-
EU153456	67.83	124.29	Palkapoulou et al. 2013	18545	Gilbert et al. 2008
EU153457	-	-	-	-	-
AP008987	-	-	-	-	-
JF912200	69.37	-154.67	Debruyne et al. 2008	41510	Enk et al. 2011

GenBank ID	Latitude	Longitude	Locality Reference	Age	Age Reference
EU153455	74.15	99.59	Palkapoulou et al. 2013	20620	Gilbert et al. 2008
EU153449	73.32	105.4	Palkapoulou et al. 2013	20380	Gilbert et al. 2008
EU153447	72.09	79.35	Palkapoulou et al. 2013	17125	Gilbert et al. 2008
KX027489	74.42	107.75	Enk et al. 2016	>48800	Enk et al. 2016
KX027490	73.75	102	Enk et al. 2016	27740	Enk et al. 2016
KX027491	65.17	-147.5	Enk et al. 2016	42764	Enk et al. 2016
KX027492	64.83	-148	Enk et al. 2016	16789	Enk et al. 2016
KX027495	70.4	143.95	Enk et al. 2016	12125	Enk et al. 2016
KX027508	68.06	-139.78	Enk et al. 2016	-	-
KX027534	40.52	-89.72	Enk et al. 2016	17510	Enk et al. 2016
KX027507	39.82	-89.53	Enk et al. 2016	20550	Enk et al. 2016
KX027526	68.9	69.5	Enk et al. 2016	41910	Enk et al. 2016
KX027531	64.05	-139.42	Enk et al. 2016	37920	Enk et al. 2016
KX027532	64.05	-139.42	Enk et al. 2016	38600	Enk et al. 2016
KX027533	63.5	142.75	Enk et al. 2016	41300	Enk et al. 2016
KX027536	42.15	-78.93	Enk et al. 2016	10350	Enk et al. 2016
KX027498	38.88	-84.75	Enk et al. 2016	13985	Enk et al. 2016
KX027499	38.88	-84.75	Enk et al. 2016	12930	Enk et al. 2016
KX027500	38.88	-84.75	Enk et al. 2016	13215	Enk et al. 2016
KX027501	38.88	-84.75	Enk et al. 2016	13950	Enk et al. 2016
KX027502	38.88	-84.75	Enk et al. 2016	13860	Enk et al. 2016
KX027564	63.73	-138.83	Enk et al. 2016	-	-
KX027565	67.48	-139.92	Enk et al. 2016	>45400	Enk et al. 2016
KX027566	63.83	-138.25	Enk et al. 2016	-	-
KX027567	63.83	-138.25	Enk et al. 2016	28960	Enk et al. 2016
KX027560	68.06	-139.78	Enk et al. 2016	-	-
KX027561	68.06	-139.78	Enk et al. 2016	-	-
DQ188829	71	145	Krause et al. 2006	12170	Krause et al. 2006
KX176757	72.68	143.52	Chang et al. 2017	40700	Chang et al. 2017
KX176751	73.34	141.31	Chang et al. 2017	43600	Chang et al. 2017
KX176755	68.733	161.383	Chang et al. 2017	42960	Chang et al. 2017
KX176773	56.51	3.52	Chang et al. 2017	40100	Chang et al. 2017
KX176793	62	58.73	Chang et al. 2017	-	-

GenBank ID	Latitude	Longitude	Locality Reference	Age	Age Reference
KX176767	51.08	83.03	Chang et al. 2017	45700	Chang et al. 2017
KX176770	51.54	7.2	Chang et al. 2017	-	-
KX176769	47.82	12.64	Chang et al. 2017	45180	Chang et al. 2017
KX176768	50.92	84.78	Chang et al. 2017	-	-
KX176785	44.8	34.29	Chang et al. 2017	-	-
EU155210	73.45	102	Poinar et al. 2006	27740	Poinar et al. 2006
MG334278	71.2489	-179.9789	Pecnerova et al. 2017	8318	Pecnerova et al. 2017
MG334266	71.2489	-179.9789	Pecnerova et al. 2017	4643	Pecnerova et al. 2017
MG334279	71.2489	-179.9789	Pecnerova et al. 2017	7470	Pecnerova et al. 2017
MG334270	71.2489	-179.9789	Pecnerova et al. 2017	4024	Pecnerova et al. 2017
MG334269	71.2489	-179.9789	Pecnerova et al. 2017	4079	Pecnerova et al. 2017
MG334265	71.2489	-179.9789	Pecnerova et al. 2017	4726	Pecnerova et al. 2017
MG334274	71.2489	-179.9789	Pecnerova et al. 2017	7336	Pecnerova et al. 2017
MG334264	71.2489	-179.9789	Pecnerova et al. 2017	4969	Pecnerova et al. 2017
MG334281	71.2489	-179.9789	Pecnerova et al. 2017	6380	Pecnerova et al. 2017
MG334280	71.2489	-179.9789	Pecnerova et al. 2017	7194	Pecnerova et al. 2017
MG334276	71.2489	-179.9789	Pecnerova et al. 2017	7060	Pecnerova et al. 2017
MG334285	74	98	Pecnerova et al. 2017	11972	Pecnerova et al. 2017
MG334277	71.2489	-179.9789	Pecnerova et al. 2017	8491	Pecnerova et al. 2017
MG334267	71.2489	-179.9789	Pecnerova et al. 2017	4354	Pecnerova et al. 2017
MG334268	71.2489	-179.9789	Pecnerova et al. 2017	4336	Pecnerova et al. 2017
MG334283	75.16	145.15	Pecnerova et al. 2017	12775	Pecnerova et al. 2017
MG334273	66.4	171	Pecnerova et al. 2017	16901	Pecnerova et al. 2017
MG334272	66.4	171	Pecnerova et al. 2017	14431	Pecnerova et al. 2017
MG334275	71.2489	-179.9789	Pecnerova et al. 2017	14408	Pecnerova et al. 2017
MG334271	71.2489	-179.9789	Pecnerova et al. 2017	41632	Pecnerova et al. 2017
MG334284	71.2489	-179.9789	Pecnerova et al. 2017	15602	Pecnerova et al. 2017
YW890206	66.76	124.12	Palkopoulou et al. 2015	44828	Palkopoulou et al. 2015
YW890205	71.2489	-179.9789	Palkopoulou et al. 2015	4436	Palkopoulou et al. 2015
MF770243	74.13	141.03	Kornienko et al. 2018	32480	Kornienko et al. 2018

Horses. The horse reference panel includes 198 modern and 239 ancient samples, with outgroups *Haringtonhippus*, *Hippidion*, and *Equus ovodovi* included for tree rooting. Most of the

samples are from the last 10kya, but a few had ages up to 100,000kya or older. There are both previously published samples and samples which were published along with the corresponding manuscript in this reference panel (Wang et al., 2021). Metadata for the horse reference samples is summarized in Supplementary Information 9 of Wang et al. (2021). Additionally for initial mapping, we added all available horse mitochondrial genomes on NCBI, which was a total of 432 at the time. This was done differently from the mammoths because of the high diversity in horses, and because the analysis yielded very few results when the mapping was done against a single horse reference. We then used only a thinned subset of the reference panel of 403 reference genomes to construct a reference phylogeny in which to place the permafrost samples.

3.2.2 Permafrost data

From 74 sites and profiles around the Arctic circle, our collaborators collected 1671 total lake and permafrost samples. These were dated with accelerator mass spectrometer radiocarbon dating (AMS 14C), electron spin resonance (ESR), and optically stimulated luminescence (OSL). Of the samples, 652 contained reads that mapped to one of the two mammoth reference genomes. Of these, 104 samples had at least one read that mapped onto the mammoth reference mitochondrial genomes using the ngsLCA method, with a total of 859 reads (Wang et al., 2022). For this subset of samples, the mean number of reads that mapped onto the mitochondrion was 8.3 , and the mean read length was 72. Distributions are shown in Figure 9. In the case of the horses, 123 samples mapped to a horse mitochondrial reference panel using the ngsLCA method, with a total of 446 reads (Wang et al., 2022). For this subset of samples, the mean number of reads that mapped onto the horse mitochondrion was 3.6, and the mean read length was 62. Distributions are shown in Figure 10.

I was provided with metadata including sample IDs, age, geographical coordinates, and region. This metadata is given in Table 4 in the appendix. A map showing the geographical distribution of the samples around the Arctic is shown in Figure 11.

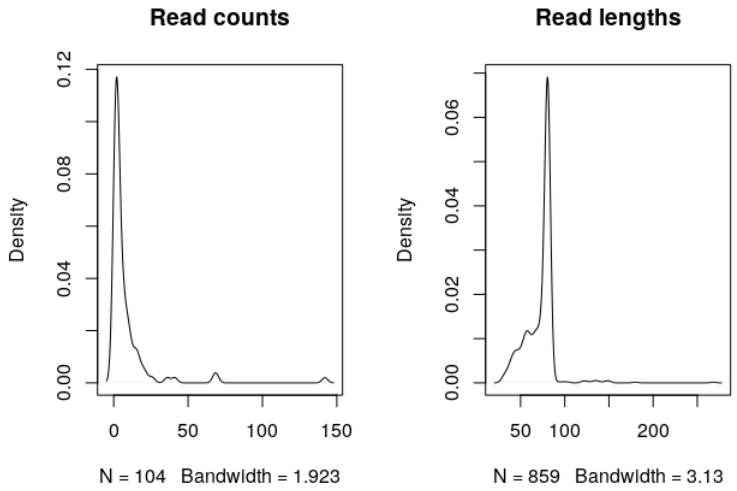


Figure 9: **Mammoths.** Left: Read counts (how many reads per sample mapped to the mammoth mitochondrial genome) per each of the 104 samples. Right: Read length distribution for the 859 total reads.

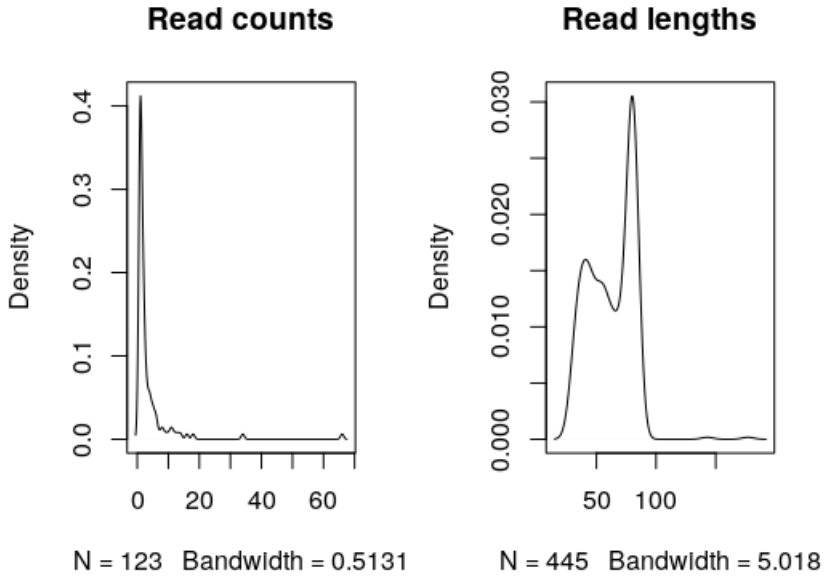


Figure 10: **Horses.** Left: Read counts (how many reads per sample mapped to the mammoth mitochondrial genome) per each of the 123 samples. Right: Read length distribution for the 446 total reads.

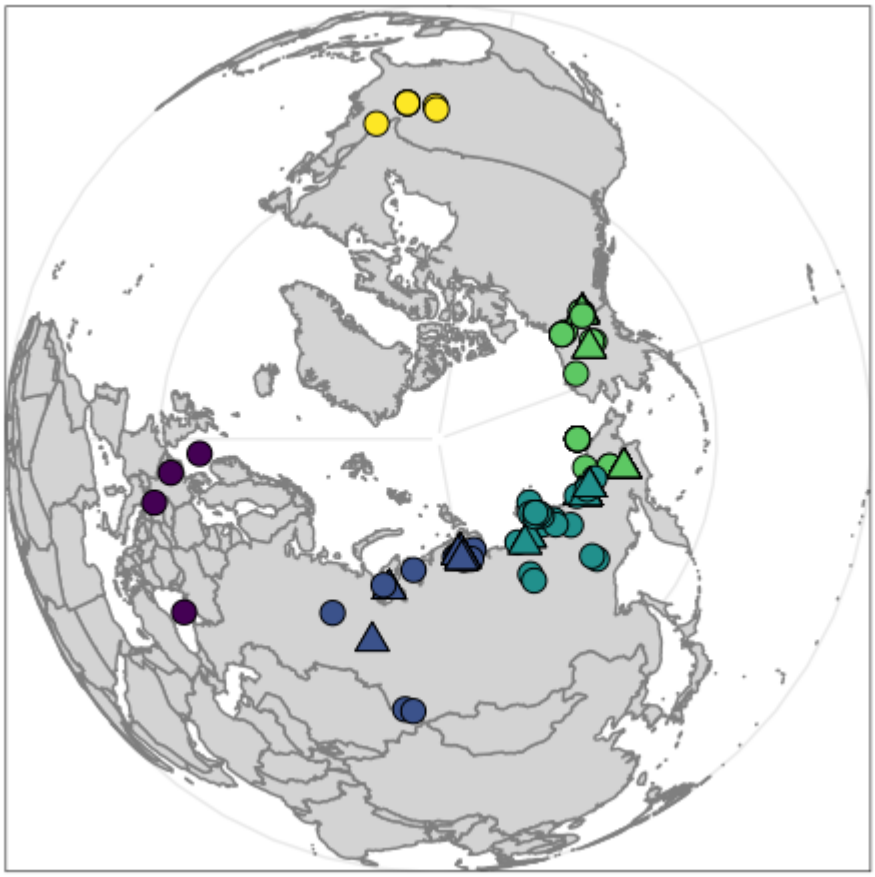


Figure 11: **Mammoths.** A map of sample locations, coloured by longitude. Here, triangles represent our new permafrost samples which had reads mapping to the mammoth reference mitochondrial genome, while circles represent previously published data.

3.2.3 Workflow

Processing the permafrost data: Mammoths. We began with 652 samples that mapped onto one of two mammoth reference genomes. 6 samples were deleted (cr7 2, tm7 8, tm7 9, cr3 32, tm6 4, tm7 5). Three of these were not present in the metadata at all, and three were marked as either "modern" or 326 years old, meaning they could not have possibly been mammoths. Extraneous text in the read names was deleted. Of the remaining samples, 104 contained reads which mapped to the mitochondrion, which was determined by searching for the chromosome name which corresponded to the mitochondrial genome in the .bam files. This name was manually changed to a Y to be compatible with PathPhynder's requirements at the time.

Processing the permafrost data: Horses. Since we mapped to a mitochondrial reference panel in the first place, we did not have to extract mitochondrial reads as in the mammoth case. Likewise to above, the name of the scaffold was manually changed to a Y.

Processing the previously published data. In the case of the mammoths, I downloaded 78 .fasta files from the NCBI database by their GenBank accession files to use as a reference panel. In the case of the horses, I used 403 horse mitochondria .fasta files as described in the previous section as a reference panel. In each case, I concatenated these to create a single file and standardized the sequences. In particular, I made all sequences uppercase, and replaced all instances of nucleotides that were not A, C, T, G or N into an N (for example, some sequences contained M, Y, etc), because most alignment programs will not accept these. I then made a multiple sequence alignment using Muscle 3.81 using default parameters (Edgar, 2004).

Building a VCF file. I then used SNPSites (Page et al., 2016) to get VCF files from the reference panel multiple sequence alignments. There were a total of 860 variable sites in this VCF file for the mammoths, and 3365 for the horses. There are significantly more in the latter case because we used a much larger and more diverse reference panel.

Building a reference tree. In each case, I used the BEAST software suite (Suchard et al., 2018) to build a tree using the VCF file. In particular, I first ran Beauti with default parameters to get a .xml file. I then ran BEAST using default parameters, including Beagle. I then used TreeAnnotator (Suchard et al., 2018) to call a consensus MCC (maximum clade credibility) tree in NEXUS format. Figtree (Rambaut, 2010) was used to convert the tree into newick format.

Processing the VCF file. The VCF output from SNPSites had multiple issues that PathPhynder is unable to handle. First, there were lines with asterisks in the alt allele field, which is supposed to represent a deletion in a newer release of GATK. However, due to the nature of the data and the high frequency of the occurrence of asterisks, I concluded that these were actually representative of missing data. The notation for missing data in a VCF file is a period, so the VCF was transformed

to reflect this. Additionally, there were a few triallelic sites, where the third allele was a C $\rightarrow$ T. This appears to reflect deamination, so the sample with the T was replaced with a missing data marker. This was done via a custom R script. The processed VCF files retained the large majority of their SNPs.

Creating a consensus sequence. Unfortunately, the multiple sequence alignment generated from the reference panels does not necessarily have the same coordinates as any single given reference sequence. Because we require a single reference sequence for PathPhynder whose coordinates match our multiple sequence alignment, we instead created a consensus sequence from the multiple sequence alignment using a Python script. I transformed all of the permafrost files that mapped to the reference panels back into .fq files, and re-aligned them to this new consensus file using bwa with default parameters, filtering for quality score 10 or above (Li and Durbin, 2009). These re-mapped reads were used going forward.

Imputation. Since the reference panel VCFs still had a significant amount of missing data, I performed imputation using the algorithm included in PathPhynder. This algorithm takes as input the newick tree file, and uses the tree in the imputation process. If a site cannot be imputed, it is left as missing data, and not included in downstream analysis.

Placing samples on the tree. PathPhynder first assigns SNPs to branches of the tree. In the case of the mammoths, the algorithm reported 188 sites with missing data and 658 informative positions. In the case of the horses, the algorithm reported 1463 SNPs with missing data and 1868 informative positions. There were 1497 positions that were not used. At this point, we also made sure that this new consensus genome and the bam files had the same contig name by renaming that in the bam files to Y to fit pathphynder's requirements at the time. I then followed a standard PathPhynder workflow to attempt to place the samples on the tree.

An example of the raw pathphynder output for a single sample, in this case ar5_18 in the mammoth phylogeny, is shown in Figure 12. Notice the numbers in the red and green dots. The green dots represent SNPs assigned to branches on the tree that are in support of placing the sample on or below that branch (i.e. the query sample contains the derived allele), and the red dots represent those that are in conflict of placing the sample on or below that branch (i.e. the query sample contains the ancestral allele). The best path is shown in green. We will call the sum of the numbers in the green dots along the best path the support of the placement, and use this measurement as an approximation for the confidence of the placement. It should be noted that almost all samples were placed to an internal node in the tree. This is partially because of the low coverage of our reads, which could lead to few or no matching SNPs on lower branches, but could also be because the sample actually diverged from the other samples in the tree at that internal node.

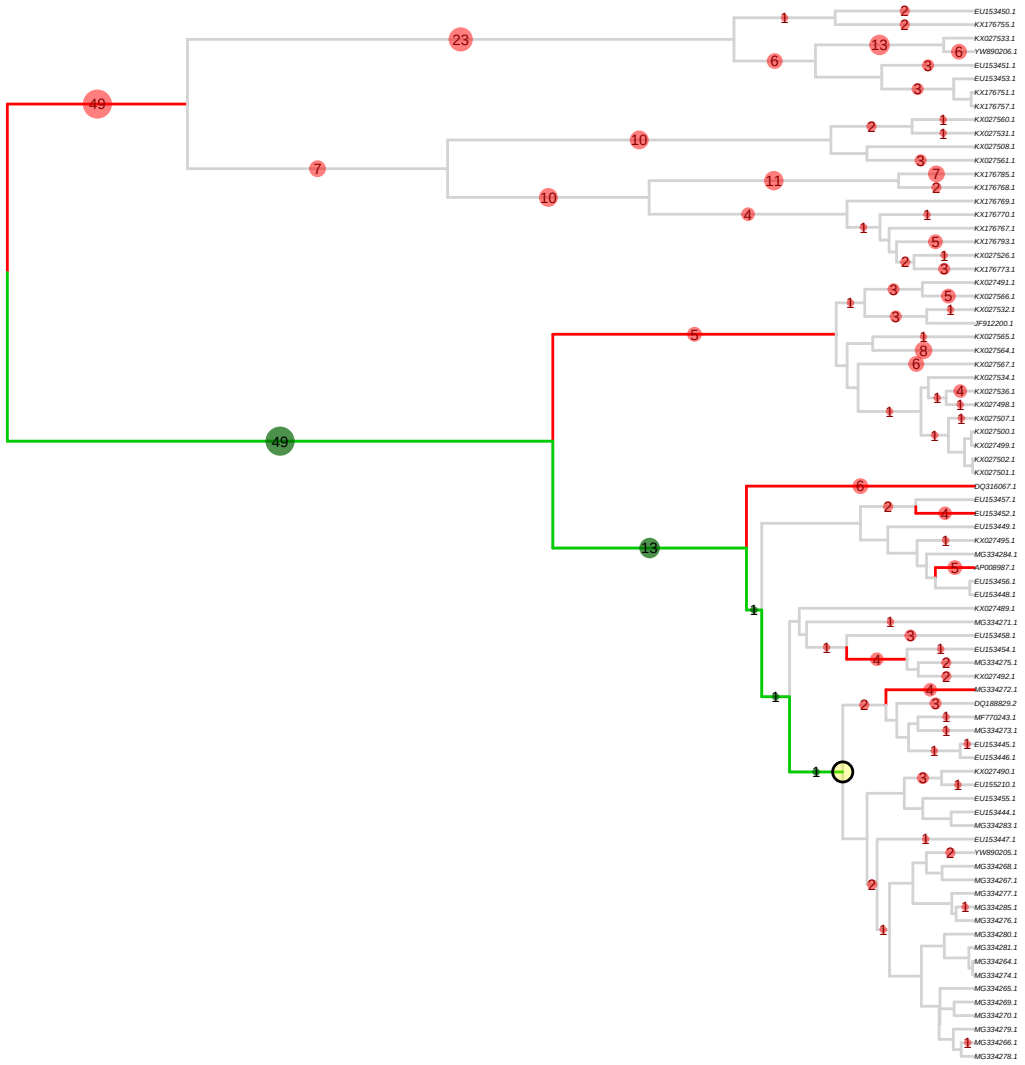


Figure 12: **Mammoths**. An example of PathPhynder output for sample ar5 18, which has been phylogenetically placed into the mammoth mitochondrial reference tree. Numbers in red circles indicate conflicting SNPs and numbers in green circles indicate supporting SNPs. The best path is shown by edges in green, and the best node is shown by a circle at the end of this green path. Larger, zoomed-in versions of this figure are given in the appendix.

Processing output and creating figures. All output from this point on was processed and plotted in R. Packages used include ggplot2 for plots, phytools and ape for tree parsing and plotting, viridis for colours, and ggmap and rworldmap for maps. Figures were edited in Inkscape.

3.3 Results and Discussion

3.3.1 Mammoths

First of all, see the bottom left of Figure 13 for a map of where each sample and previously published genome is located. We have a large number of samples around northeast Siberia. The light green dot which appears to be floating off the northeast tip of Siberia is actually Wrangel Island, the last known refuge of the woolly mammoths (Pečnerová et al., 2017) (see Introduction). Notice that we do not have any permafrost samples from America or Europe, even though mammoths have been found in these regions; for this reason, our geographical coverage is not quite as high as one might like. Therefore, we might not expect to be able to place samples deeply in the European and/or American clades. The resulting mammoth mitochondrial phylogeny, including all of the placed samples and annotated with age and location, is shown in Figure 13. The major clades 1C, 1DE, 3/B and 2/A are labelled, along with the subclades of the latter two. Horizontal dashed lines separate clades. The map is shown again on the bottom left as a colour legend.

There were no successful placements of permafrost samples into Clade 3/B. Furthermore, many of the ages of the previously published genomes here are unknown. The locations of these mammoths are also broadly scattered. Clade 3 contains four reference samples from the northern part of North America (light green in Figure 13), and the rest are from Siberia or Eurasia. This is the smallest and oldest clade, as can be seen in Figure 17, and spans across Europe to North America. The sample cr5_11 was placed at the shared root of Clades 2 and 3, and it has very good support. It is also the youngest of the mammoths in Clades 2 and 3, and quite a few conflicting SNPs internal to both Clades 2 and 3. This raises the exciting possibility that there is an undiscovered clade branching off from the common ancestor of Clades 2 and 3 that is represented by this sample. The full placement plot from PathPhynder for cr5_11 is shown in Figure 14.

In Clade 2/A, we placed 8 samples, some with very good support, and with cr 8_39 even achieving maximum depth in the tree. They agree very well with both the known geographical and age distribution of the clade, giving little new information about mammoth population dynamics, but confirming the effectiveness of the placement method. Additionally, although it is low quality data, these samples could serve to increase the known diversity of Clade 2/A, and perhaps add to future reference data.

In Clade 1, there were two long stretches of samples (in both the top of clade 1, and the top of clade 1DE) which were placed but had generally low support and mimimal depth. It's difficult to see any kind of geoegraphical or age pattern within these samples. Clade 1C is generally known as the American clade. Indeed, there is a string of yellow dots indicating the previously published

samples from the North America. We successfully placed 6 samples from Alaska within Clade 1C, all of which agreed well with both the known location and age of this clade. Further down in Clade 1DE, after the long stretch, we placed 14 samples from Siberia. These were scattered both in regards to age and location. Some samples were placed at maximum depth and found neighbours halfway across the world, such as ar5_17, indicating either fast movement of the species, or a lack of coverage of our reference tree.

Notably, no samples were placed among the Wrangel Island clade, which is the bottom string of green dots on the tree (the last 15 or so samples, all previously published data, generally beginning with MG). This might have been expected, as we did not sample any permafrost from Wrangel Island. Additionally, we were unable to place any samples under 6000 years, supporting the hypothesis that the most recently existing mammoths were on Wrangel Island.

In total, we placed 104 samples on the tree at some level. Of these, the support mean was 6.1. Figure 15 shows a distribution of supports, stratified by clade. By far the best supported placements were in clade 1de, with the highest, ar5_18, with 65 supporting SNPs.

Next we plotted the age distribution stratified by clade, as shown in Figure 16. Notably, this contains all of the previously published data that included an age reference, as well as our samples that were placed into the tree. As expected, clades 2 and 3 were quite old, while clade 1 was on average quite a bit younger.

Lastly, we plotted maps stratified by time and coloured by clade, in Figure 17. A similar figure with a different age stratification, every 10,000 years, is also given in the Figure 18. This figure significantly clarifies the geographic positioning of each of the clades. In particular, clade 3 is mostly in Europe and Western Russia. Clade 2 is in mid to East Russia. Clade 1c is scattered across America, and clade 1de is mid to West Russia, but more so along the coastline. It is also apparent that clade 1de is younger and persisted until recent years, whereas clades 2 and 3 appear to have died out somewhere between 40 and 30 ka bp (see Figure 18).

Notably, this figure also highlights a major finding: there are multiple samples from this study that are younger than 10 ka BP but do not fall on Wrangel Island. This contradicts previous literature, which claims that mainland mammoths went extinct over 10,000 years ago.

If one inspects Figure 13 for those samples, it becomes apparent that they are placed close to the root of Clade 1DE, although some have good support, e.g. cr8_33 has support 18. This suggests that these placed samples could be genetically distinct from the Wrangel Island mammoths, although low coverage of the youngest placed samples means this is still unclear.

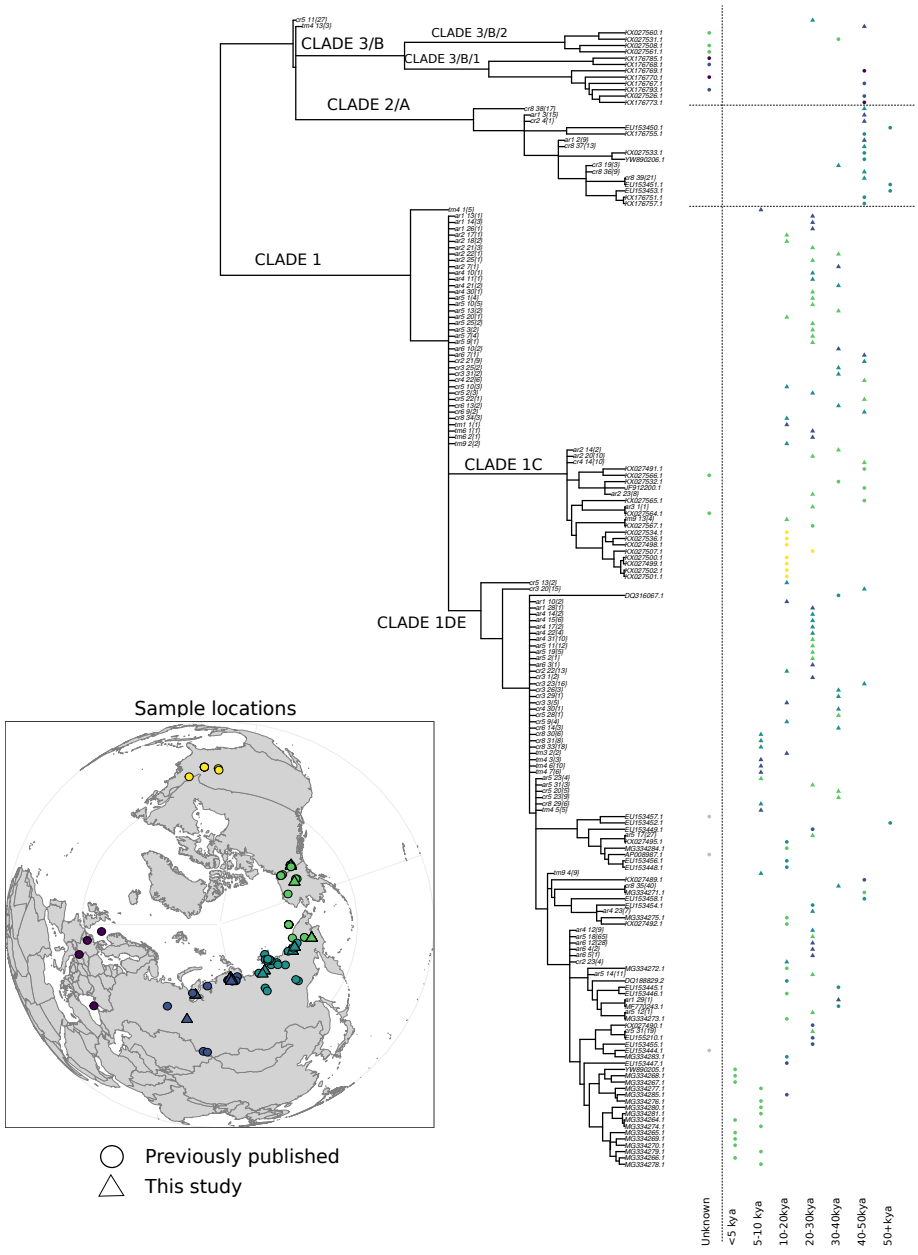


Figure 13: **Mammoths**. A phylogenetic tree of mammoth mitochondrial DNA with the permafrost samples phylogenetically placed using PathPhynder. Reference samples are named ending in ".1", whereas permafrost samples all have the number of supporting SNPs for their placement position in curly brackets. Note that permafrost samples have been phylogenetically placed and so their private edge length is arbitrary (but non-zero so they are easier to see this way), as it was not estimated by the phylogenetic placement algorithm. The tree is annotated with pre-existing clade names, which are also split by dotted lines on the right for ease of reading. On the right is the (binned) age of each sample. The map on the lower left provides a colour legend for the geographic area of each of the samples. Larger, zoomed-in versions of this figure are given in the appendix.

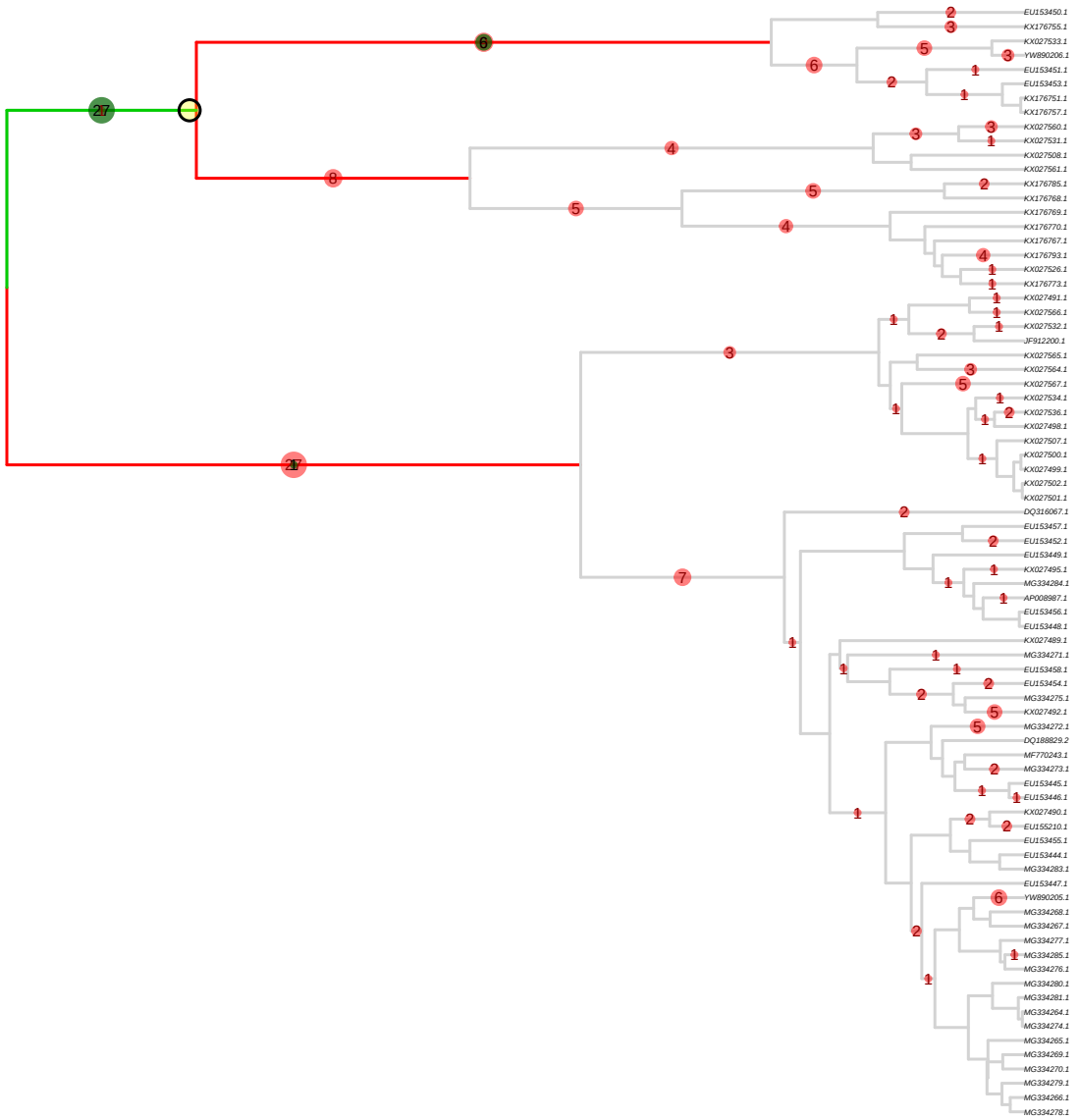


Figure 14: An example of PathPhynder output for permafrost mammoth sample cr5-11, which has been phylogenetically placed into the mammoth mitochondrial reference tree. Numbers in red circles indicate conflicting SNPs and numbers in green circles indicate supporting SNPs. The best path is shown by edges in green, and the best node is shown by a circle at the end of this green path. Larger, zoomed-in versions of this figure are given in the appendix.

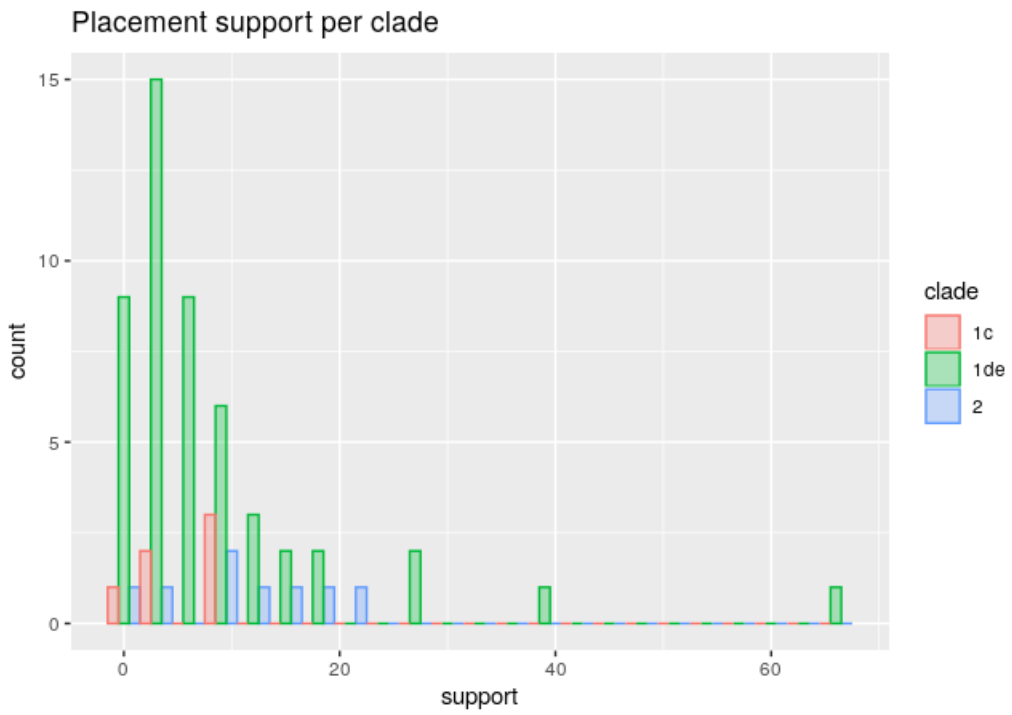


Figure 15: **Mammoths**. Placement support per clade.

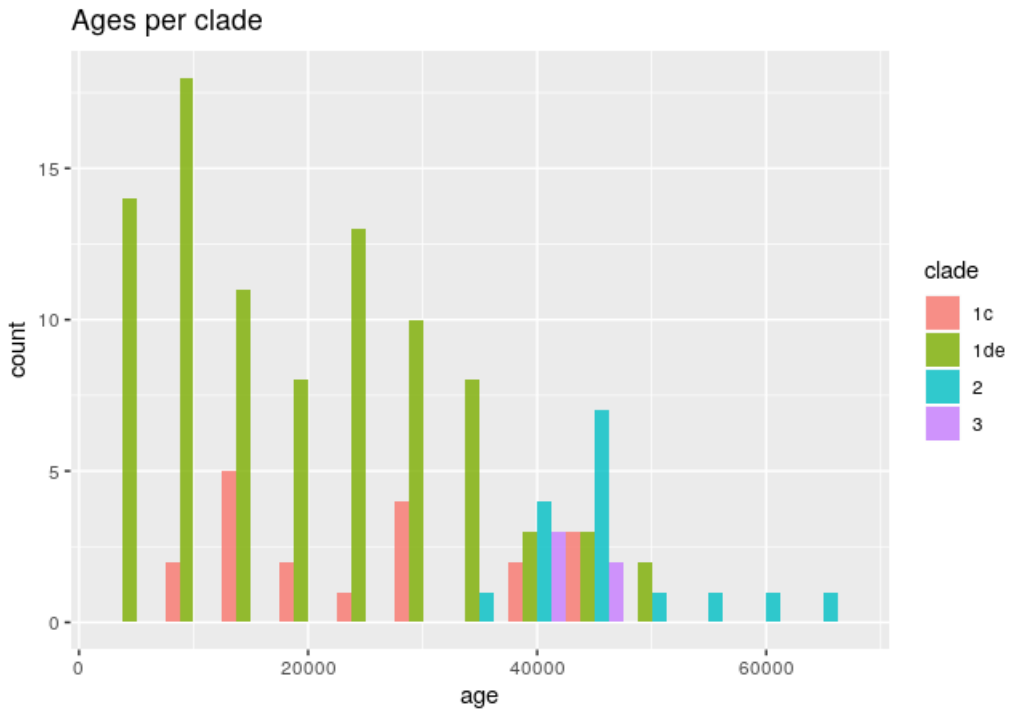


Figure 16: **Mammoths.** Age distribution stratified by clade.

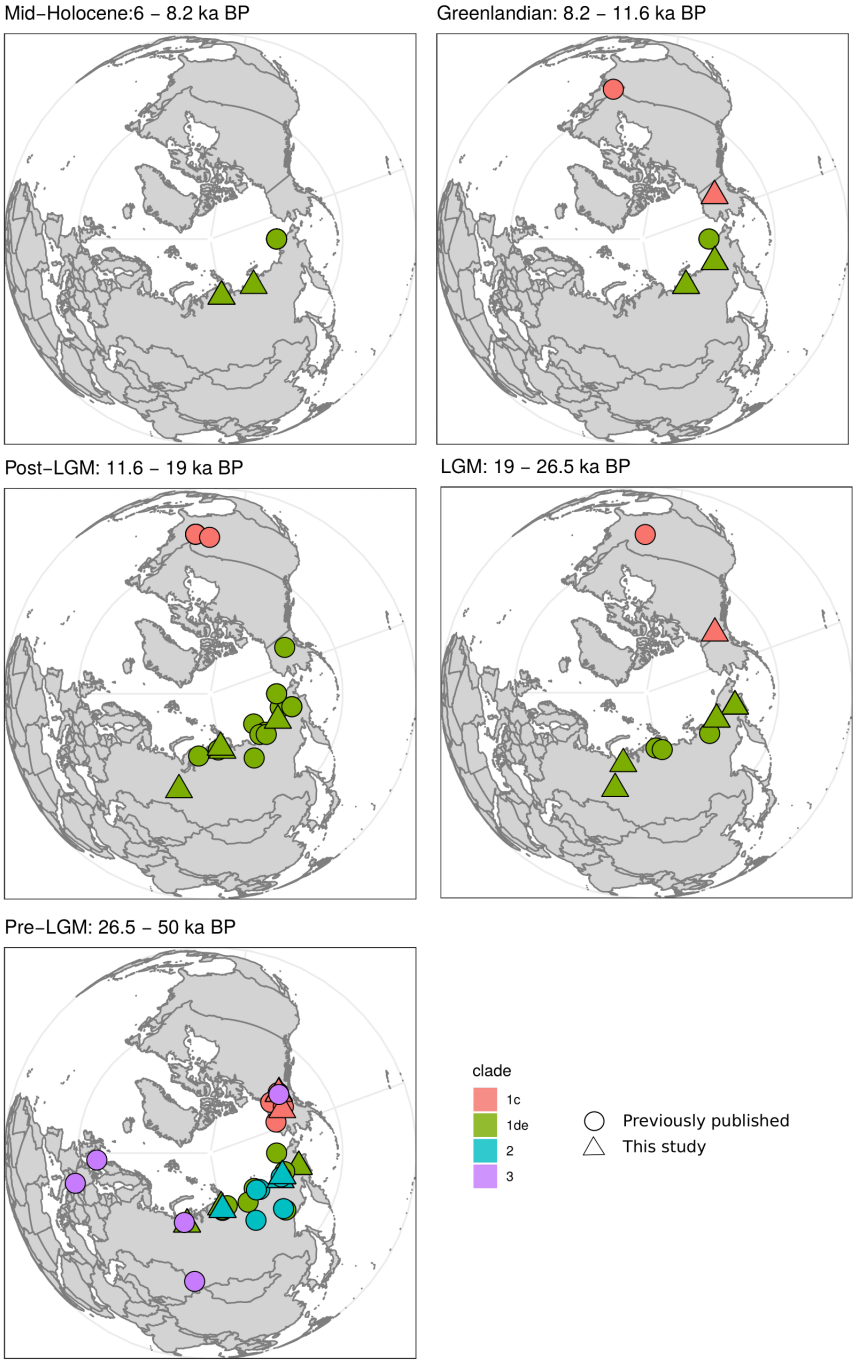


Figure 17: **Mammoths**. Geographic locations of samples, split up by age bin and stratified by clade. Maps of <6 ka BP, and >50 ka BP, are not shown here because they do not include any permafrost samples that mapped to the mammoth mitochondrial tree from this study.

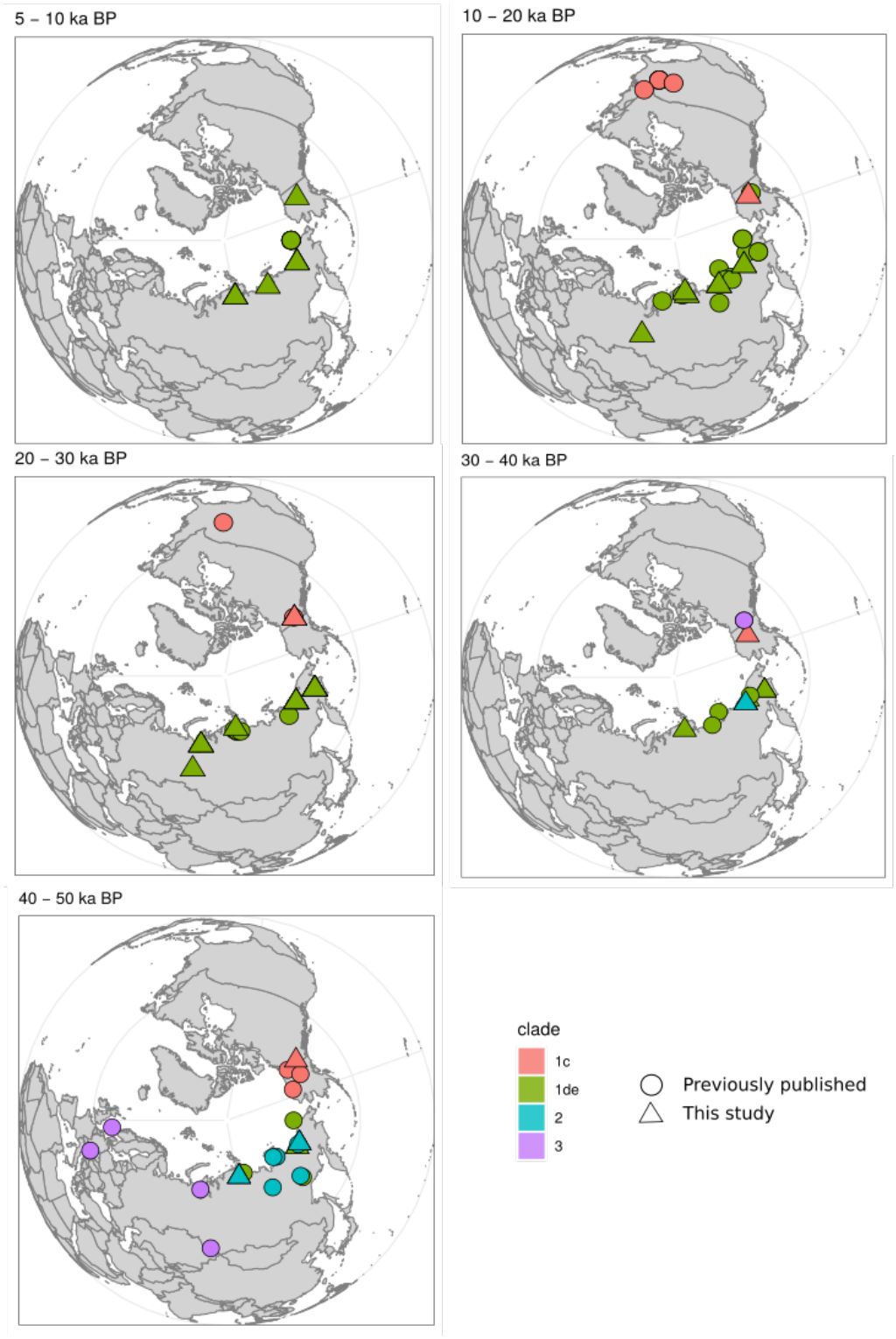


Figure 18: **Mammoths**. Geographic locations of samples, split up by age bin (every 10,000 years) and stratified by clade.

3.3.2 Horses

Of the 123 samples which uniquely mapped to horse, we placed 88 samples in the tree. However, many of these samples were placed near the root, and since the tree contained multiple outgroups to horses (such as the extinct *Hippidion*, *Haringtonhippus* and the extant *Equus ovodovi*, the latter of which includes zebras and donkeys), these placements are uninformative. In particular, many of these placements did not make it into the horse clade, giving no more information than the original read assignment algorithm from which we concluded that these ancient permafrost read sets did indeed originate from a horse.

In order to extract meaningful results and more easily visualize the information, we trimmed the tree of all outgroups (including those permafrost samples which were placed in the outgroups), and thinned the main horse clade (which consists of *Equus caballus, lambei, scotti* and *przewalskii*) by approximately half of the reference samples, taking care to leave them in if they were neighbouring a placed permafrost sample. The final tree is shown in Figure 19. After this thinning, only 22 placed permafrost samples remain (note that many placements in the original tree did not even fall in the main horse clade). The age and location of each sample is shown, indicated by its position and colour respectively. It is immediately clear that most of the samples are from the late Holocene or modern times, and that almost none of these are North American. Indeed, the mitochondrial phylogeny of horses is not known to be geographically structured in general.

However, a single ancient North American clade is evident, and is highly diverged from the rest of the tree. The reference samples in this clade are from Thistle Creek, Yukon, Canada. Somewhat surprisingly, we placed 10 samples into this clade, which consists of 7 reference horse genomes. The three oldest reference genomes, JW119, JW123 and TC21, represent the oldest known *Equus caballus* horses from North America, all dating from around or before 100kya. The other four reference genomes in this clade include two *Equus lambei*, from approximately 24 and 36kya, and two *Equus scotti* that are older than 50kya. This lineage is divergent from the main horse clade, although it is possible that these horses were in the same region, at the same time as other horses from the main clade. The environmental samples that placed into this clade range from 24 to 30kya, which might suggest they represent *Equus lambei* individuals, as the reference *lambei* samples in this clade are from this same age range. A map indicating the age bins and exact locations of this clade, including both reference and permafrost samples, is shown in Figure 20, and a phylogeny in Figure 21. One can see that all of the reference samples came from a small region in the Yukon, as do most of the placed permafrost samples. Indeed, some of our placed samples are from Goldbottom Site (GS, site ID 61), which is only $\sim$ 10 km away from the location where two fossils in this clade were found (see Figure 20). However, there is also a placed sample from the Western tip of Alaska,

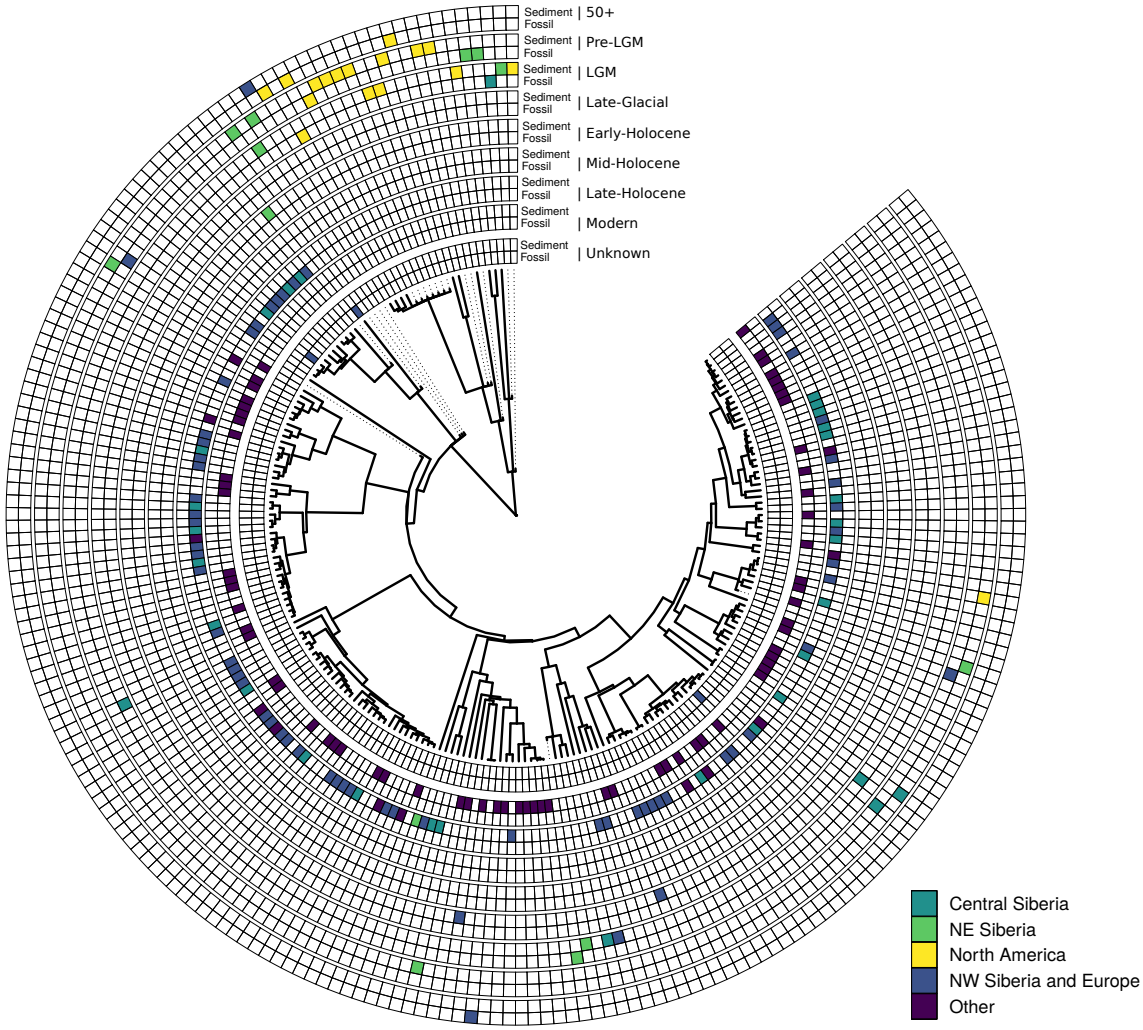


Figure 19: **Horses**: A trimmed and thinned horse phylogeny including all of the successfully placed permafrost samples, annotated with both location and age data.

which extends the known geographic range of this ancient clade of horses.

None of the other placed samples had substantial support or were placed close to a tip, and we cannot draw meaningful conclusions from them. Support plots are not shown here as they were in the mammoth case. This is because they are misleading in this case, as the support values are

heavily inflated by the presence of the highly diverged outgroups.

In summary, the placement of these new permafrost samples into the Yukon clade suggests a long-term continuity of this lineage, and extends its geographic range from near Thistle Creek, Yukon, to cover the Western tip of Alaska.

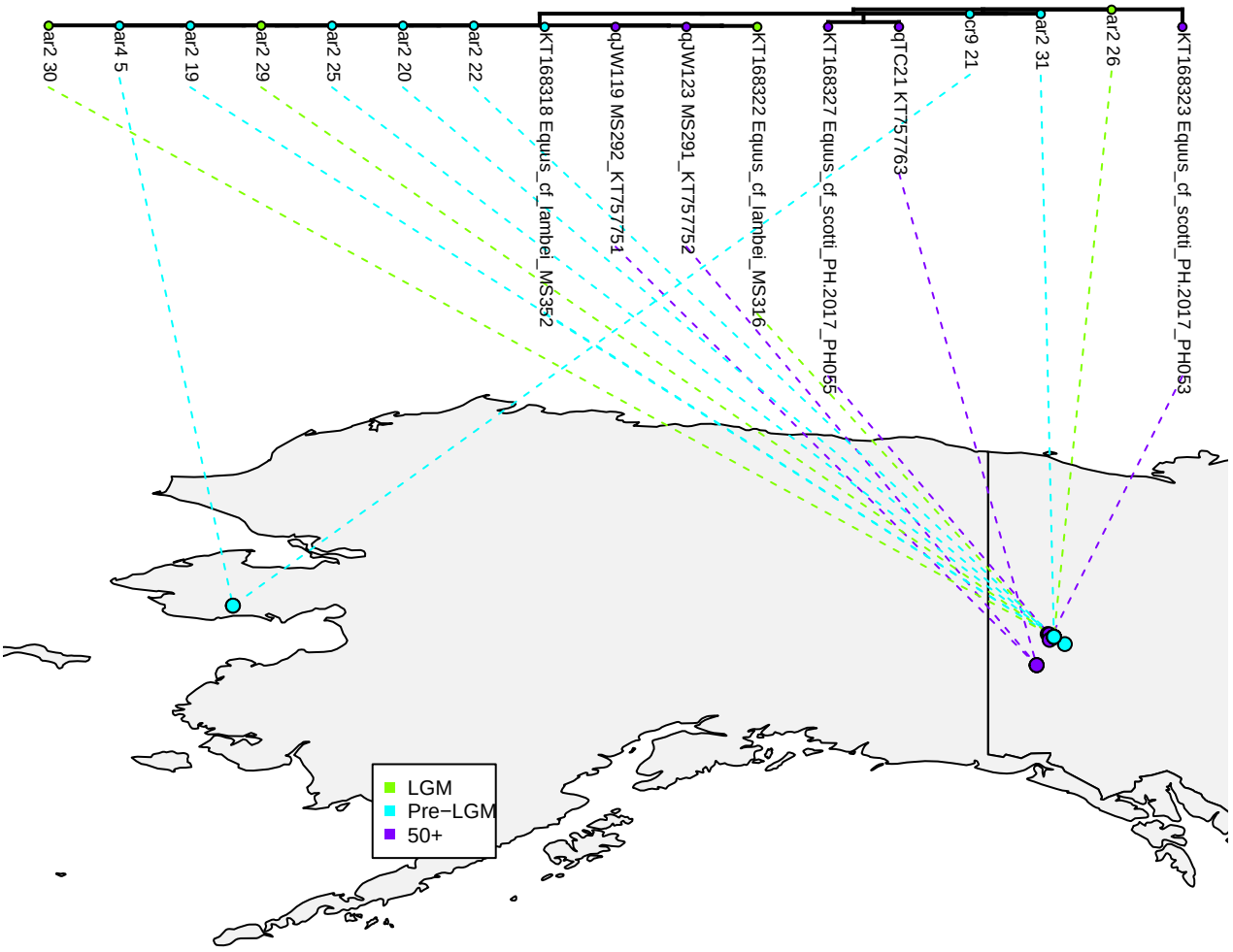


Figure 20: **Horses**: A zoomed-in phylogeny of the Yukon and Alaskan clade and their locations projected onto a map of the area.

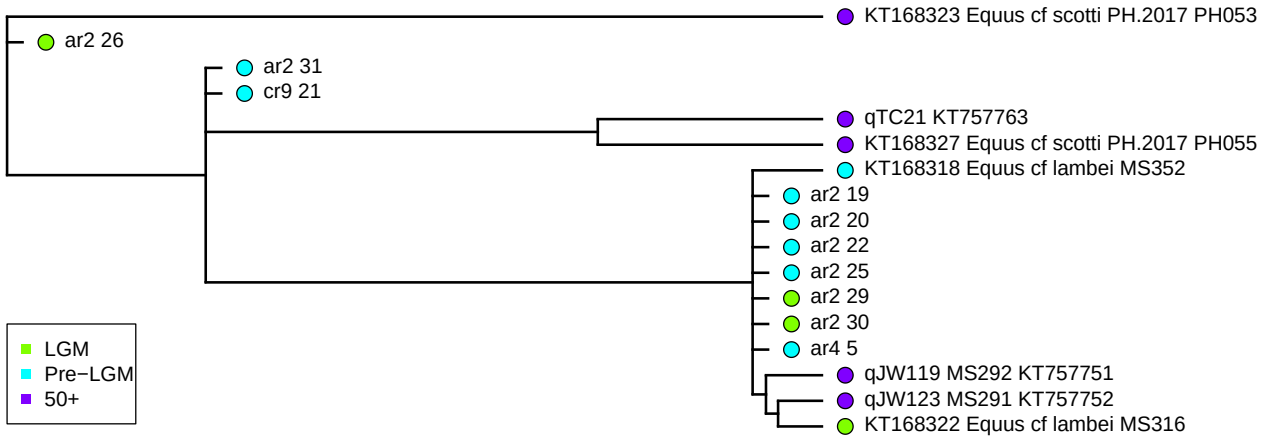


Figure 21: **Horses**:A zoomed-in phylogeny of the Yukon and Alaskan clade.

3.4 Conclusion

Mammoths: We were able to successfully place 104 ancient environmental DNA samples on the woolly mammoth mitochondrial phylogenetic tree, some of which had very good placement support. Many of the samples mapped to clades 1 and 2, but none of the samples could be placed into clade 3. Locations and ages of the samples generally matched well to previously published data, adding confidence to this method. A few samples were placed at maximum depth, i.e. closely related to previously sequenced mammoths, and a subset of these were in completely different locations but similar timepoints to their neighbours. This could indicate either fast migration or a lack of adequate samples at these branches.

No samples were placed among the Wrangel Island clade. However, the youngest samples were all placed shallowly, which means we cannot reject the hypothesis that they belong to the Wrangel Island clade. That is, perhaps we failed to sequence the relevant genetic regions from permafrost samples which would have placed them in this clade. It would make sense for the recent (6-8ka) samples to be somewhat, though not completely, genetically similar to the Wrangel Island clade, and so it is at least somewhat comforting that they were placed with high support in clade 1DE. Furthermore, since there exists no data for mainland mammoths after approximately 11ka, it is not so surprising that these young samples were placed shallowly, for their close relatives are probably not represented in the known phylogenetic tree. Geographically, these young samples are spread out along Siberia. We also found a young ($<$8.2ka BP) clade 1DE sample from Siberia, whereas other clade 1DE samples from the same time period are from Alaska, as shown in Figure 17. Lastly, we uncovered evidence for a potential new clade of mammoths represented by cr5_11, which had 27 supporting SNPs and appears to diverge on the branch leading up to clade 3/B (Figure 14).

Horses: We were able to successfully place 88 ancient environmental DNA samples onto the horse mitochondrial phylogenetic tree, of which 22 were far enough from the root to be meaningful. 10 of these permafrost samples placed into an existing ancient clade of 7 horse genomes from Thistle Creek, Yukon, whose ages range from 24kya to over 100kya, whereas our permafrost samples were between 24 to 30kya. Some permafrost samples which placed in this clade were located very close to existing reference samples, further confirming the reliability of this method. Additionally, one of the placed samples was from the Western tip of Alaska (Figure 20), extending the known geographic range of this clade significantly. These placements add to evidence for a long-term continuity of this ancient clade of horses.

Since placing aeDNA samples into a mitochondrial phylogenetic tree was successful for woolly mammoths and horses, there is no immediate reason it would not work for other species found in ancient environmental DNA. The exploration of different species in the context of this method is a

clear future direction. However, mammoths and horses are some of the best represented mammals in permafrost DNA such as that which was used here, meaning that attempts to obtain placements of other species in their trees might not succeed at the same level. That said, other species, especially those still living, might have significantly more complete mitochondrial trees, allowing for more specific and deeper placement opportunities.

Another future direction is the development of the placement method itself. For example, the usage of support as a measure of placement confidence as done in this study is not ideal, as it does not take into account the conflicting SNPs, or the supporting SNPs on other branches. Though it could be more statistically accurate to design a likelihood method of some sort to determine the best path along the tree, likelihood phylogenetic placement algorithms do not typically output which or how many SNPs are used for placement. This information is useful for gauging the reliability of the results. Lastly, the imputation algorithm could be modified to deal directly with missing data in the new GATK format, instead of requiring the user to pre-process the VCF file (see "Processing the VCF file" in Section 2.3).

4 Environmental Genomics of Late Pleistocene Black Bears and Giant Short-Faced Bears

This chapter is published in: Pedersen, M. W., De Sanctis, B., Saremi, N. F., Sikora, M., Puckett, E. E., Gu, Z., Moon, K. L., Kapp, J. D., Vinner, L., Vardanyan, Z., Ardelean, C. F., Arroyo-Cabrales, J., Cahill, J. A., Heintzman, P. D., Zazula, G., MacPhee, R. D. E., Shapiro, B., Durbin, R., and Willerslev, E. (2021). Environmental genomics of Late Pleistocene black bears and giant short-faced bears. *Current Biology*, 31(12):2728–2736.e8.

Here, I report the retrieval of three low coverage (0.03x) genomes from American black bear (Ursus americanus) and a 0.04x genome of an extinct giant short-faced bear (*Arctodus simus*) from cave sediment samples from northern Mexico dated to 16-14 thousand calibrated years before present (cal kyr BP), which I contextualize with a new high coverage (26x) and two lower coverage giant short-faced bear genomes from 22-30 cal kyr BP old Yukon fossils. Using 83 published black bear samples from across North America (Puckett et al., 2015), I show that the Late Pleistocene black bear population in Mexico is ancestrally related to the present day eastern American black bear population. Furthermore, using the new *Arctodus* fossil genomes, I conclude that the extinct giant short-faced bears present in Mexico were deeply divergent from the earlier Beringian population. These findings demonstrate the ability to separately analyse genomic-scale DNA sequences of closely related species co-preserved in environmental samples.

4.1 Introduction

In this chapter, I investigate whether it is feasible to retrieve and do population genetic analyses using genome-wide data directly from ancient environmental DNA. Cave sediment samples were obtained from Chiquihuite Cave, Astillero Mountains, North Mexico, that were screened previously for the presence of American black bear or *Ursus americanus* DNA (Ardelean et al., 2020) and selected three strata in which black bear DNA was present for further processing. The first two strata, UE1210 and UE1212, have been dated to 16-15 thousand calibrated years before present (cal kyr BP) by Ardelean et al. (2020), after the peak of the last glacial maximum (LGM) but prior to the onset of Holocene warming at $\sim$ 12.0 cal kyr BP, and radiocarbon dates from three charcoals place the last strata UE1605 between 15.0-13.0 cal kyr BP. Each of these three samples UE1210, UE1212 and UE1605 contain black bear DNA, and the sample UE1605 also contains reads from the extinct giant short-faced bear.

Currently, three bear species exist in North America: the polar bear *Ursus maritimus*, the grizzly bear (also called brown bear) *Ursus arctos* and the black bear *Ursus americanus*. Particularly the latter two have become heavily restricted in their geographical distribution in historical times, and their past population structure remains almost unknown. During the Pleistocene a fourth bear species roamed the North American steppes - the giant short-faced bear, *Arctodus simus*. This bear is the largest known carnivore of the Pleistocene, about three times larger than the grizzly bear and went extinct around 13.8-11.4 ka BP (Stuart, 2014). Fossil remains from the giant short-faced bear are rare and only 51 individuals (NEOTOMA db, (Williams et al., 2018)) have been documented, despite its continuous existence in North America for 2.58 Mya (Schubert et al., 2010). Geographically, its presence in Chiquihuite Cave, Mexico, also marks one of the most southern finding of its distribution to date, which further underlines its importance for our understanding of the Great American Biotic Interchange (GABI) and the glacial refugias that existed during the LGM. Despite its enigmatic size, little is known about its genetic affinity to the other bear species and its taxonomic relationship to extant bear species remains debated and its geographical distribution outside North America is unclear.

Here, I present results of genome-wide analysis and separation of DNA from American black bear and giant short-faced bear from three cave sediment layers. I first contextualize the ancient Mexican black bears with 83 published RADSeq American black bear genomes (Puckett et al., 2015). Using this reference panel and three ancient environmental black bear DNA samples from Chiquihuite cave, I reconstruct the evolutionary history of the black bear during the last few tens of thousands of years. Next, I compare the giant short-faced bear aeDNA sample to three higher quality fossil sequences. Lastly, I build a mitochondrial reference phylogeny of *Ursus*, and use

pathPhynder alongside a competitive mapping pipeline to separate out the mitochondrial reads and illustrate two paths on the same phylogeny leading to the two species.

4.2 Methods

4.2.1 Experimental methods and mapping pipeline

Details of the experimental methods, including the sampling, age determination, and environmental DNA laboratory methods can all be found in (Pedersen et al., 2021). I will not cover these methods here in detail it is not work that I have personally completed. Similarly, the preliminary mapping analysis was not my own work and can be found in detail in (Pedersen et al., 2021), but I cover it briefly in the paragraph below for completeness.

DNA was recovered from from a total of 48 sediment subsamples within the three stratigraphic layers of Chiquihuite cave. Extracted ancient DNA (aDNA) was converted from these samples into 65 libraries for Illumina shotgun sequencing. Competitive mapping against the RefSeq mitochondrial database (Howe et al., 2020) confirmed the presence of American black bear DNA across all three sedimentary layers UE1210, UE1212 and UE1605, and the presence of giant short-faced bear in UE1605, as shown in Figure 22. Note that the Andean bear or spectacled bear *Tremarctos ornatus* is the closest living relative to *Arctodus simus*. To determine the presence of a species in our aeDNA, one can look for both elevated 5' C>T misincorporations on the first position (since damage from deamination primarily impacts the 5' end, as discussed in Chapter 1), as shown in Figure 22a. To double check the presence of these specific species rather than related species, reads were also mapped against the mitochondria of four related bear species (giant panda, Andean bear, American black bear and polar bear) and check the fraction of reads that map with zero edit distance, as shown in Figure 22b. This confirms the presence of black bear, or Ursus americanus, DNA in all three samples, and the presence of giant short-faced bear, or *Arctodus simus*, DNA in UE1605. This preliminary analysis resulted in between 1-1.6 million reads aligning to American black bear with a coverage of 0.025x, 0.019x and 0.033x for UE1210, UE1212 and UE1605, respectively, as well as a coverage for giant short-faced bear of 0.041x for UE1605. The presence of giant short-faced bear DNA in either UE1210 or UE1212 was not concluded.

4.2.2 Black bear analysis

To contextualise the ancient black bear genome-wide data, I first had to build a reference panel. To do this, I obtained fastq RAD-seq files from Puckett et al. (2015) of 83 modern black bears from across the United States, along with metadata. Locations of these black bears are shown

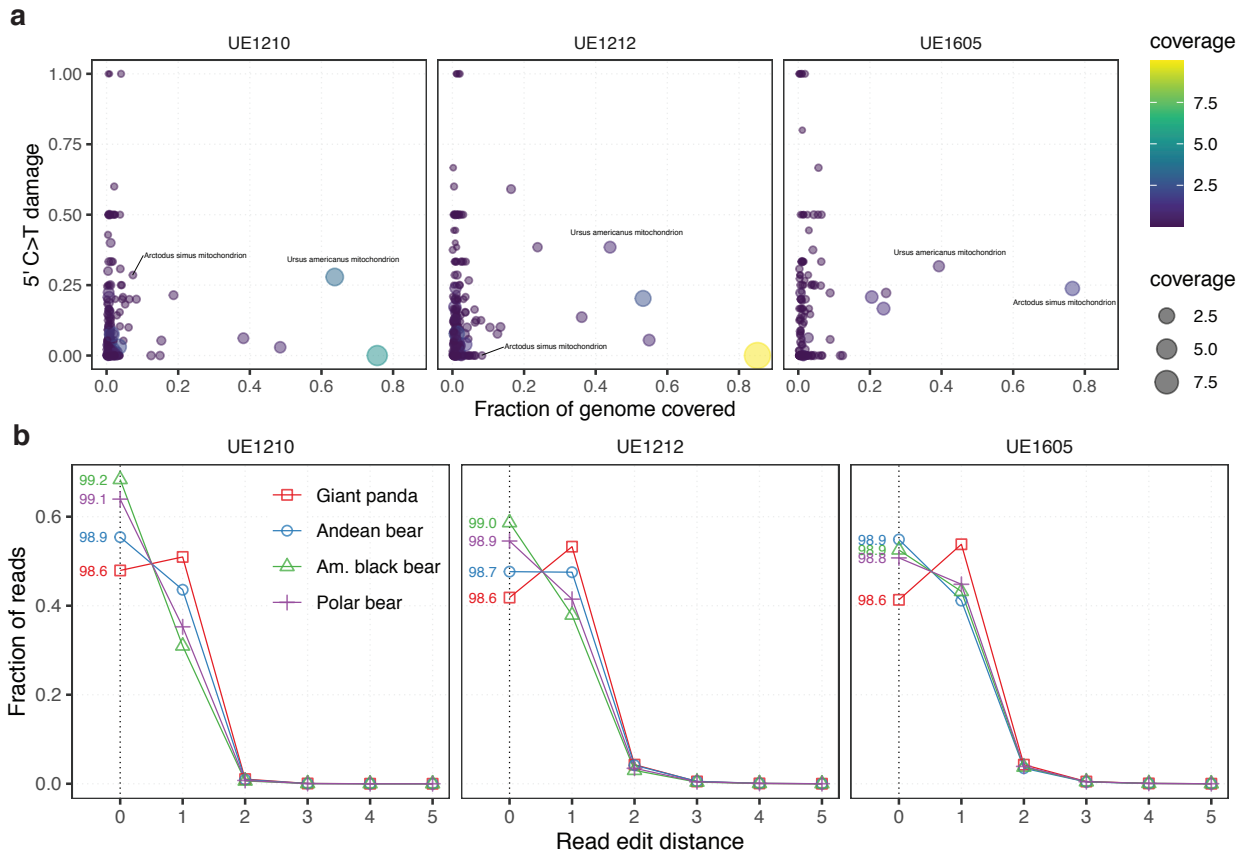


Figure 22: Exploratory mapping results. a. Mitochondrial coverage and C>T frequencies on 1st position against the RefSeq mitochondrial database. b. Read edit distances against four whole reference genomes of bears. Values on the left indicate the average nucleotide identity of the all mapped reads to the respective reference.

in Figure 24a. I realigned these original fastq files against the more recent black bear reference genome (LZNR01000000, (Srivastava et al., 2018)) using bwa aln with default parameters (Li and Durbin, 2009). I called a vcf using bcftools with default parameters (Li, 2011) and filtered for a read mapping quality of >20 and AN>150, that is, those sites which were covered by at least 90% or 75 of the 83 individuals, using samtools (Li, 2011). The latter was done to ensure I used variants for which the majority of the samples had genetic information and resulted in 101,961 SNPs to be parsed for phylogenetic analysis.

I next used Plink 1.9 (Purcell et al., 2007) to create a distance matrix of only the modern samples, then constructed a neighbour-joining tree with all modern samples in R using the phytools package (Revell, 2011). Genomic coordinates were then called using Plink to generate a .bed file of coordinates containing biallelic SNPs according to the vcf of the modern samples. On these co-

ordinates, a pileup was created on the three ancient samples and converted to Plink format using a custom Python script. This resulted in 2646 pseudo haploid SNPs for UE1210, 1927 for UE1212, and 2954 for UE1605. I next merged the modern and ancient Plink files, and used Eigenstrat's SmartPCA (Patterson et al., 2006) with shrinkmode and lsqproject options to project the ancient samples onto the modern variation. PC1 accounted for 5.13% of the variation and PC2 accounted for 2.94%. I plotted the figure in R, rotated to approximately correspond with the geographical structure of the populations (Figure 24b).

To measure the relative genetic distance of each ancient sample to each of the modern individuals, I merged all Plink files and created a pairwise genetic Hamming distance matrix on biallelic SNPs using the -dist command with the flat-missing modifier. For the missing values in the ancient samples, Plink rescales the distances to be on the same scale as the rest of the matrix. Without the flat-missing modifier, Plink assumes that the minor allele frequency is independent of the missingness proportion, which is a poor assumption in this case since all three samples with high missingness are from the same Mexican population. I then mapped the scaled distances of each ancient sample to each modern sample onto a colour scale, and plotted the colours on a phylomap (Zhang et al., 2011) plot to visualize the distance of each ancient sample to each modern sample. The phylogenetic tree shown in this plot is the neighbour-joining tree produced by a distance matrix of only the modern samples using Plink. This is shown in Figure 24c for UE1210, with additional figures in Figure 25 for UE1212 and UE1605.

Next I wanted to create an admixture graph of the black bear populations, integrating the ancient Mexican population. To calculate f4 statistics and create an admixture graph, I first needed an outgroup on the same coordinates as the black bear reference genome. I mapped two polar bear short read genomes SAMN01057659 and SAMN0105763651 (Miller et al., 2012) onto the black bear reference genome (Srivastava et al., 2018) using bwa mem (Li and Durbin, 2009) with default parameters, and filtered for read quality >30 using samtools (Li et al., 2009a). I compiled the two polar, the 83 modern, and the three ancient black bears into a single vcf file using bcftools (Li, 2011). I labelled samples as belonging to one of the following populations: Polar, Mexican, East, Southwest, Kenai (Alaska West), and SEAK (Southeast Alaska), where Mexican refers to the ancient samples and the other labels and groupings were decided using both phylogenetic and geographic factors of the modern population, and previous literature (Puckett et al., 2015). I removed the Northwest population since that population was concluded to be admixed in previous literature (Puckett et al., 2015) and therefore may have unnecessarily complicated the current analysis. I also removed three SEAK samples from the southern Alexander Archipelago that clustered separately from other SEAK samples in the PCA and phylogenetic analysis (see Figure 1b,c) as well as in pre-

vious coancestry heat maps (Puckett et al., 2015). I converted this vcf to Eigenstrat format using a custom script from Meier (2021), and used the admixtools package (Maier et al., 2022) in R to compute f4 statistics.

On the same dataset, I then used the qpGraph function in the admixtools package in R to determine an admixture graph. I used the maxmiss=1 option so as not to drop any SNPs with missing data in any of the individuals, with 500 SNP blocks for the jackknife, and default options otherwise. I first used automatic graph optimization, allowing for one admixture edge, to determine a graph using the East, Southwest, Kenai (West Alaska), SEAK (Southeast Alaska), Mexican and polar populations. Since this graph fit poorly and had excess f4 residuals with z-scores over 6, I added another admixture edge at all possible positions, resulting in seven highest-scoring graphs with similar topologies that fit the data well, each with a maximum excess f4 residual of $|Z| = 2.182$. Each of these graphs agreed on some basic structural characteristics, including a deep split between Mexican/East/Kenai and Southwest/SEAK, with Mexican basal on the Mex/East/Kenai side, and with both Southwest and SEAK admixed. Furthermore, in each graph the SEAK population took most of its admixture from the Mexican/East/Kenai clade, from a population most closely related to Kenai. I show the best-scoring of these graphs, with a score of 4.922, in Figure 24d, and the remaining six in Figure 23 (lower scores are better).

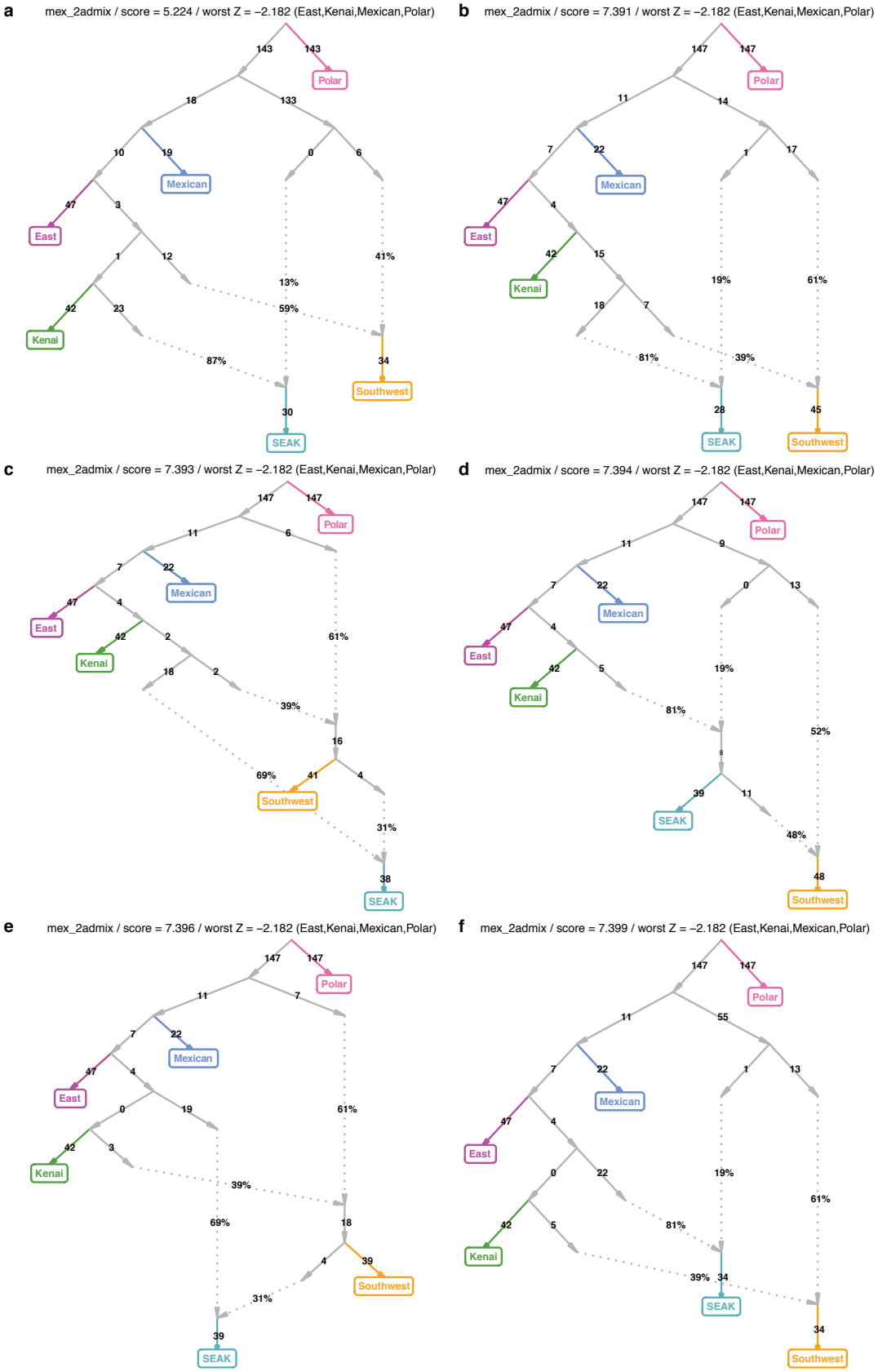


Figure 23: Our admixture analysis using admixtools produced seven best black bear admixture graphs. The best of these is shown in Figure 24d, and the remaining six are shown here (a lower score is better). Each of these has the same worst excess f4 residual z-score of -2.182 for the configuration (East,Kenai;Mexican,Polar), but scores slightly differently. All seven graphs share many common characteristics, as noted in the Methods.

4.2.3 Giant short-faced bear fossil analysis

To aid phylogenetic placement and separation of the reads from both bear species, I acquired access to three *Arctodus simus* fossil bones previously found in Yukon, Canada (See Pedersen et al. (2021) Supplementary Information, SI Text). These included a petrous bone from a complete skull found at Ophir Creek near Dawson Creek (YG 24.1), a radius bone found at Hester Creek (YG 76.4), and a right femur found in Canyon Creek (YG 546.562). DNA was extracted from these fossils was used in later analysis to contextualize the *Arctodus simus* aeDNA. Since I had no part in either the experimental or bioinformatic methods directly on the *Arctodus simus* fossils (except when contextualizing the aeDNA, as described in the next article), I will only briefly summarize these in the next few paragraphs for completeness. Details are in Pedersen et al. (2021).

The three fossil samples all yielded Late Pleistocene calibrated ages, with YG 24.1 dating 22.3 cal kyr BP, YG 76.4 dating 28.9 cal kyr BP, and YG 546.562 dating 29.8 cal kyr BP. DNA extraction and sequencing resulted in a total of 43,107,072, 50,492,295 and 758,541,872 reads aligning to the Tremarctos ornatus genome (the closest relative to *Arctodus simus* with an assembled nuclear genome) with a coverage of 1.82x, 1.66x and 26.01x, for YG 24.1, YG 76.4, and YG 546.562, respectively.

The pairwise sequentially Markovian coalescent (PSMC) model (Li and Durbin, 2011) was used to estimate the historical effective population size of YG 546.562 (Figure 27d), with a generation time of 6 years and a mutation rate of 0.6e-8 per bp per generation, based on previous estimates (Kumar et al., 2017). This is shown in Figure 27d. The high coverage short-faced bear genome also allowed for the estimation of a timeline for the divergence between tremarctine and ursine bears (Figure 27c). In particular, from the high-coverage short-faced bear genome and published genomes of the eight extant bears in the ursid family, divergence times were estimated among the bear species using an approximate likelihood calculation with MCMCTree (Yang, 2007) under an independent clock model with one fossil calibration and one tip date for the giant short-faced bear (Figure 27c).

Next, the fossil bear DNA was aligned with eight extant and three extinct bear species, using publicly available mitochondrial sequences for sun bear, sloth bear, cave bear, Asiatic black bear, giant panda, brown bear, polar bear, Andean bear, ABC island brown bear, and American black bear (FM177765.1, FM177763.1, FM177760.1, FM177759.1, EF212882.1, EU497665.1, GU573490.1, FM177764.1, JX196368.1, AF303109.1, respectively), and a mitochondrial phylogeny was created using RAxML (Kozlov et al., 2019).

4.2.4 Giant short-faced bear eDNA analyses

I sought to contextualise the giant short-faced bear eDNA sample UE1605 by placing it phylogenetically in the wider ursid tree. From the above multiple sequence alignment of 14 ursid mitochondrial genomes, which included the three *Arctodus* fossil mitochondria, I created a vcf using SNP-sites with default parameters (Page et al., 2016), and filtered out sites which contained non-ACTG bases in the reference or were not biallelic, leaving 5071 sites. I also called a consensus sequence of length 16981 on the Ursid mitochondrion multiple sequence alignment using EMBOSS cons with default parameters (Rice et al., 2000).

To phylogenetically place our ancient environmental sample UE1605, I used pathPhynder (Martiniano et al., 2022), as mentioned and used in previous chapters. Since the eDNA samples contain both black bear and giant short-faced bear DNA, I used Picard's (Broad Institute, 2019) FilterSamReads function to partition the .bam files into three sets: reads that mapped uniquely to the Andean bear reference mitochondrion (NC_011116.1,(Krause et al., 2008)), reads that mapped uniquely to the black bear reference mitochondrion (NC_003426.1,(Delisle and Strobeck, 2002)), and reads that mapped to both. I then used bedtools bamtofastq (Quinlan and Hall, 2010) to convert each of these read sets back to fastq format, and then bwa aln -l 1024 -n 0.02 (ancient DNA parameters, (Li and Durbin, 2009)) to re-map these reads to the consensus ursid sequence, because I needed the ancient sample to be on the same coordinate system as the reference multiple sequence alignment. I then gave as input to pathPhynder the ursid phylogenetic tree, the filtered ursid vcf file, the ursid mitochondrion consensus sequence, and our UE1605 read sets mapped to the consensus. I used the best-path mode and the transversion only filter (to avoid errors from deamination) and otherwise default parameters. For each read set, I ran a custom Perl script on the pathPhynder output, and thus were able to determine which biallelic transversion SNPs in our UE1605 sample mapped to Andean bear uniquely, black bear uniquely, or both, and which of each of these were in support or conflict on each branch of the phylogeny. The two best paths and three partitions were plotted in Figure 27a.

I next wanted to compare UE1605 and the three fossil giant short-faced bears on the nuclear genome. First, I called a vcf of biallelic SNPs on the three fossil samples using bcftools (Li, 2011) with default parameters, and converted this to Plink format (Purcell et al., 2007). I used Picard's (Broad Institute, 2019) FilterSamReads function on the UE1605 reads which mapped to the Andean bear reference genome to filter out the reads which also mapped to the black bear reference genome (LZNR01000000, (Srivastava et al., 2018)), to get the reads which mapped uniquely to the Andean bear reference. From these uniquely Andean bear UE1605 reads, I created an mpileup using samtools (Li, 2011) on the fossil vcf coordinates using a min-BQ of 20 and otherwise default

parameters, and used a custom Python script to convert the UE1605 mpileup to Plink format. I merged the two Plink files, so that I had a single file containing all four giant short-faced bear samples, which was then filtered to contain only transversions. This left 2505401 biallelic transversion SNP sites, on which I created a pairwise genetic Hamming distance matrix using Plink with the flat-missing modifier. These distances are shown on the off-diagonal of Figure 27b along with a hierarchical clustering tree computed on these distances using pheatmap in R (Kolde, 2018). Next, I filtered the fossil vcf to contain only transversions, and computed individual heterozygosity proportions on the same set of 2505401 biallelic transversion SNP sites using a custom script. These heterozygosity proportions are shown on the diagonal of Figure 27b.

4.3 Results and discussion

4.3.1 Black bear

Using a panel of 83 present-day American black bears, I found that the black bear genomes recovered from the three Mexican sediment layers are closely related to modern black bears from eastern North America, but also share ancestry with bears in present-day Alaska. Based on a combination of genetic data and topological features likely to impede gene flow, I assigned genomic data from 83 modern black bears (Puckett et al., 2015) to 5 geographically distinct populations in the United States: Kenai Peninsula (Alaska), Southeast Alaska (SEAK), Northwest, Southwest and East (Figure 1a). I then projected the three ancient eDNA samples into a principal components analysis of the modern black bears using SmartPCA (Patterson et al., 2006) (Figure 24b). If one does not use a method to account for this, the samples with missing information tend to get pushed to one extreme, while all the others cluster together. Here, the relative positions of the modern samples resemble the PCA done in (Puckett et al., 2015). I projected each of the ancient Mexican samples separately. All three ancient samples clustered together and are closest to the East population here. However, it should be noted that there is a potential bias in the projection principal component analysis towards the population with highest diversity (that is, the East population).

I next estimated a neighbour-joining tree of the modern samples (Figure 24c). The rooting of this tree is somewhat arbitrary, and one could argue that both Alaskan populations, Kenai and SEAK, belong in the same clade as the East population. I chose this rooting for ease of visualization. It is also important to remember that, as shown in (Puckett et al., 2015), there has been significant amounts of admixture in the black bear population. Regardless, a phylogenetic tree can be a helpful tool to visualize the structure of the population, and ours recovers expected patterns, and corresponds well to geographic structure when projected onto a map to create a phylomap

(Zhang et al., 2011), as done here. I coloured the modern samples in a phylomap according to their genetic Hamming distance from each of the ancient Mexican samples, which I rescaled using Plink to account for missing data. Each of the three samples showed very similar patterns in this regard. Mexican black bears clustered most closely to the East population (UE1212: Figure 24c, UE1210 and UE1605: Figure 25a,b), and closer to both Alaskan populations (Kenai and SEAK), than to the Northwest and Southwest populations. This reinforces what was found in the principle component analysis. Furthermore, this genetic Hamming distance analysis would not suffer from the same bias towards the population of highest diversity as the principal component analysis might.

Admixture analysis revealed that the eastern lineage, to which I find that the Mexican bears belong, was the earliest to diverge from other present-day populations of American black bears (Figure 24d). I used admixtools (Maier et al., 2022) to obtain an admixture graph using the three Mexican black bears, the modern East, Southwest, SEAK and Kenai populations, and two polar bears for an outgroup. The best-fit admixture graph (Figure 24d) indicates that the ancient Mexican population diverged from the ancestral East population after the initial divergence between the eastern and western lineages of black bears. Divergence of the eastern lineage continued into branches that produced most Alaskan ancestry. Further, this diverged eastern lineage admixed with the western lineage in an ancestral population to the modern Southwest. A second admixture event occured with a western population to produce the modern SEAK population.

These results expand and refine the working model of American black bear phylogeography, with the main hypothesis shown in Figure 26. Black bears first appeared in North America in the Late Pliocene (Wang et al., 2017; Rybczynski et al., 2013), where they live today as forest generalists able to utilize resources from diverse forest compositions ranging from subtropica to boreal. Previous work reported that American black bears cluster into two major lineages in the eastern and western parts of the continent, and estimated that these lineages diverged 67 cal kyr BP, possibly becoming separated by expanding grasslands across the central continent (Puckett et al., 2015). However, genomic similarities between black bears in the East and those living in the most northerly population in Alaska (Puckett et al., 2015) suggested that the lineages may have remained connected during the Late Pleistocene, perhaps by forested habitat that spanned latitudinally across the northern continent, as they are today (Pelletier et al., 2012). The population admixture graph shown here supports this hypothesis, and gives a lineage diverging from the East that constitutes the Kenai population and contributes a large portion of genetic ancestry to SEAK following admixture from western lineage populations (Figure 24d). This inferred earlier divergence between the Mexican population and the population ancestral to both the East and Alaskan populations (Figure 24d) suggests either that there may have been two waves of colonization of the eastern range, or

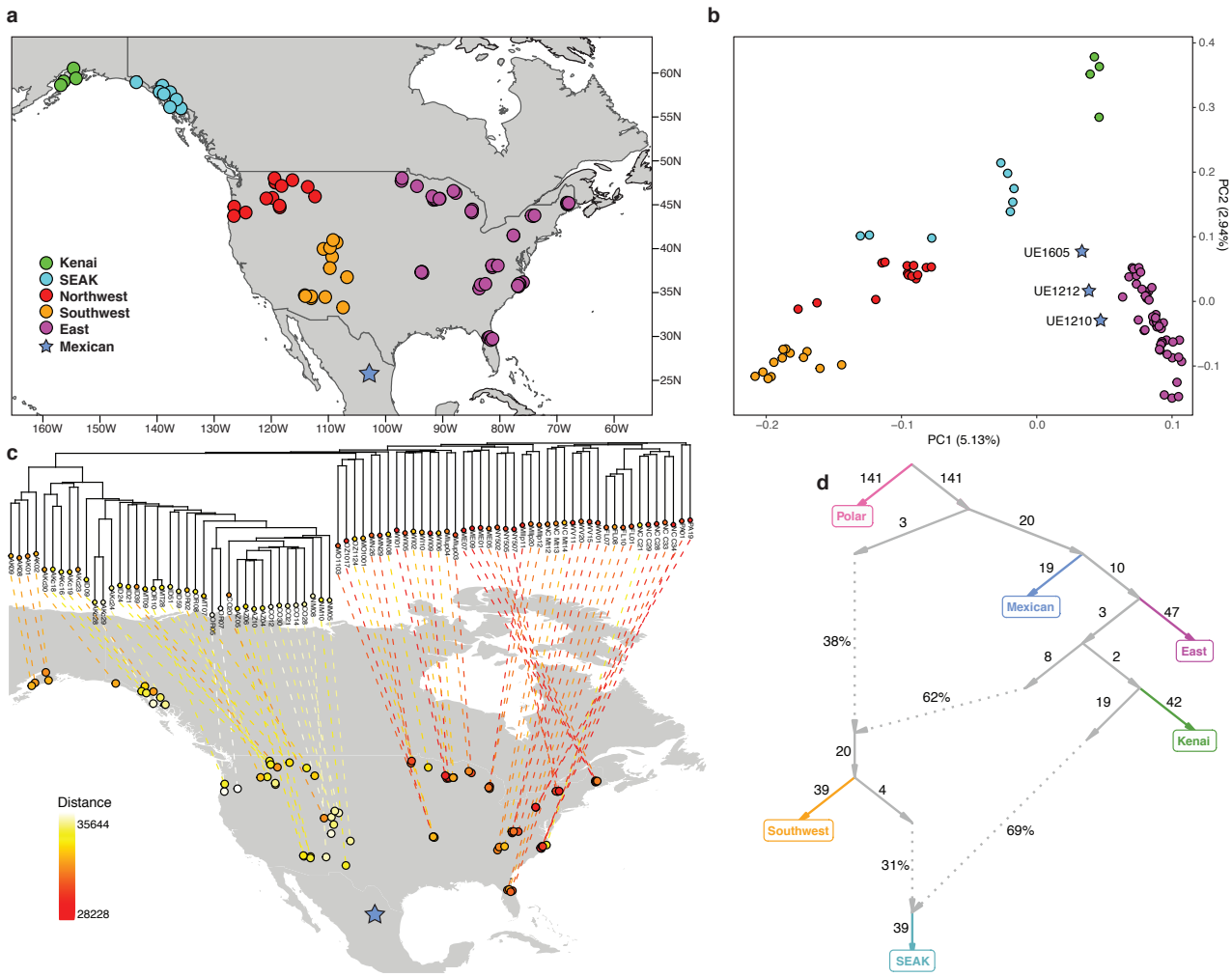


Figure 24: American black bear phylogeny. (a) Map showing the black bear samples used. (b) Principal component analysis using SmartPCA, which accounts for the high amount of missing data in the Mexican samples by projecting the ancient samples onto a PCA created from the modern samples. (c) Genetic Hamming distance of UE1212, to each of the modern samples on biallelic SNPs, scaled to account for missing data, mapped to a colour scale and plotted on a phylomap using a neighbour-joining tree of the modern samples (results for UE1210 and UE1605 show very similar patterns, see (Pedersen et al., 2021) Figure S3a,b). (d) Inferred admixture graph, using two polar bear genomes as an outgroup in our admixture analysis. All data were parsed and plotted using admixtools (Maier et al., 2022). I determined seven best-fitting graphs with highly similar topologies and many shared characteristics. The best of these is shown here, with a score of 4.922, and with a worst excess f4 residual of -2.182 for the configuration (East,Kenai;Mexican,Polar), and the rest are shown in Figure S4.

alternatively that the East and Alaskan populations are descendants of a northward range expansion from a southern population. A PCA (Figure 24b) shows a signature of range expansion in

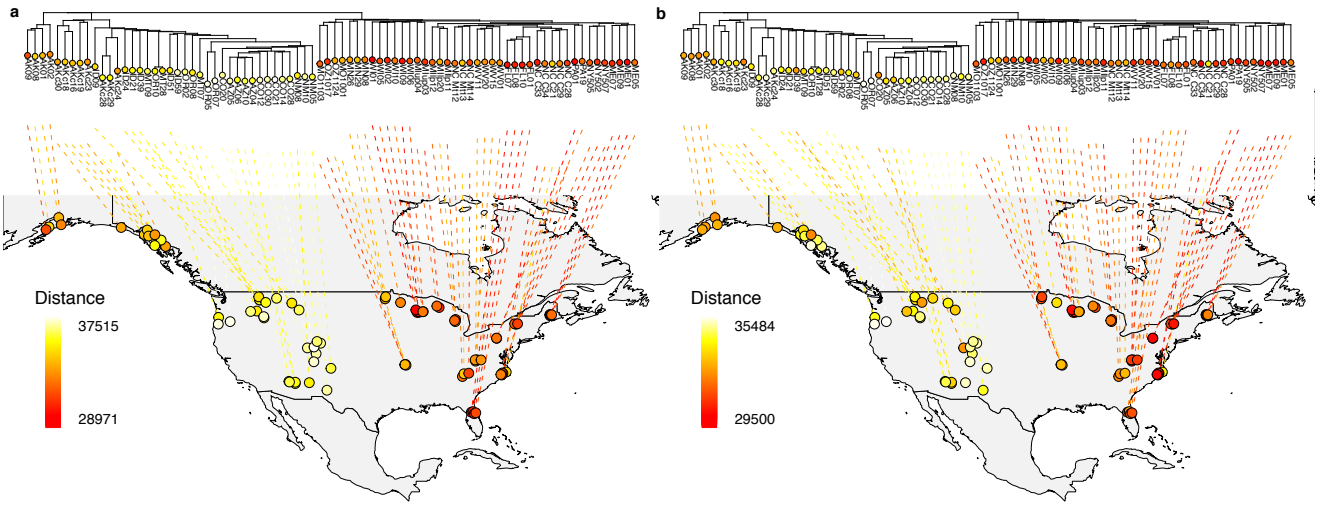


Figure 25: Genetic Hamming distance of ancient Mexican black bear samples UE1210 and UE1605 (shown in (a) and (b) respectively), mapped to a colour scale and plotted on a phylomap using a neighbour-joining tree of the modern samples.

the east, which may be explained in two, non-mutually exclusive, ways. First, range expansion into the eastern mountain ranges may have begun in the north and proceeded southward, resulting in isolation-by-distance or population structure. When glaciers advanced toward the peak of the ice age, northern bear populations contracted southward into the Southeast refugium (Figure 26a), where they maintained geographically structured populations rather than mixing with established bears. In this case, the leading edge of the northward expansion after the peak of the ice age would have comprised the descendents of the northern populations. Alternatively, northern populations may have been extirpated (or panmixia occurred) and the range expansion signal reflects expansion of the refugial population during post-glacial reforestation. The substructuring in the East may also be influenced by more recent processes; specifically, admixture from the Northwest into populations around the Great Lakes (Puckett et al., 2015) (Figure 26b) which has resulted in higher diversity (Puckett and Davis, 2021).

Contemporary Mexico has isolated bear populations in both the Sierra Madre Occidental and Sierra Madre Oriental mountain ranges, and is the only range state where black bears are considered endangered. Assuming population continuity in Mexico over the past 16 cal kyr BP, our results provide the first direct evidence linking eastern Mexican populations to the eastern lineage. Mitochondrial haplotype analyses identified clades A-west and A-east, respectively in the Occidental and Oriental ranges (Onorato et al., 2004, 2006; Varas-Nelson, 2011), yet mitochondrial-nuclear discordance has been observed between bear species and in black bear populations. Combined, the

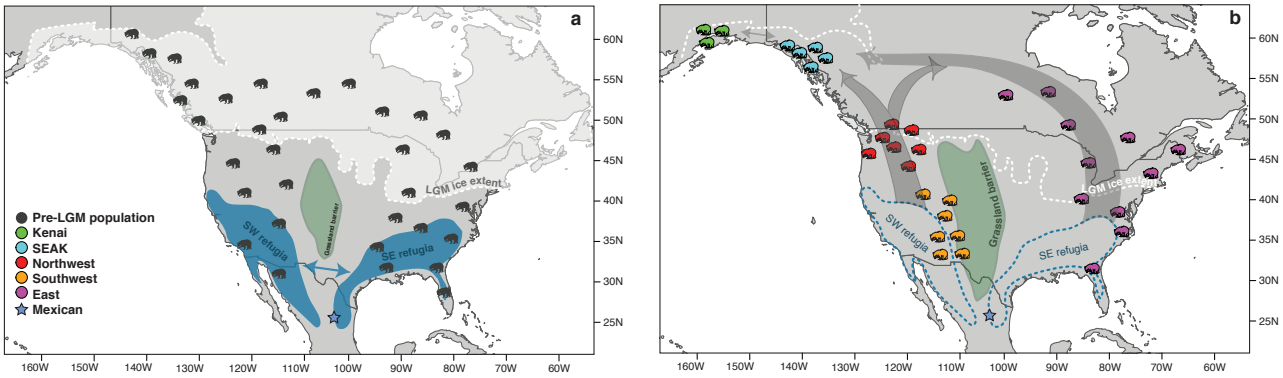


Figure 26: Working model of American black bear phylogeography. a. Pre-LGM – LGM conditions, with the ice-sheet extending at 18kya (uncalibrated), and the hypothesized refugia to which the pre-LGM black bear population was suppressed. b. Post-LGM conditions, with grey arrows indicating the northward recolonization of ice-free areas.

data suggest two colonizations of the Mexican mountain ranges by black bears, and that the Chihuahuan Desert may have been a barrier to east-west gene flow. Here it is shown that the ancient Mexican population diverged before the East and Alaskan populations split; thus, given previous population divergence times from nuclear genomic data, one can infer the Mexican population diverged 67-31 cal kyr BP (Puckett et al., 2015).

4.3.2 Giant short-faced bear

Exploratory analyses revealed that stratum UE1605 contained what appeared to be a mixture of DNA from two bear species (Figure 22). When mapping reads recovered from this layer, some reads better aligned to the mitogenome of the giant short-faced bear (Arctodus simus) than to the reference genome of the black bear, with both showing an equally edit distance and high amount of DNA damage (Figure 22). To contextualize the giant short-faced bear aeDNA in UE1605, ancient DNA was used from three Late Pleistocene short-faced bear fossils from Yukon, Canada (YG 24.1 (22.3 cal kyr BP), YG 76.4 (28.9 cal kyr BP), YG 546.562 (29.8 cal kyr BP). Complete mitochondrial genomes were assembled and nuclear genomic data sets from all three, including a 26-fold coverage nuclear genome for YG 564.562 (see Methods in (Pedersen et al., 2021)).

Mitochondrial DNA analyses confirmed that the additional bear represented in UE1605 was a giant short-faced bear. A mitochondrial phylogeny was estimated using whole mitogenomes of the eight extant bears of the family Ursidae as well as three extinct bear lineages: cave bears (U. spelaeus), and the two extinct tremarctine bears, the North American giant short-faced bear, *Arctodus*, and the South American giant short faced bear, Arctotherium, which I reassembled using

the Andean bear as reference (Figure 27a; also see methods in (Pedersen et al., 2021) for details). To assign reads from UE1605 to this phylogeny, I implemented a competitive mapping approach in which I simultaneously mapped each read to both black bear and Andean bear mitochondrial genomes and partitioned them into read sets that: (1) mapped uniquely to black bear; (2) mapped uniquely to Tremarctos; or (3) mapped to both. I then used pathPhynder (Martiniano et al., 2022) to assign biallelic transversion SNPs onto the mitochondrial tree and to determine which SNPs in each read set either supported or conflicted with each branch of the phylogeny (Figure 27a). Apart from a single SNP from a read that mapped uniquely to the black bear and supports the Andean bear clade, which I assume is due to noise, this analysis supports two distinct paths on the mitochondrial phylogeny, one leading to the giant short-faced bear and the other to American black bear, confirming that the competitive mapping approach can separate two related species co-recovered from an eDNA sample. Note that only 18 biallelic transversion SNPs assigned to branches mapped to both black bear and Andean bear mitochondrial genomes, despite their being species that diverged only 13.4 million years ago (Mya) (Figure 27c).

Although the mitochondrial data from UE1605 were too sparse to infer the evolutionary relationships between the Mexican and Yukon giant short-faced bear populations (only 197 reads mapped uniquely to the Andean bear mitochondrion), the nuclear data suggest that the two populations were genetically distinct. After filtering the UE1605 reads to obtain only those that mapped uniquely to the Andean bear reference genome, I computed a pairwise genetic Hamming distance matrix on biallelic transversion SNPs on this filtered UE1605 read set and reads from three Yukon short-faced bears, rescaling to account for missing data in the eDNA sample. These distances are shown on the off-diagonal of Figure 27b, along with a hierarchical clustering tree on the distance matrix. Numbers on the diagonal represent individual heterozygosities for the three fossil samples, calculated on the same sites. The heterozygosity for YG 546.562 is somewhat higher than the other two samples, probably because of its greater sequencing depth. In terms of relative genetic distance, the Mexican UE1605 appears to be deeply divergent from the Yukon fossils, approximately 3.5-4.4x more diverged than the Yukon fossils are to each other.

Using the relative distances in Figure 27b and the high coverage Yukon giant short-faced bear genome, I estimate that the Yukon and Mexican short-faced bear populations diverged roughly 200-150 thousand years ago (kya) as follows. A PSMC plot for YG 546.562 (Figure 27d; Methods) suggested an average effective population size for the Yukon population of around 5000 individuals over the last 100 kya. Assuming a generation time of 6 years (Kumar et al., 2017), I estimate a mean coalescence time within the Yukon giant short-faced bear population of 60,000 years, and thus can approximate a divergence time for the Mexican and Yukon populations around 200-150 kya

(assuming a similar population size and generation time within the common ancestor). Given that the most recent time points in the PSMC model (typically younger than 10 kya) are less accurate (Li and Durbin, 2011), demographic estimates younger than 40 kya in Figure 27d have been ignored.

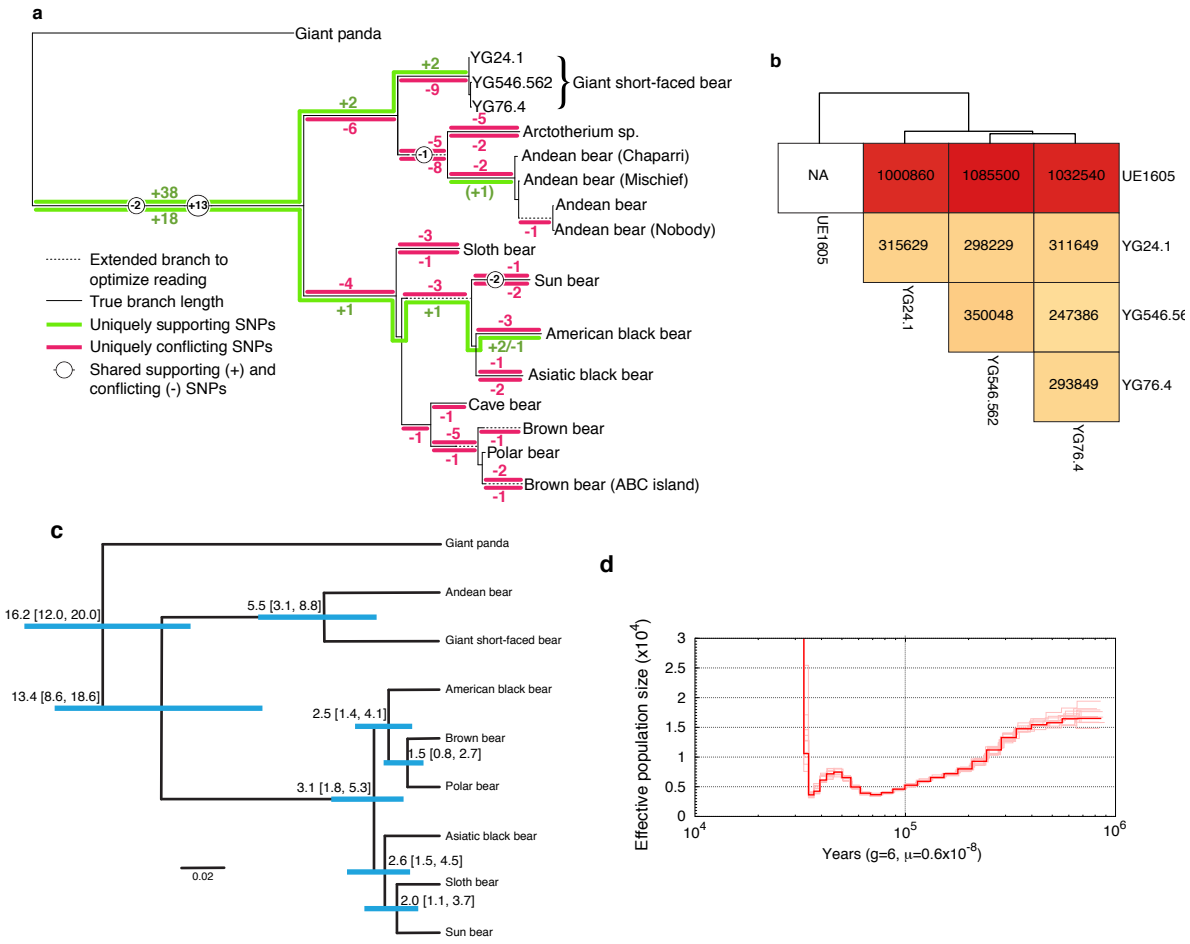


Figure 27: Giant short-faced bear genomic and population estimates. (a) Biallelic transversion SNPs in UE1605, partitioned by read mapping (uniquely to the black bear mitochondrion, uniquely to Andean bear, or shared), and placed onto a mitochondrial Ursid tree. Lines above the black backbone lines of the tree indicate SNPs mapping uniquely to Andean bear, lines below the tree indicate mapping uniquely to black bear. The (+1) indicates a single supporting SNP in the black bear mapping leading to the Andean bear clade. (b) Nuclear genome distances and heterozygosities on all giant short-faced bear specimens and the UE1605 reads which mapped uniquely to the Andean bear genome but not the black bear genome, calculated on biallelic transversion SNPs. Numbers on the diagonal refer to individual heterozygosities, whereas the off-diagonal shows pairwise genetic Hamming distances between samples. I performed hierarchical clustering on the pairwise genetic distance matrix to obtain the tree. (c) Phylogenetic tree and divergence times of the eight extant bear species and the extinct giant short-faced bear, as inferred from analysis of nuclear genomes. Branch lengths represent time before present (million years ago (Mya)). The mean age of each node is shown, with 95% credibility intervals in parentheses and depicted as blue bars around each node. (d) PSMC plot for YG 546.562.

4.4 Conclusion

Here I have presented one of the first eDNA studies to show that it is possible to separate genomic-wide sequences from closely related species that are present in the same environmental samples, and then to make population genetic inferences on at least one of these species, as long as reference data exist for the taxa in question. I further showcased how an "environmental genome" can be used in population genomic and phylogenetic studies. This opens the possibility of analyzing DNA from environmental samples in a similar manner as is currently done for DNA from fossil remains. As fossils are valuable, DNA analyses are destructive, and most species and populations of the past are poorly represented in, or even absent from, fossil records, the analysis of ancient environmental genomes directly from eDNA will allow improved insights compared to what can be addressed by DNA from fossils alone.

It is not generally possible to determine how many individuals contributed to these environmental genomes. As a pseudo-haploid sequence was used for these analyses, which was obtained by selecting a random read at each position, these results concern the population from which these reads came. This is true of all low-coverage genomes: because even a single individual is diploid, a pseudo-haploid genome from a fossil samples from two genomes from the population. All analyses used are robust to operating on a random sample of alleles from the population. It should also be noted that if the sample arose from multiple individuals from different populations, for example due to gene flow and/or replacement, then the analyses would report results as for an admixed population. With sufficiently deep coverage it might in principle be possible to use linkage disequilibrium to distinguish a mixture of individuals from recent genetic admixture. In this case, neither the black bear nor Arctodus results suggested admixture.

5 Molecular dating of a *Betula* chloroplast aeDNA sequence from Northern Greenland

This chapter has been published as part of: Kjaer, K.H, Pedersen, M.W, De Sanctis, B., De Cahsan, B., Korneliussen, T.S., Michelsen, C.S., Sand, K.K., Jelavić1, S., Ruter, A.H., Bonde, A.M.Z, Kjeldsen, K.K., Tesakov, A.S, Snowball, I., Gosse, J.C., Alsos, I.G., Wang, Y., Dockter, C., Rasmussen, M., Jørgensen, M.E., Skadhauge, B., Prohaska, A., Kristensen, J.A., Bjerager, M., Allentoft, M.E., Coissac, E., PhyloNorway Consortium, Rouillard, A., Simakova, A., Fernandez-Guerra, A., Bowler, C., Macias-Fauria, M., Vinner, L., Welch, J.J., Hidy, A.J., Sikora, M., Collins, M.J., Durbin, R., Larsen, N.K. and Willerslev, E. A 2-Million-year-old ecosystem in Greenland uncovered by Environmental DNA. *Nature* 612, 283–291 (2022).

Until this point, the oldest DNA sequenced was from a mammoth fossil specimen (van der Valk et al., 2021). Here, I report the oldest ancient DNA record to date, from an environmental sample describing the rich plant and animal assemblages of the Kap København Formation in North Greenland. Various geological dating techniques have narrowed down the date of this site to 1.9-2.1 million years ago. At this point in history, the climate in the region resembled that forecasted under future warming, and paleoclimactic records show mean annual temperatures of 11-19°C above contemporary values. The eDNA record shows an open boreal forest ecosystem with mixed vegetation of poplar, birch, and thuja trees as well as a variety of Arctic and boreal shrubs and herbs, many of which had not previously been detected at the site from macrofossil and pollen records. In this chapter, I use the ancient environmental reads assigned to the birch tree chloroplast genome to molecularly date the sample in order to confirm the geological dates. I do this in two different ways: first, using pathPhynder phylogenetic placement and an explicit SNP-counting/branch-shortening approach, then by using BEAST (Suchard et al., 2018). Though results differ somewhat between the two methods, both successfully yield 95% high posterior density (HPD) intervals which overlap with the geological dates.

5.1 Introduction

The competition to reliably sequence the oldest DNA has been going since the first ancient DNA was sequenced. As discussed in Chapter 1, much of the 1990s was spent without proper anti-contamination protocols, and without the use of next-generation sequencing, false claims were made of ancient DNA that was many millions of years old. Nowadays, the oldest confirmed DNA until this study was a $\sim$ 1.3 million year old mammoth fossil, sequenced just last year (van der Valk et al., 2021). Before that, a $\sim$700,000 year old horse bone held the record for almost 8 years Orlando et al. (2013b). Just a few months ago, (Armbrecht et al., 2022) provided an ancient environmental DNA diatom record in Antarctica which may be up to 1 million years old.

The current study relies on a new dataset of ancient environmental DNA from a formation in Peary Land, Northern Greenland (82°24' N 22°12' W). Though the area is now a polar desert as can be seen in Figure 28(a), around 2 million years ago the mean annual temperature was 11-19°C above modern values (Brigham-Grette et al., 2013), which is similar to what might occur in the future due to the climate crisis (Chylek et al., 2022). The formation has been the subject of multiple previous studies which have shown the existence of a coniferous boreal forest. These studies have dated the deposit to approximately 2.4 million years ago. Though traces of insects have been found and there must have been animals inhabiting the region, the only direct evidence from fossils is of the family *Lagomorpha* (hares, rabbits and pikas). In the Arctic in general, fossils are relatively rare and the faunal communities in the past are not well understood (Matthews et al., 2019).

An ancient environmental DNA record therefore provides deep, novel insight into this ecosystem, uncovering which animals and plants lived there in the past and hinting at what the area could look like in the future. The current study uncovered the presence of mastodons, hare, reindeer, rodents, geese, horseshoe crab, willow, mountain avens, birch, poplar, sedges, horsetails and more in sediment cores collected from Kap Kopenhagen (see Figure 28(b)). An illustrated reconstruction of this ecosystem is shown in Figure 29. Many of these taxa are not present in the Arctic today because of their need for higher temperatures or different conditions. Indeed, there is no modern analogue of this entire ecosystem. The current study used geological methods to date the site to either $\sim$1.9 or $\sim$2.1 million years ago. DNA was extracted from five different sites in the deposit, which were combined for the purposes of this chapter. Reads were assigned to individual taxa using methods outlined briefly in the Methods below, and in more detail in the full manuscript.

To validate the geological date, I aimed to perform molecular dating on one or more of the read sets assigned to individual taxa. Molecular dating has been used numerous times on ancient DNA extracted from bone, for example in van der Valk et al. (2021), but never to my knowledge from ancient environmental DNA, which presents a number of additional difficulties. Environmental

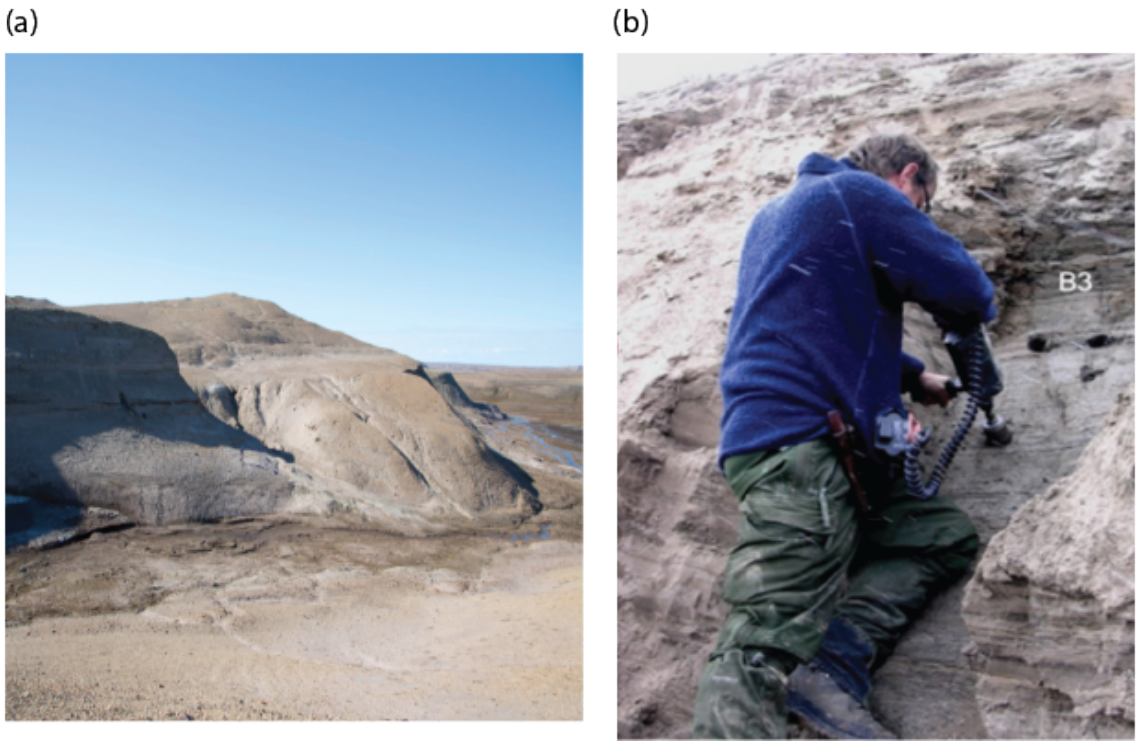


Figure 28: (a) A picture of the Kap Kopenhagen site now. (b) A collaborator, Nicolaj Larsen, collecting sediment cores from the Kap Kopenhagen site.

DNA samples will tend to contain a mixture of individuals from each species or genus, potentially from genetically different populations. This can confound SNP calling at variable positions. In order to determine if samples here represent enough of a mixture as to be problematic for molecular dating, I used the pathPhynder phylogenetic placement method first. This can reveal when read sets contain a mixtures of species, at least, because the sample will contain supporting SNPs on multiple different paths. Furthermore, eDNA read sets will be biased in the regions of the genome that are covered because of the competitive mapping algorithm required to assign reads to individual taxa. In particular, more conserved regions of the genome will have fewer reads assigned to them because those reads will also map to similar taxa with higher rates. In this chapter, I consider the three plant taxa from the Kap Kopenhagen aeDNA dataset with the most chloroplast reads assigned, and with a sufficient number of reference chloroplast genomes to theoretically perform molecular dating: *Salix* or willow tree, *Betula* or birch tree and *Populus* or poplar tree.

Figure 29: A reconstruction of the Kap Kopenhagen site 2 million years ago according to the taxa revealed by ancient environmental DNA. Artist: Beth Zaiken.

5.2 Methods

5.2.1 Extracting and mapping reads

First I sought to phylogenetically place the set of ancient plant taxa with the most abundant number of chloroplast reads assigned, and with a sufficient number of reference sequences to build a phylogeny. Although the evolution of the chloroplast genome is somewhat less stable than that of the plant mitochondrial genome, it has a faster rate of evolution and hence is more likely to contain more informative sites for our analysis than the plant mitochondria (Huang et al., 2014). Like the mitochondrial genome, the chloroplast genome also has a high copy number, so that we would expect a high number of sedimentary reads mapping to it, and it is non-recombining.

Raw sequence data for the ancient environmental samples is available through the ENA project accession PRJEB55522. The raw reads were run through standard filtering and quality control steps, then mapped against a large reference database using ngsLCA by my collaborators. Of the taxa with the most abundant number of reads assigned in the reference database, the three genera which also had sufficient chloroplast reference sequences were *Salix* or willow tree, *Populus* or poplar tree, and *Betula* or birch tree. From the damage-filtered ngsLCA output (Wang et al., 2022),

I extracted all read IDs uniquely classified to reference sequences within *Salix*, *Populus*, and *Betula* or assigned to any common ancestor inside these taxonomic groups with a minimum sequence similarity of 90% or higher, and converted these back to fastq files using seqtk (Li, 2018). For the purpose of molecular dating on the scale we are considering, it is appropriate to consider these read sets as a single sample, and so I merged the resulting bam files from all sites and layers to create a single read set for each taxon using samtools (Li et al., 2009a).

Next I needed to build reference databases and phylogenies. For each of *Salix*, *Betula* and *Populus*, I downloaded a representative set of whole chloroplast genome fasta sequences from NCBI's Genbank (Howe et al., 2020), including a single representative sequence from a recently diverged outgroup. For the *Betula* genus, I also included three chloroplast genomes from the PhyloNorway database (Wang et al., 2021). A list of the species and accession IDs is given in Table 3. Since chloroplast sequences are circular, downloaded sequences may not always be in the same orientation or at the same starting point as is necessary for alignment, so I used custom code (https://github.com/miwipe/KapCopenhagen) that uses an anchor string to rotate the reference sequences to the same orientation and start them all from the same point. I changed all ambiguous (non-ACGT) bases in the fasta files to Ns. I used MAFFT (Rozewicki et al., 2019) to align each of these sets of reference sequences, and inspected multiple sequence alignments in NCBI's MSAViewer to confirm quality (Howe et al., 2020). The BEAST suite (Suchard et al., 2018) was used with default parameters to create ultrametric phylogenetic trees for each of the three taxa from the multiple sequence alignments (MSAs) of reference sequences, which were converted from Nexus to Newick format in Figtree (Rambaut, 2010). I then passed the multiple sequence alignments to a custom script using the python module AlignIO from BioPython (Cock et al., 2009) to create a reference chloroplast consensus fasta sequence for each set of taxa. Next, I used SNPSites (Page et al., 2016) to create a vcf file from each of the MSAs. Since SNPSites outputs a slightly different format for missing data than needed for downstream analysis with pathPhynder, I wrote a custom R script to modify the vcf format appropriately. I also filtered out non-biallelic SNPs at this point.

Since the extracted ancient environmental reads were mapped against a reference database including multiple sequences from each taxon, the output files were not on the same coordinate system. To circumvent this issue and avoid mapping bias, we re-mapped each read set to the consensus sequence generated above for that taxon using bwa (Li and Durbin, 2009) with ancient DNA parameters (bwa aln -o 2 -n 0.001 -t 20). We converted these reads to bam files, removed unmapped reads, and filtered for mapping quality >25 using samtools (Li et al., 2009b). This yielded 103,042, 39,306, and 91,272 chloroplast reads for the *Salix*, *Populus*, and *Betula* respectively, with mean

depths 27x, 57x and 24x (although coverage is extremely uneven across the chloroplast genome; see Figure 30, for example). I used bcftools (Li et al., 2009a) to make an mpileup and call a vcf file, using options for haploidy and disabling the default calling algorithm, which can slightly bias the calls towards the reference sequence, in favour of a majority call on bases that passed the default base quality cutoff of 13. I included the default option which filters according to base alignment qualities, which I found greatly reduced the read depths of some bases and removed spurious SNPs around indel regions. Lastly, I filtered the vcf file to include only single nucleotide variants, because I do not believe other variants such as insertions or deletions in an ancient environmental sample of this type to be of sufficiently high confidence to include in molecular dating, and including these types of mutations in a molecular dating analysis is nontrivial regardless.

I performed molecular dating two days. First, I used phylogenetic placement and SNP-counting, which is more explicit but does not allow for variable mutation rates between sites or branches. Second, I used the BEAST software, and partitioned sites into sets which one might expect to have different rates. This more relaxed molecular clock is likely to be more correct, but I had to give BEAST only a subset of sites (in which I could confidently make calls for the ancient sample) for the entire analysis, whereas the pathPhynder SNP-counting approach did not have this limitation. These methods both have pros and cons, and gave somewhat different answers, as discussed below.

5.2.2 Phylogenetic placement

I next used pathPhynder (Martiniano et al., 2022), a phylogenetic placement algorithm that identifies informative markers on a phylogeny from a reference panel, evaluates SNPs in the ancient sample overlapping these markers, and traverses the tree to place the ancient sample according to its derived and ancestral SNPs on each branch. I investigated the pathPhynder output in each taxon set to determine the phylogenetic placement of the ancient samples.

In theory, all three genera *Betula*, *Populus* and *Salix* had both sufficiently high chloroplast genome coverage to attempt molecular dating on these samples. However, the *Populus* sample clearly contained a mixture of individuals from different species, as seen from its inconsistent placement in the pathPhynder output as can be seen in Figure 34. In particular, there were multiple supporting SNPs to both *Populus balsamifera* and *Populus trichocarpa*, and both supporting and conflicting SNPs on branches above. It therefore appears that there are multiple species of *Populus* contributing to this ancient eDNA sample. Because of this, I continued with only *Betula* and *Salix*.

5.2.3 Molecular dating with phylogenetic placement and SNP-counting

Point estimate. First, I wanted to use an explicit SNP-counting approach with a constant molecular clock for dating. This procedure is sometimes called branch shortening, where the missing amount of evolutionary change is proportional to the age of the ancient sample. In particular, this means using the phylogenetic placement and an estimated length for the private branch to estimate how far back into the tree the ancient sample lies, then converting this to years by calibrating using a known outgroup divergence time. The logic for this for *Betula* is shown in Figure 32b.

To estimate the private branch length, I needed to count private SNPs. For each of these two ancient samples *Betula* and *salix*, continuing from the phylogenetic placement above, I first counted the number of private SNPs in each sample with a custom R script. To do this, I called a private SNP if the majority of the reads at the position agreed and did not match the reference allele, and if that site was monomorphic for the reference allele in the reference panel. Of course, at low coverage, a site which appears to have a private SNP may in reality stem from deamination or other errors such as from the mapping pipeline. To avoid these false positives for private SNPs, I decided to implement a depth cutoff, count the number of private SNPs which had depth above this cutoff, and extrapolate to the rest of the chloroplast genome. This is shown in Figure 31.

For *Salix*, I found an extremely high number of private SNPs at any reasonable depth cutoff (eg. 147 private SNPs at a depth cutoff of 20), which is highly inconsistent with its age, especially considering that the number of SNPs assigned to the edges of the phylogenetic tree leading to other *Salix* sequences is much lower. I am unsure what causes this inconsistency, but hypothesize that the sample contains multiple *Salix* species which diverged from the same placement branch on the phylogenetic tree at different time periods. This is supported by looking at all of the reads that cover these private SNP sites, which generally appear to be from a mixed sample, with reads containing both alternate and reference alleles present at a high proportion in almost all cases. This could also partially result from the high number of nuclear plastid DNA sequences (NUPTs) (a sequence transposed from the chloroplast into the nuclear DNA) in *Salix*, which could cause nuclear DNA sequences to map to the chloroplast, potentially adding extra variation (Huang et al., 2017). Because of this uncertainty, I only continued the molecular dating analysis for *Betula* from this point on.

For *Betula*, analyzing Figure 31 combined with the high depths of SNPs overlapping other markers on the phylogenetic tree in pathPhynder led to a decision to use a depth cutoff of ≥ 20. This left 8 private SNPs, with a mean depth of 41, of which 4 were transitions and 4 were transversions. Since only 30.99% of the *Betula* chloroplast genome was covered at a depth of >=20, we extrapolated to the rest of the chloroplast genome to estimate 8/.3099 = 25.81 total private SNPs for our ancient *Betula* sample. This is given by (2) in Figure 32b.

The ancient *Betula* sample was placed on edge 3 in the phylogenetic tree, leading from node 34 to 35. Next I had to determine the average number of SNPs between node 35 and the sample placement, ie. leading inwards to the sample placement on edge 3 (that is, (1) in Figure 32b). To do this, I extracted from the pathPhynder output (see Phylogenetic Placement section) the number of SNPs assigned to each edge of the phylogenetic tree. Edge 3, on which the ancient sample was placed, had 109 total SNPs, and as can be seen in Figure 36, the ancient sample had 29 of these supporting placement on this branch, and 13 in conflict (as shown in Figure 32b). Since the fraction (29+13)/109 = 0.385 is consistent with the fraction of the chloroplast genome called, or 0.3099 as stated above, we used the latter number for our analysis. Therefore, we estimated that our sample was 13/.3099 = 41.95 SNPs into edge 3, out of the 109 total assigned.

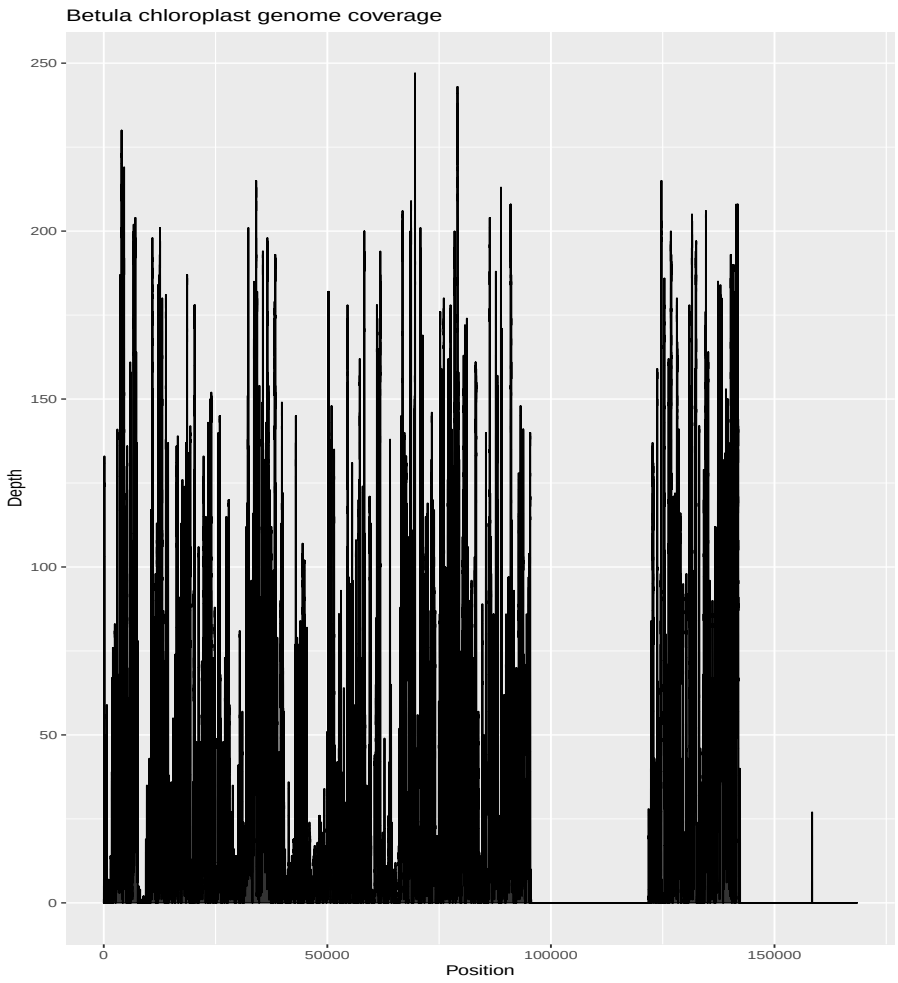


Figure 30: Coverage along the *Betula* chloroplast genome has a mean of 24x, but is very uneven across the genome, partly due to the mapping procedure. Regions which are more unique to *Betula* as a genus will generally have higher coverage, as ancient reads can be more confidently assigned to *Betula* on these regions. The two regions with no coverage are the inverted repeats in the chloroplast.

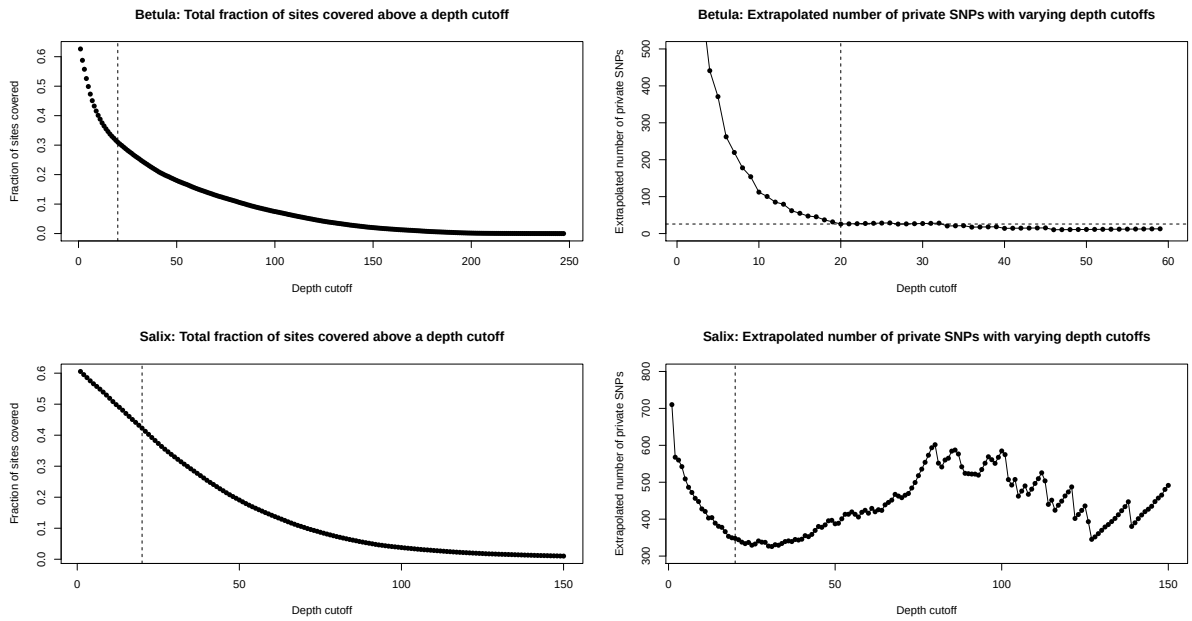


Figure 31: Top left: Number of ancient *Betula* chloroplast sites, mapped against the modern chloroplast consensus reference sequence, covered above different depth cutoffs. Top right: Extrapolated number of ancient *Betula* chloroplast private SNPs, compared to a reference panel of modern *Betula* samples, covered above different depth cutoffs. The extrapolated number of private SNPs was obtained by counting the actual number of private SNPs with a given depth cutoff, and dividing by the fraction of the genome covered at that depth. Since we would like the number of extrapolated private SNPs to be more or less stable above our chosen depth cutoff, we choose a depth cutoff of 20 (vertical dotted line), leading to 25.81 extrapolated private SNPs (horizontal dotted line). Bottom left: Number of ancient *Salix* chloroplast sites, mapped against the modern chloroplast consensus reference sequence, covered above different depth cutoffs. Bottom right: Extrapolated number of ancient *Salix* chloroplast private SNPs, compared to a reference panel of modern *Salix* samples, covered above different depth cutoffs. The extrapolated number of private SNPs was obtained by counting the actual number of private SNPs with a given depth cutoff, and dividing by the fraction of the genome covered at that depth.

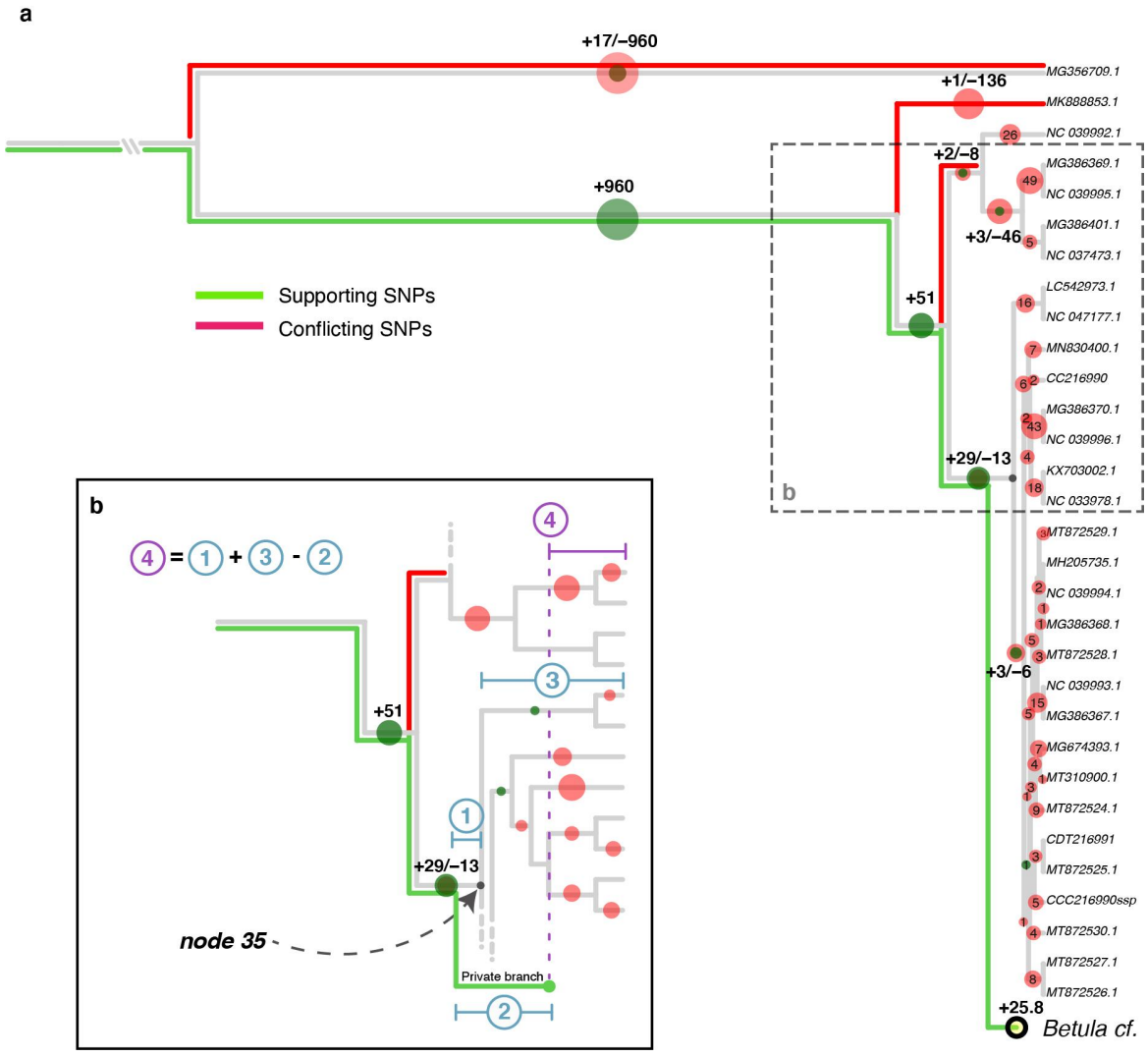


Figure 32: (a) An illustration of the *Betula* phylogenetic placement into its reference chloroplast phylogeny, along with its private branch length. (b) Zoomed into the square in (a); An illustration of how the molecular dating was done using phylogenetic placement and SNP-counting, further explained in the text.

Lastly, I needed the average number of SNPs both below node 35, that is, (3) in Figure 32b, and below the root node, or node 32. To do this, I again used the number of SNPs assigned to each edge of the phylogenetic tree, again extracted from the pathPhynder output. Since pathPhynder cannot discriminate between SNPs assigned to each of the two edges attached to the root node, I manually balanced these edges, assigning a proportion of the total number of assigned SNPs to these two edges according to their branch length. To count the average number of SNPs below a node, I used a phylogenetically aware distance measure by averaging the number of SNPs over each pair of edges, adding that average to the parent edge, and repeating, working from the tips to the node in consideration. This gave an average of 33.68 SNPs below node 32, and of 1385.15 below the root node.

Combining these, I estimate that our ancient Betula sample lies 33.68 + 41.95 - 25.81 = 49.81 SNPs inwards to its chloroplast phylogenetic tree (estimating (4) in Figure 32b), or a fraction of 49.81 / 1385.15 = 0.0360 into the phylogenetic tree, from the tips to the root. Next, I needed to convert this estimate to years. I used an *Alnus rubra* whole chloroplast sequence as an outgroup to the phylogenetic tree for this purpose. A recent paper (Yang et al., 2019) dated the Alnus-Betula chloroplast divergence at 61.1Mya (HPD 58.7-64.3Mya), and so using this, the point estimate for the ancient Betula sample converts to to 0.0360 * 61.1e6 = 2197265, or approximately 2.2 million years old.

Parametric bootstrap. Next, I wanted to obtain both a confidence interval and a variance estimate for this point estimate. First, I describe the parametric bootstrap model used to obtain a confidence interval. Generally, I model the number of SNPs on each branch of the phylogenetic tree as Poisson random variables, and introduce additional Binomial random variables to account for the uncertainty that arises from only using a fraction of the chloroplast genome with sufficient depth.

First, I sampled from a Poisson distribution with rate 25.81 (the point estimate for private branch length from above), then sampled from this according to a Binomial distribution with success probability 0.3099, the covered fraction of the genome. I divided the result by 0.3099, and used this as a re-estimate of the private branch length. Similarly, I sampled from a Poisson distribution with rate 41.95 (the point estimate for number of SNPs into edge 3 from above), then sampled from this according to a Binomial distribution with success probability 0.3099. I again divided this result by 0.3099, resulting in a re-estimate of the number of SNPs into edge 3. Continuing in the same way, I sampled the number of SNPs on every edge of the phylogenetic tree from Poisson distributions with rates according to the original number of SNPs on each edge, and re-calculated the average number of SNPs from the tips to node 35, and to the root node 32, using the phylogenetic distance

measure described above. Last, as done to obtain the point estimate, I added together the average number of SNPs under node 35 and the estimated number of SNPs into edge 3, subtracted the private branch length, and divided the total by the average number of SNPs from tip to root to obtain a re-estimate of the fraction of SNPs into the phylogenetic tree at which the ancient sample sits, including its private branch. I also output an estimate for the number of SNPs into the tree, rather than the fraction. Notably, this estimate will not account for the variation in the edges of the tree above the placement edge.

I iterated this parametric bootstrap procedure 1,000,000 times to create a distribution of the number and fraction of SNPs into the tree at which our ancient Betula sample lies, including its private branch, using an R script. The former distribution, of the number of SNPs, gives a 95% confidence interval of [20.3,80.0], computed using the quantile function in R. I multiplied the latter distribution of the fraction of SNPs by 61.1e6, the point estimate for the Alnus-Betula divergence. A 95% confidence interval for this distribution is [0.89, 3.5] Mya. Lastly, I wanted to include the uncertainty in the divergence date estimate, which is given as a 95% HPD [58.7,64.3] in (Yang et al., 2019). This interval is approximately 4% on either side around the mean, and incorporating this extra uncertainty into our confidence interval conservatively gives a range of [0.86, 3.7] Mya.

Variance. Using similar assumptions, I can also estimate the variance of the number of private SNPs, the number of SNPs into edge 3, and the average number of SNPs under node 35. Under the assumption of a Poisson distribution for the number of SNPs on each branch given its length, and that the called fraction 0.3099 is fixed, one can get $8/(.3099^2) = 83.29$ for the variance of the number of private SNPs and $13/(.3099^2) = 135.35$ for the variance of the number of SNPs into edge 3. I can also use the Poisson assumption to get the variance of the average number of SNPs under node 35 by writing out the equation for the algorithm given above to determine its expected value, and get 12.73 for this variance.

Altogether I get a variance of $135.35 + 12.73 + 83.29 = 231.38$, and therefore a standard deviation of 15.21, for the number of SNPs into the tree where our ancient Betula sample lies, including its private branch. This gives an approximate 95% confidence interval of 49.81 $\pm$ 1.96*15.21 = [20.0, 79.6]. This is reassuringly similar to the confidence interval calculated in the parametric bootstrap procedure (which was [20.3,80.0]). Indeed, we expect the confidence interval here to be exactly the same as the one calculated from the parametric bootstrap , since they both assume a Poisson distribution with the same mean.

To be absolutely sure there was no bias or error from deamination, I also performed this entire analysis with only transversions. As expected, this gave similar results with a slightly larger confidence interval (point estimate 2.7 Mya, 95% CI [0.98, 4.7] Ma).

5.2.4 Molecular dating with BEAST

In this section, I describe molecular dating of the ancient birch or *Betula* chloroplast genome using BEAST v1.10.4 (Suchard et al., 2018). For reasons outlined in the previous section, I did not attempt to molecularly date *Populus* or *Salix* using BEAST.

I first needed to partition sites into sets which might have different mutation rates. For example, the last nucleotide in a codon is known to mutate faster than the first (Bofkin and Goldman, 2006). To partition sites, then, I needed to annotate them. I downloaded the gff3 annotation file for the longest *Betula* reference sequence, MG386368.1, from NCBI. Using custom R code, I parsed this file and the associated fasta to label individual sites as protein-coding regions (in which I labelled the base with its position in the codon according to the phase and strand noted in the gff3 file), RNA, or neither coding nor RNA. We extracted the coding regions and checked in Seqotron (Fourment and Holmes, 2016) and R that they translated to a protein alignment well (e.g. no premature stop codons), both in the reference sequence and the associated positions in the ancient sequence. Though the modern reference sequence's coding regions translated to a high quality protein alignment, translating the associated positions in the ancient sequence with no depth cutoff leads to premature stop codons and an overall poor quality protein alignment. On the other hand, when using a depth cutoff of 20 and replacing sites in the ancient sequence which did not meet this filter by Ns, I saw a high quality protein alignment (except for the Ns). I also interrogated any positions in the ancient sequence which differed from the consensus, and found that any suspicious regions (e.g. with multiple SNPs clustered closely together spatially in the genome) were removed with a depth cutoff of 20. Because of this, I moved forward only with sites in both the ancient and modern samples which met a depth cutoff of at least 20 in the ancient sample, which consisted of just over 30% of the total sites.

Next, we parsed this annotation through the multiple sequence alignment to create partitions for BEAST (Suchard et al., 2018). After checking how many polymorphic and total sites were in each, I decided to use four partitions: (1) sites belonging to protein-coding positions 1 and 2, (2) coding position 3, (3) RNA, or (4) non-coding and non-RNA. To ensure that these were high confidence sites, each partition also only included those positions had fewer than three total missing sites in the multiple sequence alignment. This gave partitions which had 11,668, 5,828, 2,690 and 29,538 sites respectively. I used these four partitions to run BEAST, with unlinked substitution models for each partition and a strict clock, with a different relative rate for each partition. (There was insufficient information in these data to infer between-lineage rate variation from a single calibration). I assigned an age of 0 to all of the reference sequences, and used a normal distribution prior with mean 61.1 million years and standard deviation 1.633 million years for the root height

(Yang et al., 2019) (the standard deviation was obtained by converting the 95% HPD to z-scores). For the overall tree prior, we selected the coalescent model. The age of the ancient sequence was estimated following the overall procedures of Shapiro et al. (2010). To assess sensitivity to prior choice for this unknown date, I used two different priors, namely a Gamma distribution biased towards a younger age (shape=1, scale=1.7); and a uniform prior on the range [0,10MA]. I also compared two different models of rate variation among sites and substitution types within each partition, namely a GTR+G with four rate categories, and base frequencies estimated from the data, and the much simpler Jukes Cantor model, which assumed no variation between substitution types nor sites within each partition. All other priors were set at their defaults. Neither rate model nor prior choice had a qualitative effect on results (see Figure 33). I also ran the coding regions alone, since they translated correctly and are therefore highly reliable sites, and found that they gave the same median and a much larger confidence interval, as expected when using fewer sites (Figure 33). I ran each MCMC for a total of 100 million iterations. After removing a burn-in of the first 10%, I verified convergence in Tracer (Rambaut et al., 2018) (apparent stationarity of traces, and all parameters having an Effective Sample Size > 100). I also verified that the resulting MCC (maximum clade credibility) tree from TreeAnnotator (Suchard et al., 2018) had placed the ancient sequence phylogenetically identically to the pathPhynder placement, which is shown in Figure 36. For the major results, I report the uniform ancient age prior, and the GTR+G4 model applied to each of the four partitions. The 95% HPD was [0.68,2.02] Mya for the age of the ancient Betula chloroplast sequence, with a median estimate of 1.3 million years, as shown in Figure 33.

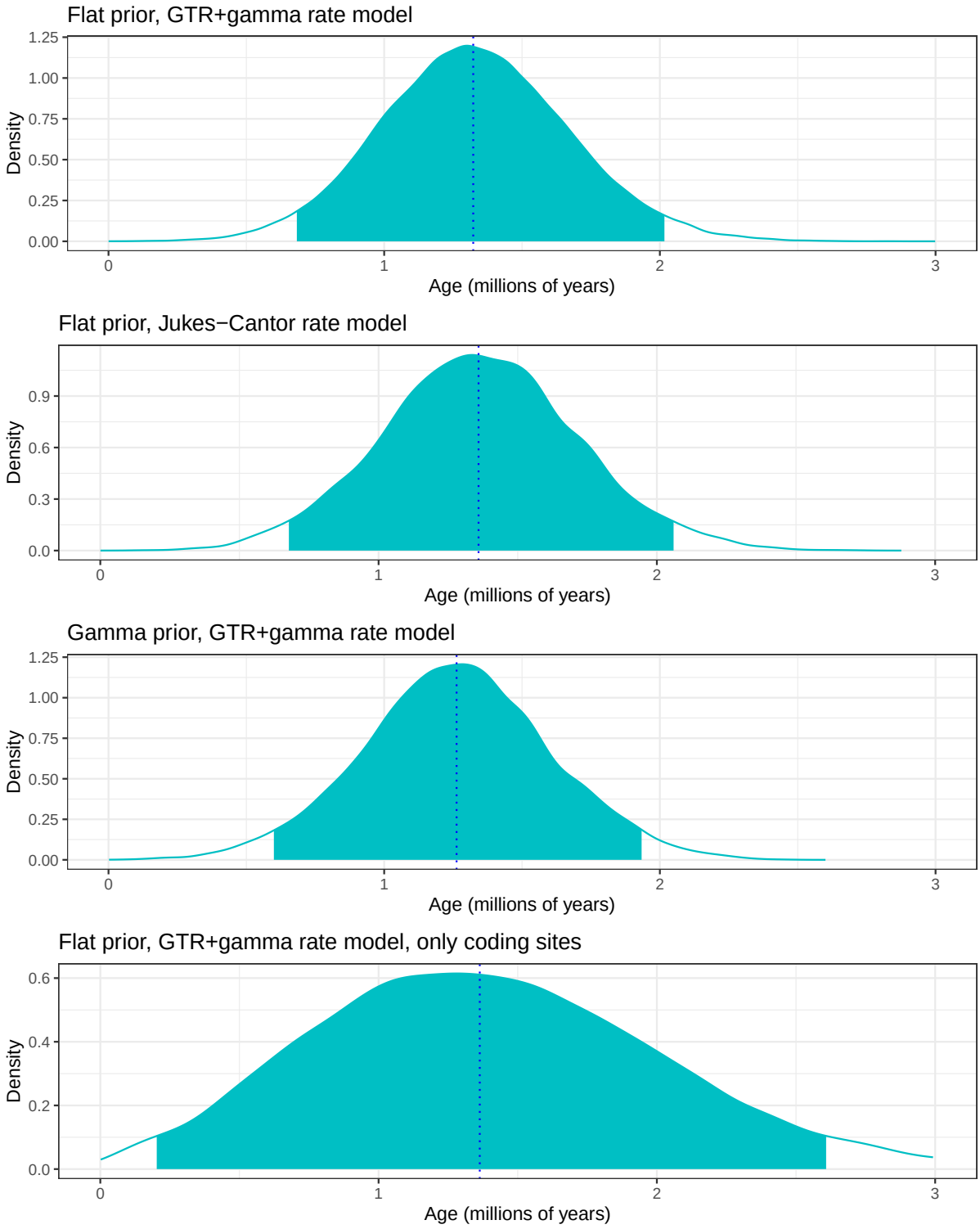


Figure 33: Molecular dating results for *Betula* using different prior distributions for site partitions in BEAST.

5.3 Results and discussion

5.3.1 Phylogenetic placements

Populus. As can be seen from the pathPhynder results in Figure 34, the ancient *Populus* sample likely contains a mixture of species, and so it is hard to make any conclusive statements without further analysis. The most likely placement is on the edge above *Populus trichocarpa* (NC 009143.1) and *Populus balsamifera* (NC 024735.1), with +71/-15 supporting and conflicting SNPs. However, we find some support for both of the branches directly leading to these species as well. *Populus balsamifera* and *trichocarpa* are considered sister species. They are both distributed in North America, as far North as Alaska, are known to hybridize both among themselves and other *Populus* species, and are morphologically very similar (Huang et al., 2014; Levsen et al., 2012; Huang et al., 2017). Previous analyses found a very recent nuclear genome divergence time of only 75,000 years ago for *Populus trichocarpa* and *balsamifera* (Levsen et al., 2012), but a deep chloroplast genome divergence time of at least 6-7Mya (Huang et al., 2014), which is an uncommon pattern in plants. The ancient *Populus* sample here could contain individuals either ancestral to, or hybridized from ancestors of the modern *Populus trichocarpa* and *balsamifera* species, or their relatives.

I also find 13 supporting (and 47 conflicting) SNPs leading to the related clade which contains *Populus trinervis* (MT482538.1), *Populus tremula* (MT482535.1), *Populus rotundifolia* (MT482542.1) and *Populus davidiana* (MT407464.1). This clade has an estimated root time of 12.46Mya (Zhou et al., 2021), which is considerably older than the age of our sample, raising the possibility that our ancient *Populus* sample also contains material from one or more extinct species basal to this clade which diverged a long time ago.

Salix. As can be seen in Figure 35, our ancient Salix sample is phylogenetically placed, with 356 supporting SNPs and 22 conflicting SNPs, on a basal branch leading to the main *Salix* clade. Though the *Salix* chloroplast phylogeny is not considered fully resolved, the difficulties in resolution lie underneath our placement branch (Huang et al., 2017), and this along with the high number of SNPs on the placement branch allow us to be reasonably confident in the placement branch, at least. The clade underneath the placement contains *Salix dasyclados* (MT551160.1), *Salix rorida* (MG262368.1, NC037428.1), *Salix minjiangensis* (NC037425.1), *Salix hypoleuca* (NC037423.1), *Salix argyracea* (MT551159.1), *Salix suchowensis* (MT551163.1), *Salix eriocephala* (MT551161.1), *Salix taoensis* (MG262369.1, NC037429.1), *Salix rehderiana* (NC037427.1, MG262367.1), *Salix integra* (MT551162.1), and *Salix magnifica* (NC 037424.1), whereas the other *Salix* clade contains *Salix chaenomeloides* (NC 037422.1) and *Salix paraplesia* (NC 037426.1). The chloroplast phylogeny inferred here agrees roughly with (Zhang et al., 2018), which estimated a divergence date

between these two main *Salix* clades at 16.9Mya, and a root age of the first clade, to which our ancient sample is basal, of 8.1Mya. It is reasonable, then, to conclude that the ancient *Salix* read set is at least 8.1Mya diverged from modern *Salix* species, and probably represents an extinct species, or a pool thereof. As noted in the Methods section, our ancient *Salix* sample also contains a surprisingly high number of private SNPs, which could indicate a mixed sample of individuals diverging at different points along this branch, high diversity in the ancient *Salix* population, selection on this ancient *Salix* population, or any combination of these.

Betula. Like *Salix*, the ancient *Betula* sample was placed basal to a main present day *Betula* clade, as can be seen in Figure 36. The placement was based on 29 supporting and 13 conflicting SNPs on its placement branch, and along with very low numbers of supporting SNPs elsewhere in the tree other than those leading to this branch, this indicates high confidence in the placement of our ancient *Betula* sample on this branch. The clade directly outside the placement edge contains *Betula cordifolia* (NC037473.1, MG386401.1), *Betula populifolia* (NC 039995.1, MG386369.1) and *Betula lenta* (NC 039992.1), the outgroup within all of the *Betula* samples contains *Betula alnoides* (MG386401.1), and every other *Betula* sample (*Betula nana, costata, chibuensis, microphylla, platphylla, pubescens, halophila, occidentalis*, and *fruticosa*) lies in a clade underneath our placement edge. The *Betula cordifolia-nana* divergence time is estimated at 6.63Mya in (Yang et al., 2019), which corresponds to the node at the top of the placement edge, so one can conclude that the ancient *Betula* sample diverged more recently than 6.63Mya.

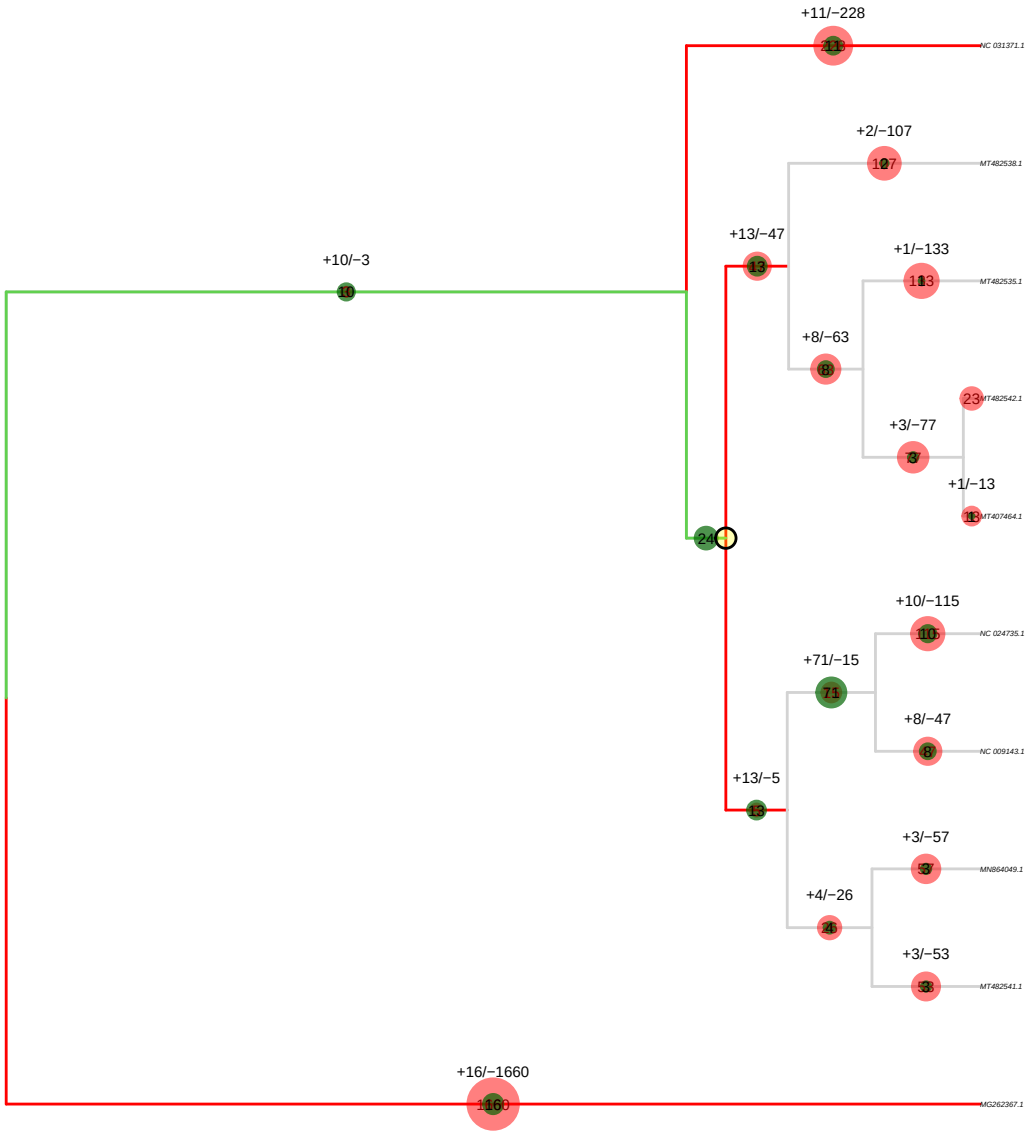


Figure 34: Phylogenetic placement results for the *Populus* chloroplast reads, using both transition and transversion SNPs, and using reads merged from all layers and sites. The numbers on each edge represent the number of supporting (+) and conflicting (-) SNPs in the ancient *Populus* environmental genome overlapping the reference SNPs assigned to that edge. The ancient *Populus* environmental genome clearly contains a mixture of different species.

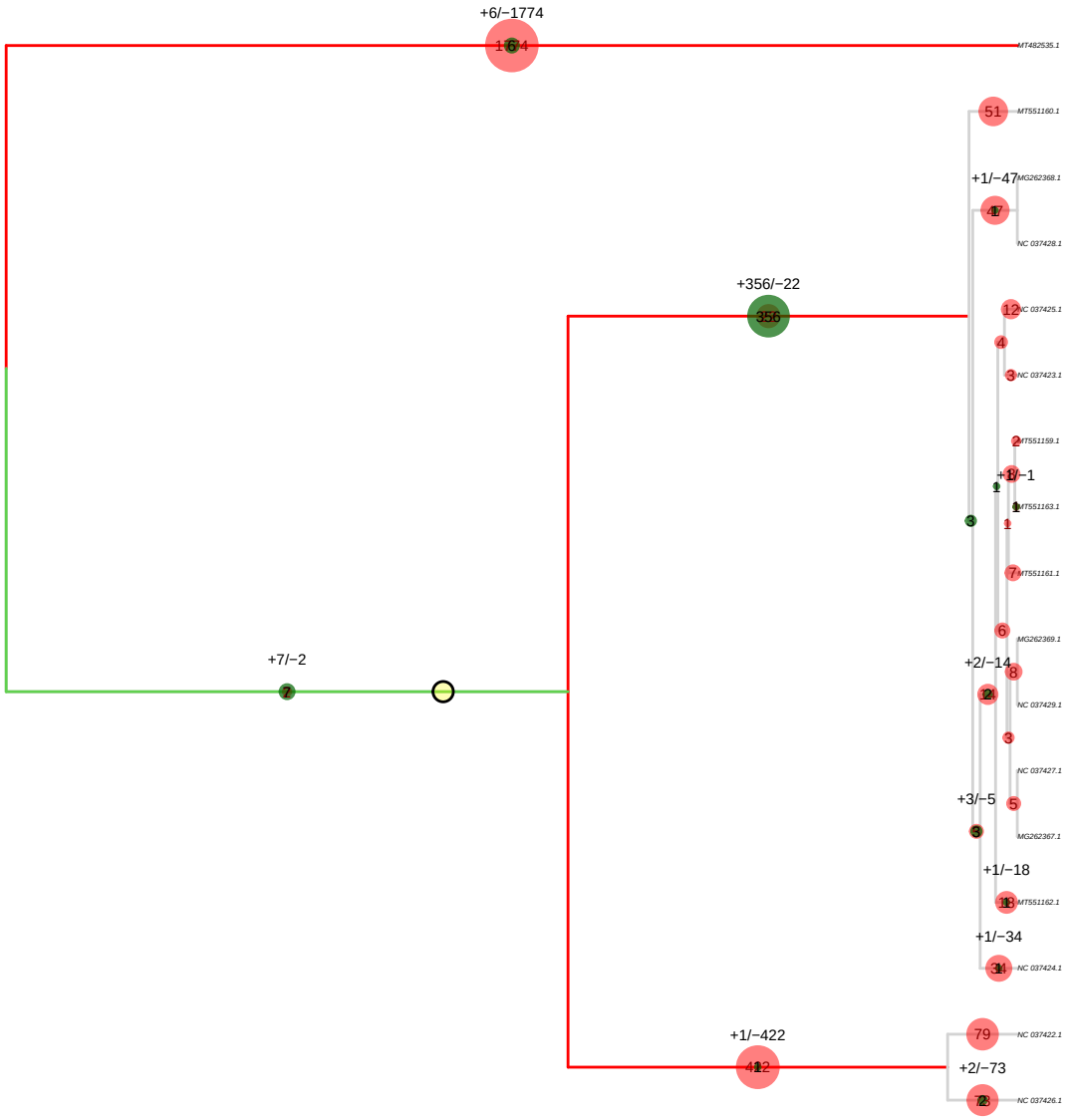


Figure 35: Phylogenetic placement results for the *Salix* chloroplast reads, using both transition and transversion SNPs, and using reads merged from all layers and sites. The numbers on each edge represent the number of supporting (+) and conflicting (-) SNPs in the ancient *Salix* environmental genome overlapping the reference SNPs assigned to that edge. The ancient *Salix* environmental genome falls basal to a main *Salix* clade, and therefore likely diverged more than 8.1Mya.

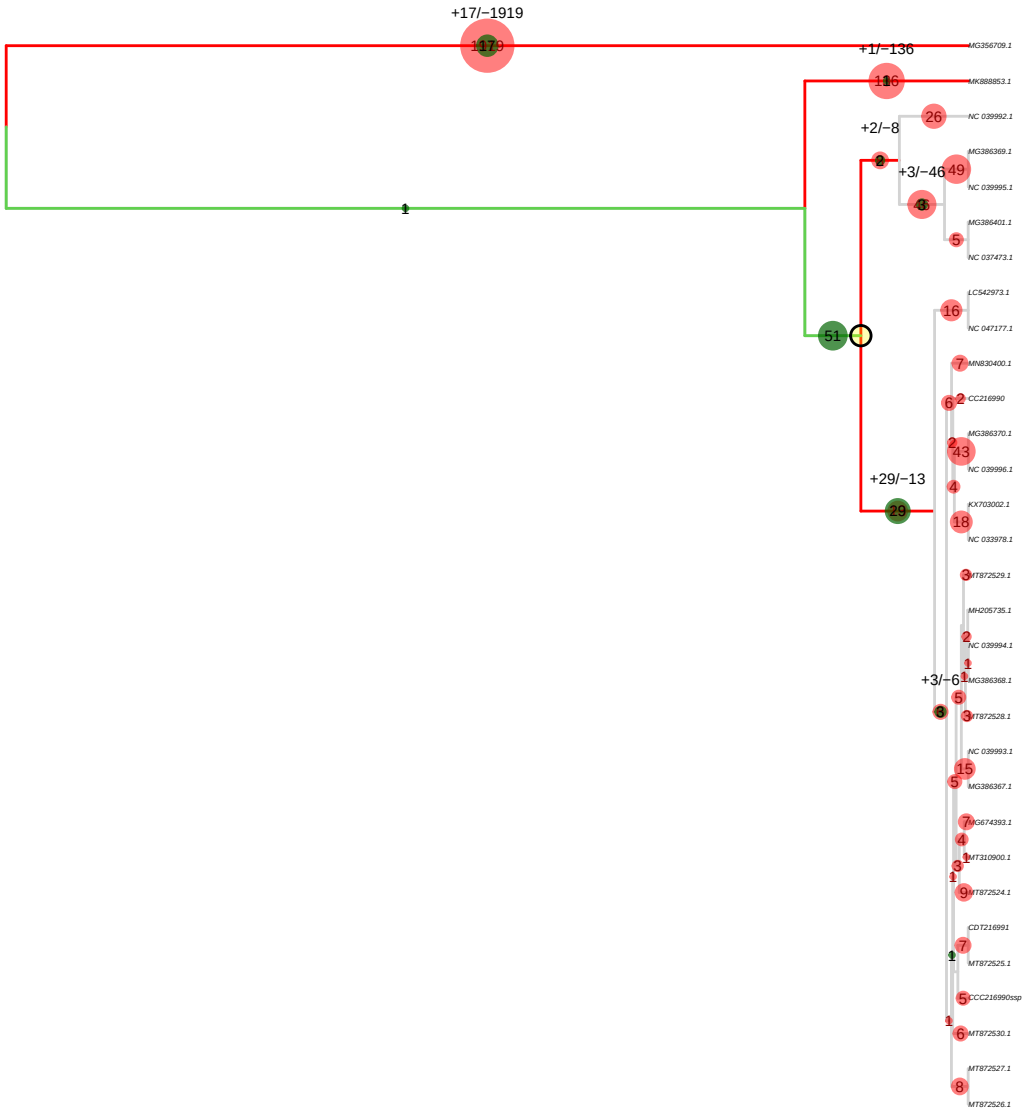


Figure 36: Phylogenetic placement results for the *Betula* chloroplast reads, using both transition and transversion SNPs, and using reads merged from all layers and sites. The numbers on each edge represent the number of supporting (+) and conflicting (-) SNPs in the ancient *Betula* environmental genome overlapping the reference SNPs assigned to that edge. There is a clear placement for the ancient *Betula* environmental genome on the edge marked +29/-13.

5.3.2 Molecular dating

Here I performed molecular dating on an ancient environmental *Betula* sequence using two different methods. The first, using pathPhynder and explicit SNP-counting, yielded a point estimate of 2.2 million years old, and a 95% HPD of [0.86, 3.72] Mya. On the other hand, using BEAST yielded a median estimate of 1.3 million years and a 95% HPD of [0.68, 2.02] Mya. The reason for the discordance between these estimates is not entirely clear, but could be due to multiple factors. First, I only ran BEAST using approximately a third of the sites in the chloroplast genome. This was necessary because only this fraction of sites could be treated as reliable in the ancient genome, and BEAST does not deal well with large amounts of missing data. In particular, it will integrate over all possibilities for a missing site, flattening the posterior. As discussed in Chapter 1, reconstructing entire phylogenies including ancient environmental DNA is generally difficult due to high amounts of missing data. However, these sites in the fraction of the genome with high coverage in the ancient sample have highly confident calls for the ancient sample, as validated by investigating the annotation and resulting protein alignment. On the other hand, using pathPhynder explicitly does not have this limitation, so that estimates can be made concerning branch lengths in units of SNPs using all of the variation in the modern reference panel, and only requires limiting to the third of the ancient chloroplast genome when necessary (e.g. when estimating the private branch length), then extrapolating.

That said, the extrapolation done in the SNP-counting analysis is somewhat crude and probably biased, in that it does not account for the fact that regions that meet the depth cutoff may be fundamentally different than those that fall below the depth cutoff, in a number of important ways. First, regions which meet the depth cutoff are more likely to contain variants unique to the *Betula* genus. This is a direct consequence of the competitive mapping pipeline used to assign reads to each taxon when using environmental DNA samples. In particular, the mapping pipeline will inevitably classify reads from conserved regions shared with plants outside the genus *Betula* to higher taxonomic levels, and so the coverage on the ancient *Betula* chloroplast will be lower in more conserved regions. If the fraction of the genome used is more likely to be variable than the rest of it, the extrapolated number of private variants could be an overestimate. Furthermore, these more conserved regions are probably less informative on average regarding evolution within the *Betula* phylogeny as well, so that the BEAST analysis may well be capturing more than just a third of the evolutionary information within the phylogenetic tree when using just a third of the sites. Another consequence of the competitive mapping pipeline is that reads assigned to *Betula* must meet a 90% similarity threshold. This means that reads which are highly divergent in the ancient sample will not be assigned to the genus or included in the analysis at all. It is difficult to

weigh the relative impact that all of these factors might have on the simple extrapolation done in the pathPhynder/SNP-counting analysis, and their consequences on the final molecular date.

Even though BEAST could only be used on approximately a third of the chloroplast genome, these sites could be partitioned into sets with variable rates. The pathPhynder/SNP-counting approach does not allow for this relaxed clock, which is arguably more correct. That said, using the Jukes-Cantor model within BEAST, which assumes no variation between substitution types nor sites within each partition, gave a similar result in terms of the inferred age of the *Betula* sample (see Figure 33), so although the relaxed clock is technically more correct, it could not have had a very large effect on results. Lastly, BEAST uses a Bayesian approach and integrates over many possibilities, which gives a posterior distribution rather than a point estimate.

Ultimately, I felt that the BEAST method was more reliable, because it did not require an extrapolation procedure which is likely biased, because it allowed for variable rates, and because I ran it on only sites which met a depth cutoff in the ancient sample. Though the 1.3 million year old median date estimate was younger than expected and than determined by geological dating methods, the 95% HPD does include two million years. Therefore, we can conclude that the molecular dating analysis here is consistent with the geological date.

5.4 Conclusion

In this chapter, I sought to date chloroplast DNA from an ancient environmental sample. This was to my knowledge the first time a sequence from environmental DNA has been molecularly dated, and was complicated by the potential presence of mixed species or individuals in the sample. For example, I could not perform molecular dating on the ancient *Populus* sequences, which rather clearly contained a mixture of reads from multiple species. Furthermore, an attempt to date the ancient *Salix* sequences was thwarted by an extremely high apparent number of private SNPs, due to deep divergence from modern *Salix* and potentially to high diversity in the species that contributed to the read set.

However, I was able to use the *Betula* sequence for molecular dating, and both methods used in this chapter yielded HPD intervals that overlapped with the geological date estimate, lending confidence to the use of molecular dating on sequences extracted from eDNA to supplement geological dating methods. In theory, this also means that many independent dates could be obtained from different taxa in ancient environmental DNA samples when sufficient sequence data is available, as should increasingly be the case in the near future, and when certain conditions are met (e.g. the sample does not contain a mixture of divergent species or individuals). Additionally, molecular dating is not necessarily limited to the mitochondrial or chloroplast genomes and is often performed

on nuclear genes. The potential to obtain many independent dates both from different taxa and independent non-recombining loci from ancient environmental DNA could be a powerful tool to determine or validate the age of sediment samples.

Table 3: List of chloroplast reference sequences and accession IDs

Accession ID	**Species**	**Source**
MG262367.1	*Salix rehderiana*	NCBI Genbank
MG262368.1	*Salix rorida*	NCBI Genbank
MG262369.1	*Salix taoensis*	NCBI Genbank
MT551159.1	*Salix argyracea*	NCBI Genbank
MT551160.1	*Salix dasyclados*	NCBI Genbank
MT551161.1	*Salix eriocephala*	NCBI Genbank
MT551162.1	*Salix integra*	NCBI Genbank
MT551163.1	*Salix suchowensis*	NCBI Genbank
NC037422.1	*Salix chaenomeloides*	NCBI Genbank
NC037423.1	*Salix hypoleuca*	NCBI Genbank
NC037424.1	*Salix magnifica*	NCBI Genbank
NC037425.1	*Salix minjiangensis*	NCBI Genbank
NC037426.1	*Salix paraplesia*	NCBI Genbank
NC037427.1	*Salix rehderiana*	NCBI Genbank
NC037428.1	*Salix rorida*	NCBI Genbank
NC037429.1	*Salix taoensis*	NCBI Genbank
MN830400.1	*Betula costata*	NCBI Genbank
NC037473.1	*Betula cordifolia*	NCBI Genbank
MG386401.1	*Betula cordifolia*	NCBI Genbank
NC047177.1	*Betula chichibuensis*	NCBI Genbank
LC542973.1	*Betula chichibuensis*	NCBI Genbank
NC033978.1	*Betula nana*	NCBI Genbank
KX703002.1	*Betula nana*	NCBI Genbank
MK888853.1	*Betula alnoides*	NCBI Genbank
MT872524.1	*Betula nana*	NCBI Genbank
MT872530.1	*Betula nana*	NCBI Genbank
MT872529.1	*Betula nana*	NCBI Genbank
MT872528.1	*Betula nana*	NCBI Genbank

MT872527.1	*Betula nana*	NCBI Genbank
MT872526.1	*Betula nana*	NCBI Genbank
MT872525.1	*Betula nana*	NCBI Genbank
MT310900.1	*Betula microphylla*	NCBI Genbank
MH205735.1	*Betula platyphylla*	NCBI Genbank
NC039996.1	*Betula pubescens*	NCBI Genbank
NC039995.1	*Betula populifolia*	NCBI Genbank
MG674393.1	*Betula halophila*	NCBI Genbank
NC039994.1	*Betula platyphylla*	NCBI Genbank
NC039993.1	*Betula occidentalis*	NCBI Genbank
MG386370.1	*Betula pubescens*	NCBI Genbank
NC039992.1	*Betula lenta*	NCBI Genbank
MG386369.1	*Betula populifolia*	NCBI Genbank
MG386368.1	*Betula platyphylla*	NCBI Genbank
MG386367.1	*Betula occidentalis*	NCBI Genbank
-	*Betula fruticosa*	PhyloNorway
-	*Betula nana*	PhyloNorway
-	*Betula nana*	PhyloNorway
MG356709.1	*Alnus rubra*	NCBI Genbank
KJ664927.1	*Populus balsamifera*	NCBI Genbank
MT482535.1	*Populus tremula*	NCBI Genbank
NC009143.1	*Populus trichocarpa*	NCBI Genbank
NC024735.1	*Populus balsamifera*	NCBI Genbank
MT482542.1	*Populus rotundifolia*	NCBI Genbank
MT482541.1	*Populus wilsonii*	NCBI Genbank
MT482538.1	*Populus trinervis*	NCBI Genbank
MN864049.1	*Populus koreana*	NCBI Genbank
NC031371.1	*Populus ilicifolia*	NCBI Genbank
MT407464.1	*Populus davidiana*	NCBI Genbank

6 Conclusion

This work shows the potential of ancient environmental DNA to go further than simply studying the presence of organisms in ancient ecosystems, in particular that the actual sequence content and variation in aeDNA can be exploited to make both population genetic and phylogenetic inferences of the taxa in the sample. In this thesis, I presented four studies. First, in Chapter 2, I constructed a theoretical framework to estimate the accuracy of taxa assignment. This framework is a simplified version of reality, assuming that there are only two reference taxa, though it does consider a wide variety of population genetic and reference parameters. In the future, it would be helpful to find a way to extend it to consider entire reference databases. Doing this rigorously would require extensive multiple integrals and would likely not be computationally feasible, but I think it would be possible to find a reasonable approximation. It would be even more helpful to implement it as part of an existing competitive mapping software such as ngsLCA (Wang et al., 2022). This way, researchers could use accuracy estimates as a cutoff to assign reads to taxa, instead of less reliable measures like sequence similarity or uniqueness of mapping. Knowledge of an accuracy estimate for read sets assigned to individual taxa would also inform the reliability of downstream population genetic or phylogenetic analyses. Lastly, reference databases could be constructed in such a way to maximize accuracy.

In the third chapter, I presented a phylogenetic placement method for ancient DNA and applied it to data from the Arctic aeDNA read sets assigned to mammoth and horse from the last 50,000 years. A phylogenetic placement method, in particular one which provides a visualization based on individual SNPs instead of (or in addition to) likelihood, is especially good for low coverage and mixed data such as aeDNA. This is because the visualization can expose the presence of mixed species or populations by showing supporting SNPs on multiple paths, whereas a likelihood method would simply place the sample basally with no further information. Additionally, knowledge of individual SNPs gives a more concrete measure of reliability for sample placement. Using this method on mammoths and horse read sets yielded the possibility of a previously undiscovered new clade of mammoths, and extended the time span in which an existing clade was known to exist. In the future, a pipeline using pathPhynder could be more streamlined to work on multiple read sets assigned to individual taxa at once.

In the fourth chapter, I performed a phylogenetic and population genetic analyses on black bear and giant short-faced bear DNA extracted from 14-16,000 year old cave soil in Northern Mexico. As mentioned in the last paragraph, these two species are close enough that I could use pathPhynder to deconvolute their signal on the mitochondrial phylogenetic Ursid tree and see the presence of both species in the reads. Next, I used 83 modern black bear samples from across the United States

to contextualize the ancient black bear DNA. Using a principal component analysis and genetic Hamming distance, I determined that the ancient Mexican samples are most genetically related to modern samples located in the Eastern United States. Similarly, an admixture graph analysis showed that the Mexican sample might be ancestrally related to the Eastern United States samples, which may have then given rise to the modern Kenai population in Alaska. Using these results, I inferred a working model of the recent demographic history of black bears across America in relation to the ice sheets in the Last Glacial Maximum. Lastly, I used three new high quality giant short-faced bear genomes to contextualize my ancient giant short-faced bear read set. I found that the ancient Mexican sample was much further away from the three modern samples than the latter were to each other, in terms of genetic Hamming distance. This was one of the first studies which used the sequence content of a read set from aeDNA to make population genetic inferences about the demographic and evolutionary history of individual species (also see Zavala et al. (2021)).

Lastly in the fifth chapter, I provide a molecular dating analysis of aeDNA extracted from a deposit in Northern Greenland, using reads assigned to the *Betula* or birch tree genus. This deposit was dated to approximately 2 million years ago using geological methods, and I wanted to confirm this date using molecular methods on the chloroplast genome. I considered three genera originally: *Betula*, *Salix* and *Populus*, because of their high number of assigned reads and the existence of a sufficient number of reference chloroplast genomes. *Populus* was excluded due to an initial phylogenetic placement that determined it contained a mixture of species, and *Salix* was later excluded due to its strangely high number of private SNPs, which may also indicate a mixed sample. I then used two different methods to date the ancient *Betula* read set, both calibrated with an existing divergence date of *Betula* and *Alnus* at the root of the phylogeny. First, I built on the phylogenetic placement and used a SNP-counting method to explicitly estimate how far into the phylogeny the ancient sample lay. Second, I used BEAST, a Bayesian phylogenetic software. As discussed in the chapter, both methods had advantages and disadvantages and gave somewhat different answers, but both 95% posterior intervals overlapped the estimate of 2 million years, consistent with the geological date.

These studies offer only a start on using ancient environmental DNA for more than observing the presence of species in ancient ecosystems. I am particularly interested in the huge potential in the near future to apply ancient environmental DNA to study and inform how we approach the current biodiversity and climate crisis. First of all, studying the evolutionary history of different species can help with questions of species delimitation, which is relevant because conservation policy is often set per individual species. Similarly, since genetic diversity is a recognized form of biodiversity, directly demonstrating its decline using ancient DNA can impact policy (Jensen

et al., 2022). Understanding the demographic history of species and how they react to large climactic or environmental shifts such as the Holocene-Pleistocene boundary or the retreat of the ice sheets after the Last Glacial Maximum can help predict how they will react to imminent changes. It can also inform strategies such as re-introductions into areas in which species used to live but are no longer present due to recent habitat loss. Functional mutations and selective effects can be inferred and used to preserve species by giving them a better chance at withstanding upcoming environmental changes. For example, functional mutations which encode for heat resistance could be preferentially selected in the genotypes of new trees in replanted forests. Similarly, one could infer the relationship between genetic load, fitness and the risk of extinction by using extinct species, which might inform the use of fitness or genetic load measurements to categorize living species as threatened or extinct (Bertorelle et al., 2022). Lastly, extinct genetic diversity could conceptually be reintroduced into extant species which have undergone a bottleneck and in which low diversity is a threat to their ongoing existence.

Many of the applications listed above are, in theory, possible using ancient DNA isolated from fossils when they are available. However, on top of being useful when fossils do not exist and the potential longer persistence of DNA in some environmental sources as opposed to fossils, one of the benefits of using ancient environmental DNA is its ability to capture genetic information from the entire ecosystem at once. Importantly, mean temperatures in the past have sometimes been much higher than temperatures in the present day, but greenhouse gas emissions are causing temperatures to rise quickly, which may create an environment similar to those which existed at certain times in the past. Especially in a time series context, studying the way that entire ecosystems react to large climactic shifts could help us determine which ones will be the most resilient in the future. On the other hand, we could attempt to predict which species will be less likely to survive, and prioritize their conservation. Ancient environmental DNA sources such as sediment cores also often contain information about more local events such as volcanic eruptions, fires or changes in diatom content, so that sequencing DNA in a time series after these events could provide a real-time record as to their past effects on local ecosystems, which could inform predictions as to their future impacts as well. Again, this would be especially useful when considering the correlations between different species, for example to determine that the regrowth of certain fungal taxa after fire requires the presence of their plant symbiotes or vice versa.

Overall, ancient environmental DNA has so far been limited by a suite of factors including fragmentation, damage, low coverage and a mixture of samples. The present work offers a start to overcoming these limitations and inferring evolutionary histories using aeDNA, both by giving improved computational methods for taxa assignment and by presenting some of the first studies

to use aeDNA for phylogenetic and population genetic purposes. I hope that this can inspire future studies of this kind, especially those which can be applied to help us respond to the climate crisis and play some small role in safeguarding the rich biodiversity of our world.

References

Achilli, A., Olivieri, A., Soares, P., Lancioni, H., Kashani, B. H., Perego, U. A., Nergadze, S. G., Carossa, V., Santagostino, M., Capomaccio, S., Felicetti, M., Al-Achkar, W., Penedo, M. C. T., Verini-Supplizi, A., Houshmand, M., Woodward, S. R., Semino, O., Silvestrelli, M., Giulotto, E., Pereira, L., Bandelt, H.-J., and Torroni, A. (2012). Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. *Proceedings of the National Academy of Sciences*, 109(7):2449–2454.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Anari, S. S. (2020). Towards comparative pan-genomics. *PhD thesis*, page 113.

Anthony, D. W. (2010). *The horse, the wheel, and language*. Princeton University Press, Princeton, NJ.

Arcones, A., Ponti, R., and Vieites, D. R. (2021). Mitochondrial substitution rates estimation for divergence time analyses in modern birds based on full mitochondrial genomes. *Ibis*, 163(4):1463–1471.

Ardelean, C. F., Becerra-Valdivia, L., Pedersen, M. W., Schwenninger, J.-L., Oviatt, C. G., Macías-Quintero, J. I., Arroyo-Cabrales, J., Sikora, M., Ocampo-Díaz, Y. Z. E., Rubio-Cisneros, I. I., Watling, J. G., de Medeiros, V. B., Oliveira, P. E. D., Barba-Pingarón, L., Ortiz-Butrón, A., Blancas-Vázquez, J., Rivera-González, I., Solís-Rosales, C., Rodríguez-Ceja, M., Gandy, D. A., Navarro-Gutierrez, Z., Rosa-Díaz, J. J. D. L., Huerta-Arellano, V., Marroquín-Fernández, M. B., Martínez-Riojas, L. M., López-Jiménez, A., Higham, T., and Willerslev, E. (2020). Evidence of human occupation in mexico around the last glacial maximum. *Nature*, 584(7819):87–92.

Armbrecht, L., Weber, M. E., Raymo, M. E., Peck, V. L., Williams, T., Warnock, J., Kato, Y., Hernández-Almeida, I., Hoem, F., Reilly, B., Hemming, S., Bailey, I., Martos, Y. M., Gutjahr, M., Percuoco, V., Allen, C., Brachfeld, S., Cardillo, F. G., Du, Z., Fauth, G., Fogwill, C., Garcia, M., Glüder, A., Guitard, M., Hwang, J.-H., Iizuka, M., Kenlee, B., O'Connell, S., Pérez, L. F., Ronge, T. A., Seki, O., Tauxe, L., Tripathi, S., and Zheng, X. (2022). Ancient marine sediment DNA reveals diatom transition in antarctica. *Nature Communications*, 13(1).

Bell, K. L., Petit, R. A., Cutler, A., Dobbs, E. K., Macpherson, J. M., Read, T. D., Burgess, K. S., and Brosi, B. J. (2021). Comparing whole-genome shotgun sequencing and DNA metabarcoding approaches for species identification and quantification of pollen species mixtures. *Ecology and Evolution*, 11(22):16082–16098.

Bertorelle, G., Raffini, F., Bosse, M., Bortoluzzi, C., Iannucci, A., Trucchi, E., Morales, H. E., and van Oosterhout, C. (2022). Genetic load: genomic estimates and applications in non-model animals. *Nature Reviews Genetics*, 23(8):492–503.

Boddé, M., Makunin, A., Ayala, D., Bouafou, L., Diabaté, A., Ekpo, U. F., Kientega, M., Goff, G. L., Makanga, B. K., Ngangue, M. F., Omitola, O. O., Rahola, N., Tripet, F., Durbin, R., and Lawniczak, M. K. (2022). High resolution species assignment of anopheles mosquitoes using k-mer distances on targeted sequences. *eLife*, 11.

Bofkin, L. and Goldman, N. (2006). Variation in evolutionary processes at different codon positions. *Molecular Biology and Evolution*, 24(2):513–521.

Brennan, G. L., , Potter, C., de Vere, N., Griffith, G. W., Skjøth, C. A., Osborne, N. J., Wheeler, B. W., McInnes, R. N., Clewlow, Y., Barber, A., Hanlon, H. M., Hegarty, M., Jones, L., Kurganskiy, A., Rowney, F. M., Armitage, C., Adams-Groom, B., Ford, C. R., Petch, G. M., and Creer, S. (2019). Temperate airborne grass pollen defined by spatio-temporal shifts in community composition. *Nature Ecology &amp Evolution*, 3(5):750–754.

Briggs, A. W., Stenzel, U., Johnson, P. L. F., Green, R. E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M. T., Lachmann, M., and Pääbo, S. (2007). Patterns of damage in genomic DNA sequences from a neandertal. *Proceedings of the National Academy of Sciences*, 104(37):14616–14621.

Brigham-Grette, J., Melles, M., Minyuk, P., Andreev, A., Tarasov, P., DeConto, R., Koenig, S., Nowaczyk, N., Wennrich, V., Rosén, P., Haltia, E., Cook, T., Gebhardt, C., Meyer-Jacob, C., Snyder, J., and Herzschuh, U. (2013). Pliocene warmth, polar amplification, and stepped pleistocene cooling recorded in NE arctic russia. *Science*, 340(6139):1421–1427.

Broad Institute (2019). Picard toolkit. `https://broadinstitute.github.io/picard/`.

Brown, E. A., Chain, F. J. J., Crease, T. J., MacIsaac, H. J., and Cristescu, M. E. (2015). Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities? *Ecology and Evolution*, 5(11):2234–2251.

Brys, R., Halfmaerten, D., Neyrinck, S., Mauvisseau, Q., Auwerx, J., Sweet, M., and Mergeay, J. (2020). Reliable eDNA detection and quantification of the european weather loach (misgurnus fossilis). *Journal of Fish Biology*, 98:399–414.

Cai, P., Huang, Q.-Y., and Zhang, X.-W. (2006). Interactions of DNA with clay minerals and soil colloidal particles and protection against degradation by DNase. *Environmental Science &amp Technology*, 40(9):2971–2976.

Capo, E., Giguet-Covex, C., Rouillard, A., Nota, K., Heintzman, P. D., Vuillemin, A., Ariztegui, D., Arnaud, F., Belle, S., Bertilsson, S., Bigler, C., Bindler, R., Brown, A. G., Clarke, C. L., Crump, S. E., Debroas, D., Englund, G., Ficetola, G. F., Garner, R. E., Gauthier, J., Gregory-Eaves, I., Heinecke, L., Herzschuh, U., Ibrahim, A., Kisand, V., Kjær, K. H., Lammers, Y., Littlefair, J., Messager, E., Monchamp, M.-E., Olajos, F., Orsi, W., Pedersen, M. W., Rijal, D. P., Rydberg, J., Spanbauer, T., Stoof-Leichsenring, K. R., Taberlet, P., Talas, L., Thomas, C., Walsh, D. A., Wang, Y., Willerslev, E., van Woerkom, A., Zimmermann, H. H., Coolen, M. J. L., Epp, L. S.,

Domaizon, I., Alsos, I. G., and Parducci, L. (2021). Lake sedimentary DNA research on past terrestrial and aquatic biodiversity: Overview and recommendations. *Quaternary*, 4(1):6.

Carew, M. E., Pettigrove, V. J., Metzeling, L., and Hoffmann, A. (2013). Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species. *Frontiers in Zoology*, 10(45).

Casas-Marce, M., Marmesat, E., Soriano, L., Martinez-Cruz, B., Lucena-Perez, M., Nocete, F., Rodriguez-Hidalgo, A., Canals, A., Nadal, J., Detry, C., Bernáldez-Sánchez, E., Fernández-Rodríguez, C., Pérez-Ripoll, M., Rodriguez, A., Revilla, E., Delibes, M., and Godoy, J. A. (2017). Spatiotemporal dynamics of genetic variation in the iberian lynx along its path to extinction reconstructed with ancient DNA. *Molecular Biology and Evolution*, 34:2893–2907.

Chang, D., Knapp, M., Enk, J., Lippold, S., Kircher, M., Lister, A., MacPhee, R. D. E., Widga, C., Czechowski, P., Sommer, R., Hodges, E., Stümpel, N., Barnes, I., Dalén, L., Derevianko, A., Germonpré, M., Hillebrand-Voiculescu, A., Constantin, S., Kuznetsova, T., Mol, D., Rathgeber, T., Rosendahl, W., Tikhonov, A. N., Willerslev, E., Hannon, G., Lalueza-Fox, C., Joger, U., Poinar, H., Hofreiter, M., and Shapiro, B. (2017). The evolutionary and phylogeographic history of woolly mammoths: a comprehensive mitogenomic analysis. *Scientific Reports*, 7(1).

Chylek, P., Folland, C., Klett, J. D., Wang, M., Hengartner, N., Lesins, G., and Dubey, M. K. (2022). Annual mean arctic amplification 1970–2020: Observed and simulated by CMIP6 climate models. *Geophysical Research Letters*, 49(13).

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2015). GenBank. *Nucleic Acids Research*, 44(D1):D67–D72.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

de Filippo, C., Meyer, M., and Prüfer, K. (2018). Quantifying and reducing spurious alignments for the analysis of ultra-short ancient DNA sequences. *BMC Biology*, 16(1):121.

De Sanctis, B., Money, D., Pedersen, M. W., and Durbin, R. (2022). A theoretical analysis of taxonomic binning accuracy. *Molecular Ecology Resources*, 22(6):2208–2219.

Debruyne, R., Chu, G., King, C. E., Bos, K., Kuch, M., Schwarz, C., Szpak, P., Gröcke, D. R., Matheus, P., Zazula, G., Guthrie, D., Froese, D., Buigues, B., de Marliave, C., Flemming, C., Poinar, D., Fisher, D., Southon, J., Tikhonov, A. N., MacPhee, R. D., and Poinar, H. N. (2008). Out of america: Ancient DNA evidence for a new world origin of late quaternary woolly mammoths. *Current Biology*, 18(17):1320–1326.

Delisle, I. and Strobeck, C. (2002). Conserved primers for rapid sequencing of the complete mitochondrial genome from carnivores, applied to three species of bears. *Molecular Biology and Evolution*, 19(3):357–361.

Djurhuus, A., Closek, C. J., Kelly, R. P., Pitz, K. J., Michisaki, R. P., Starks, H. A., Walz, K. R., Andruszkiewicz, E. A., Olesin, E., Hubbard, K., Montes, E., Otis, D., Muller-Karger, F. E., Chavez, F. P., Boehm, A. B., and Beritbart, M. (2020). Environmental DNA reveals seasonal shifts and potential interactions in a marine community. *Nature Communications*, 11(254).

dos Reis, M., Inoue, J., Hasegawa, M., Asher, R. J., Donoghue, P. C. J., and Yang, Z. (2012). Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proceedings of the Royal Society B: Biological Sciences*, 279(1742):3491–3500.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5):e88.

Duc, D. L., Velluva, A., Cassatt-Johnstone, M., Olsen, R.-A., Baleka, S., Lin, C.-C., Lemke, J. R., Southon, J. R., Burdin, A., Wang, M.-S., Grunewald, S., Rosendahl, W., Joger, U., Rutschmann, S., Hildebrandt, T. B., Fritsch, G., Estes, J. A., Kelso, J., Dalén, L., Hofreiter, M., Shapiro, B., and Schöneberg, T. (2022). Genomic basis for skin phenotype and cold adaptation in the extinct Steller's sea cow. *Science Advances*, 8(5):eabl6496.

Dugal, L., Thomas, L., Jensen, M. R., Sigsgaard, E. E., Simpson, T., Jarman, S., Thomsen, P. F., and Meekan, M. (2022). Individual haplotyping of whale sharks from seawater environmental dna. *Molecular Ecology Resources*, 22(1):56–65.

Dussex, N., Bergfeldt, N., Prado, V., Dehasque, M., Díez del Molino, D., Ersmark, E., Kanellidou, F., Larsson, P., Lemež, □., Lord, E., Mármol-Sánchez, E., Meleg, I., Måsviken, J., Naidoo, T., Studerus, J., Vicente, M., von Seth, J., Götherström, A., Dalén, L., and Heintzman, P. (2021). Integrating multi-taxon palaeogenomes and sedimentary ancient DNA to study past ecosystem dynamics. *Proceedings of the Royal Society B: Biological Sciences*, 288:20211252.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.

Enk, J., Devault, A., Widga, C., Saunders, J., Szpak, P., Southon, J., Rouillard, J.-M., Shapiro, B., Golding, G. B., Zazula, G., Froese, D., Fisher, D. C., MacPhee, R. D. E., and Poinar, H. (2016). Mammuthus population dynamics in late pleistocene north america: Divergence, phylogeography, and introgression. *Frontiers in Ecology and Evolution*, 4.

Everett, R., Cribdon, B., Kistler, L., Ware, R., and Allaby, R. (2021). MetaDamage tool: Examining post-mortem damage in sedaDNA on a metagenomic scale. *The 23rd EGU General Assembly*.

Felsenstein, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology*, 22(3):240–249.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology*, 27(4):401–410.

Feuerborn, T. R., Palkopoulou, E., van der Valk, T., von Seth, J., Munters, A. R., Pečnerová, P., Dehasque, M., Ureña, I., Ersmark, E., Lagerholm, V. K., Krzewińska, M., Rodríguez-Varela, R., Götherström, A., Dalén, L., and Díez-del Molino, D. (2020). Competitive mapping allows for the identification and exclusion of human DNA contamination in ancient faunal genomic datasets. *BMC Genomics*, 21(1):844.

Fisher, R. A. (1923). XXI.—on the dominance ratio. *Proceedings of the Royal Society of Edinburgh*, 42:321–341.

Fourment, M. and Holmes, E. C. (2016). Seqotron: a user-friendly sequence editor for mac OS x. *BMC Research Notes*, 9(1).

François, O. and Jay, F. (2020). Factor analysis of ancient population genomic samples. *Nature Communications*, 11(1).

Fulton, T. L. and Shapiro, B. (2019). Setting up an ancient DNA laboratory. pages 1–13. Springer New York.

Gaunitz, C., Fages, A., Hanghøj, K., Albrechtsen, A., Khan, N., Schubert, M., Seguin-Orlando, A., Owens, I. J., Felkel, S., Bignon-Lau, O., de Barros Damgaard, P., Mittnik, A., Mohaseb, A. F., Davoudi, H., Alquraishi, S., Alfarhan, A. H., Al-Rasheid, K. A. S., Crubézy, E., Benecke, N., Olsen, S., Brown, D., Anthony, D., Massy, K., Pitulko, V., Kasparov, A., Brem, G., Hofreiter, M., Mukhtarova, G., Baimukhanov, N., Lõugas, L., Onar, V., Stockhammer, P. W., Krause, J., Boldgiv, B., Undrakhbold, S., Erdenebaatar, D., Lepetz, S., Mashkour, M., Ludwig, A., Wallner, B., Merz, V., Merz, I., Zaibert, V., Willerslev, E., Librado, P., Outram, A. K., and Orlando, L. (2018). Ancient genomes revisit the ancestry of domestic and przewalski's horses. *Science*, 360(6384):111–114.

Gilbert, M. T. P., Jenkins, D. L., Götherstrom, A., Naveran, N., Sanchez, J. J., Hofreiter, M., Thomsen, P. F., Binladen, J., Higham, T. F. G., Yohe, R. M., Parr, R., Cummings, L. S., and Willerslev, E. (2008). DNA from Pre-Clovis human coprolites in Oregon, North America. *Science*, 320(5877):786–789.

Gillette, D. D. and Madsen, D. B. (1993). The Columbian mammoth, imammuthus columbi/i, from the Wasatch Mountains of central Utah. *Journal of Paleontology*, 67(4):669–680.

Ginolhac, A., Rasmussen, M., Gilbert, M. T. P., Willerslev, E., and Orlando, L. (2011). mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics*, 27(15):2153–2155.

Gire, S. K., Goba, A., Andersen, K. G., Sealfon, R. S. G., Park, D. J., Kanneh, L., Jalloh, S., Momoh, M., Fullah, M., Dudas, G., Wohl, S., Moses, L. M., Yozwiak, N. L., Winnicki, S., Matranga, C. B., Malboeuf, C. M., Qu, J., Gladden, A. D., Schaffner, S. F., Yang, X., Jiang, P.-P., Nekoui, M., Colubri, A., Coomber, M. R., Fonnie, M., Moigboi, A., Gbakie, M., Kamara, F. K., Tucker, V., Konuwa, E., Saffa, S., Sellu, J., Jalloh, A. A., Kovoma, A., Koninga, J., Mustapha, I., Kargbo, K., Foday, M., Yillah, M., Kanneh, F., Robert, W., Massally, J. L. B., Chapman, S. B., Bochicchio, J., Murphy, C., Nusbaum, C., Young, S., Birren, B. W., Grant, D. S., Scheiffelin,

J. S., Lander, E. S., Happi, C., Gevao, S. M., Gnirke, A., Rambaut, A., Garry, R. F., Khan, S. H., and Sabeti, P. C. (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202):1369–1372.

Gokhman, D., Lavi, E., Prüfer, K., Fraga, M. F., Riancho, J. A., Kelso, J., Pääbo, S., Meshorer, E., and Carmel, L. (2014). Reconstructing the DNA methylation maps of the Neandertal and the Denisovan. *Science*, 344(6183):523–527.

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., Spear, S. F., McKee, A., Oyler-McCance, S. J., Cornman, R. S., Laramie, M. B., Mahon, A. R., Lance, R. F., Pilliod, D. S., Strickler, K. M., Waits, L. P., Fremier, A. K., Takahara, T., Herder, J. E., and Taberlet, P. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution*, 7(11):1299–1307.

Gorički, S., Stanković, D., Snoj, A., Kuntner, M., Jeffery, W. R., Trontelj, P., Pavićević, M., Grizelj, Z., Năpăruş-Aljančič, M., and Aljančič, G. (2017). Environmental DNA in subterranean biology: range extension and taxonomic implications for proteus. *Scientific Reports*, 7(45054).

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

Green, R. E., Malaspinas, A.-S., Krause, J., Briggs, A. W., Johnson, P. L., Uhler, C., Meyer, M., Good, J. M., Maricic, T., Stenzel, U., Prüfer, K., Siebauer, M., Burbano, H. A., Ronan, M., Rothberg, J. M., Egholm, M., Rudan, P., Brajković, D., Kućan, Ž., Gušić, I., Wikström, M., Laakkonen, L., Kelso, J., Slatkin, M., and Pääbo, S. (2008). A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell*, 134(3):416–426.

Haile, J., Froese, D. G., MacPhee, R. D. E., Roberts, R. G., Arnold, L. J., Reyes, A. V., Rasmussen, M., Nielsen, R., Brook, B. W., Robinson, S., Demuro, M., Gilbert, M. T. P., Munch, K., Austin, J. J., Cooper, A., Barnes, I., Möller, P., and Willerslev, E. (2009). Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *Proceedings of the National Academy of Sciences*, 106(52):22352–22357.

Haller, B. C. and Messer, P. W. (2019). SLiM 3: Forward genetic simulations beyond the wright–fisher model. *Molecular Biology and Evolution*, 36(3):632–637.

Herbig, A., Maixner, F., Bos, K., Zink, A., Krause, J., and Huson, D. (2016). MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. *bioRxiv*.

Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A., and Wilson, A. C. (1984). DNA sequences from the quagga, an extinct member of the horse family. *Nature*, 312(5991):282–284.

Hong, C., Manimaran, S., Shen, Y., Perez-Rogers, J. F., Byrd, A. L., Castro-Nallar, E., Crandall, K. A., and Johnson, W. E. (2014). PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome*, 2(1).

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Fioretto, L. D. R., Davidson, C., Dodiya, K., Houdaigui, B. E., Fatima, R., Gall, A., Giron, C. G., Grego, T., Guijarro-Clarke, C., Haggerty, L., Hemrom, A., Hourlier, T., Izuogu, O. G., Juettemann, T., Kaikala, V., Kay, M., Lavidas, I., Le, T., Lemos, D., Martinez, J. G., Marugán, J. C., Maurel, T., McMahon, A. C., Mohanan, S., Moore, B., Muffato, M., Oheh, D. N., Paraschas, D., Parker, A., Parton, A., Prosovetskaia, I., Sakthivel, M. P., Salam, A. I. A., Schmitt, B. M., Schuilenburg, H., Sheppard, D., Steed, E., Szpak, M., Szuba, M., Taylor, K., Thormann, A., Threadgold, G., Walts, B., Winterbottom, A., Chakiachvili, M., Chaubal, A., Silva, N. D., Flint, B., Frankish, A., Hunt, S. E., IIsley, G. R., Langridge, N., Loveland, J. E., Martin, F. J., Mudge, J. M., Morales, J., Perry, E., Ruffier, M., Tate, J., Thybert, D., Trevanion, S. J., Cunningham, F., Yates, A. D., Zerbino, D. R., and Flicek, P. (2020). Ensembl 2021. *Nucleic Acids Research*, 49(D1):D884–D891.

Huang, D. I., Hefer, C. A., Kolosova, N., Douglas, C. J., and Cronk, Q. C. B. (2014). Whole plastome sequencing reveals deep plastid divergence and cytonuclear discordance between closely related balsam poplars, populus balsamifera and trichocarpa. *New Phytologist*, 204(3):693–703.

Huang, Y., Wang, J., Yang, Y., Fan, C., and Chen, J. (2017). Phylogenomic analysis and dynamic evolution of chloroplast genomes in salicaceae. *Frontiers in Plant Science*, 8.

Hübler, R., Key, F. M., Warinner, C., Bos, K. I., Krause, J., and Herbig, A. (2019). HOPS: automated detection and authentication of pathogen DNA in archaeological remains. *Genome Biology*, 20(1).

Huelsenbeck, J. P. and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755.

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

Hunter, M. E., Oyler-McCance, S. J., Dorazio, R. M., Fike, J. A., Smith, B. J., Hunter, C. T., Reed, R. N., and Hart, K. M. (2015). Environmental DNA (eDNA) sampling improves occurrence and detection estimates of invasive burmese pythons. *PLOS One*, 10(4):e0121655.

iBOL (2022). International barcode of life project (iBOL).

Ishige, T., Miya, M., U. M., Sado, T., Ushioda, M., Maebashi, K., Yonechi, R., Lagan, P., and Matsubayashi, H. (2017). Environmental dna reveals seasonal shifts and potential interactions in a marine community. *Biological Conservation*, 210:281–285.

ISOGG (2022). List of forensic and ancient DNA laboratories. *ISOGG Wiki*, https://isogg.org/wiki/List_of_forensic_and_ancient_DNA_laboratories.

Jensen, E. L., Díez-del Molino, D., Gilbert, M. T. P., Bertola, L. D., Borges, F., Cubric-Curik, V., de Navascués, M., Frandsen, P., Heuertz, M., Hvilsom, C., Jiménez-Mena, B., Miettinen, A., Moest, M., Pečnerová, P., Barnes, I., and Vernesi, C. (2022). Ancient and historical DNA in conservation policy. *Trends in Ecology & Evolution*.

Jensen, J. D., Payseur, B. A., Stephan, W., Aquadro, C. F., Lynch, M., Charlesworth, D., and Charlesworth, B. (2018). The importance of the neutral theory in 1968 and 50 years on: A response to kern and hahn 2018. *Evolution*, 73(1):111–114.

Jiao, X., Flouri, T., and Yang, Z. (2021). Multispecies coalescent and its applications to infer species phylogenies and cross-species gene flow. *National Science Review*, 8(12).

Jones, E. D. (2022). *Ancient DNA*. Yale University Press.

Kapp, J. D., Green, R. E., and Shapiro, B. (2021). A fast and efficient single-stranded genomic library preparation method optimized for ancient DNA. *Journal of Heredity*, 112(3):241–249.

Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, 12(5):e1004842.

Key, F. M., Posth, C., Krause, J., Herbig, A., and Bos, K. I. (2017). Mining metagenomic data sets for ancient DNA: Recommended protocols for authentication. *Trends in Genetics*, 33(8):508–520.

Kircher, M. (2011). Analysis of high-throughput ancient DNA sequencing data. In *Methods in Molecular Biology*, pages 197–228. Humana Press.

Kircher, M. (2012). Analysis of high-throughput ancient DNA sequencing data. *Methods in Molecular Biology*, 840:197–228.

Kolde, R. (2018). pheatmap: Pretty heatmaps.

Kornienko, I. V., Faleeva, T. G., Oreshkova, N. V., Grigoriev, S. E., Grigoreva, L. V., Simonov, E. P., Kolesnikova, A. I., Putintseva, Y. A., and Krutovsky, K. V. (2018). Complete mitochondrial genome of a woolly mammoth (mammuthus primigenius) from Maly Lyakhovsky Island (New Siberian islands, russia) and its phylogenetic assessment. *Mitochondrial DNA Part B*, 3(2):596–598.

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21):4453–4455.

Krause, J., Dear, P. H., Pollack, J. L., Slatkin, M., Spriggs, H., Barnes, I., Lister, A. M., Ebersberger, I., Pääbo, S., and Hofreiter, M. (2005). Multiplex amplification of the mammoth mitochondrial genome and the evolution of elephantidae. *Nature*, 439(7077):724–727.

Krause, J., Unger, T., Noçon, A., Malaspinas, A.-S., Kolokotronis, S.-O., Stiller, M., Soibelzon, L., Spriggs, H., Dear, P. H., Briggs, A. W., Bray, S. C., O'Brien, S. J., Rabeder, G., Matheus, P., Cooper, A., Slatkin, M., Pääbo, S., and Hofreiter, M. (2008). Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. *BMC Evolutionary Biology*, 8(1):220.

Krause-Kyora, B., Susat, J., Key, F. M., Kühnert, D., Bosse, E., Immel, A., Rinne, C., Kornell, S.-C., Yepes, D., Franzenburg, S., Heyne, H. O., Meier, T., Lösch, S., Meller, H., Friederich, S., Nicklisch, N., Alt, K. W., Schreiber, S., Tholey, A., Herbig, A., Nebel, A., and Krause, J. (2018). Neolithic and medieval virus genomes reveal complex evolution of hepatitis b. *eLife*, 7:e36666.

Krukov, I., de Sanctis, B., and de Koning, A. J. (2017). Wright-fisher exact solver (WFES): Scalable analysis of population genetic models without simulation or diffusion theory. *Bioinformatics*, 33(9):1416–1417.

Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., Stecher, G., and Hedges, S. B. (2022). TimeTree 5: An expanded resource for species divergence times. *Molecular Biology and Evolution*, 39(8).

Kumar, V., Lammers, F., Bidon, T., Pfenninger, M., Kolter, L., Nilsson, M. A., and Janke, A. (2017). The evolutionary history of bears is characterized by gene flow across species. *Scientific Reports*, 7(1).

Köster, K., Aaltonen, H., Berninger, F., Heinonsalo, J., Köster, E., Ribeiro-Kumara, C., Sun, H., Tedersoo, L., Zhou, X., and Pumpanen, J. (2021). Impacts of wildfire on soil microbiome in Boreal environments. *Current Opinion in Environmental Science & Health*, 22:100258.

Ladoukakis, E. D. and Zouros, E. (2017). Evolution and inheritance of animal mitochondrial DNA: rules and exceptions. *Journal of Biological Research-Thessaloniki*, 24(1).

Lammers, Y., Heintzman, P. D., and Alsos, I. G. (2021). Environmental palaeogenomic reconstruction of an ice age algal population. *Communications Biology*, 4(1):1–11.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–359.

Lemmon, A. R., Brown, J. M., Stanger-Hall, K., and Lemmon, E. M. (2009). The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and bayesian inference. *Systematic Biology*, 58(1):130–145.

Levsen, N. D., Tiffin, P., and Olson, M. S. (2012). Pleistocene speciation in the genus Populus (Salicaceae). *Systematic Biology*, 61(3):401.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993.

Li, H. (2018). Seqtk: Toolkit for processing sequences in fasta/q formats.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009a). The sequence alignment/map format and SAMtools. 25(16):2078–2079.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009b). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Librado, P., Gamba, C., Gaunitz, C., Sarkissian, C. D., Pruvost, M., Albrechtsen, A., Fages, A., Khan, N., Schubert, M., Jagannathan, V., Serres-Armero, A., Kuderna, L. F. K., Povolotskaya, I. S., Seguin-Orlando, A., Lepetz, S., Neuditschko, M., Thèves, C., Alquraishi, S., Alfarhan, A. H., Al-Rasheid, K., Rieder, S., Samashev, Z., Francfort, H.-P., Benecke, N., Hofreiter, M., Ludwig, A., Keyser, C., Marques-Bonet, T., Ludes, B., Crubézy, E., Leeb, T., Willerslev, E., and Orlando, L. (2017). Ancient genomic changes associated with domestication of the horse. *Science*, 356(6336):442–445.

Lister, A. M. and Sher, A. V. (2015). Evolution and dispersal of mammoths across the northern hemisphere. *Science*, 350(6262):805–809.

Liu, Y. and Mittler, J. E. (2008). Selection dramatically reduces effective population size in HIV-1 infection. *BMC Evolutionary Biology*, 8(1):133.

Lynggaard, C., Bertelsen, M. F., Jensen, A. V., Johnson, M. S., Frøslev, T. G., Tange, M., and Bohmann, K. (2022). Airborne environmental DNA for terrestrial vertebrate community monitoring. *Current Biology*, 32(3):701–707.e5.

MacHugh, D. E., Larson, G., and Orlando, L. (2017). Taming the past: Ancient DNA and the study of animal domestication. *Annual Review of Animal Biosciences*, 5(1):329–351.

Maier, R., Flegontov, P., Flegontova, O., Changmai, P., and Reich, D. (2022). On the limits of fitting complex models of population history to genetic data.

Martiniano, R., De Sanctis, B., Hallast, P., and Durbin, R. (2022). Placing ancient DNA sequences into reference phylogenies. *Molecular Biology and Evolution*, 39(2).

Matsen, F. A., Kodner, R. B., and Armbrust, E. V. (2010). pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 11(1).

Matthews, J. V., Telka, A. J., and Kuzmina, S. A. (2019). Late Neogene insect and other invertebrate fossils from Alaska and Arctic/Subarctic Canada. *Invertebrate Zoology*, 16(1):126–153.

Meier, J. (2021). Speciation population genomics: a how-to-guide.

Mikheyev, A., Zwick, A., Magrath, M. J., Grau, M. L., Qui, L., Su, Y. N., and Yeates, D. (2017). Museum genomics confirms that the Lord Howe Island stick insect survived extinction. *Current Biology*, 27(20):3157–3161.e4.

Miller, W., Schuster, S. C., Welch, A. J., Ratan, A., Bedoya-Reina, O. C., Zhao, F., Kim, H. L., Burhans, R. C., Drautz, D. I., Wittekindt, N. E., Tomsho, L. P., Ibarra-Laclette, E., Herrera-Estrella, L., Peacock, E., Farley, S., Sage, G. K., Rode, K., Obbard, M., Montiel, R., Bachmann, L., Ingólfsson, Ó., Aars, J., Mailund, T., Wiig, Ø., Talbot, S. L., and Lindqvist, C. (2012). Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proceedings of the National Academy of Sciences*, 109(36).

Mizumoto, H., Mitsuzuka, T., and Araki, H. (2020). An environmental DNA survey on distribution of an endangered Salmonid species, Parahucho perryi, in Hokkaido, Japan. *Frontiers in Ecology and Evolution*, 8.

Mondol, S., Bruford, M. W., and Ramakrishnan, U. (2013). Demographic loss, genetic structure and the conservation implications for Indian tigers. *Proceedings of the royal society B*, 280(1762).

Murray, G. G. R., Soares, A. E. R., Novak, B. J., Schaefer, N. K., Cahill, J. A., Baker, A. J., Demboski, J. R., Doll, A., Fonseca, R. R. D., Fulton, T. L., Gilbert, M. T. P., Heintzman, P. D., Letts, B., McIntosh, G., O'Connell, B. L., Peck, M., Pipes, M.-L., Rice, E. S., Santos, K. M., Sohrweide, A. G., Vohr, S. H., Corbett-Detig, R. B., Green, R. E., and Shapiro, B. (2017). Natural selection shaped the rise and fall of passenger pigeon genomic diversity. *Science*, 358(6365):951–954.

Nielsen, R. and Matz, M. (2006). Statistical approaches for DNA barcoding. *Systematic Biology*, 55(1):162–169.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. (2015). Reference sequence (RefSeq)

database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745.

Onorato, D. P., Hellgren, E. C., Bussche, R. A. V. D., and Doan-Crider, D. L. (2004). Phylogeographic patterns within a metapopulation of black bears (ursus. 85(1):140–147.

Onorato, D. P., Hellgren, E. C., Bussche, R. A. V. D., Doan-Crider, D. L., and Skiles, J. R. (2006). Genetic structure of American black bears in the desert southwest of North America: conservation implications for recolonization. *Conservation Genetics*, 8(3):565–576.

Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P. W., Ávila Arcos, M. C., Fu, Q., Krause, J., Willerslev, E., Stone, A. C., and Warinner, C. (2021). Ancient DNA analysis. *Nature Reviews Methods Primers*, 1(1):1–26.

Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., Johnson, P. L. F., Fumagalli, M., Vilstrup, J. T., Raghavan, M., Korneliussen, T., Malaspinas, A.-S., Vogt, J., Szklarczyk, D., Kelstrup, C. D., Vinther, J., Dolocan, A., Stenderup, J., Velazquez, A. M. V., Cahill, J., Rasmussen, M., Wang, X., Min, J., Zazula, G. D., Seguin-Orlando, A., Mortensen, C., Magnussen, K., Thompson, J. F., Weinstock, J., Gregersen, K., Røed, K. H., Eisenmann, V., Rubin, C. J., Miller, D. C., Antczak, D. F., Bertelsen, M. F., Brunak, S., Al-Rasheid, K. A. S., Ryder, O., Andersson, L., Mundy, J., Krogh, A., Gilbert, M. T. P., Kjær, K., Sicheritz-Ponten, T., Jensen, L. J., Olsen, J. V., Hofreiter, M., Nielsen, R., Shapiro, B., Wang, J., and Willerslev, E. (2013a). Recalibrating equus evolution using the genome sequence of an early middle pleistocene horse. *Nature*, 499(7456):74–78.

Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., Johnson, P. L. F., Fumagalli, M., Vilstrup, J. T., Raghavan, M., Korneliussen, T., Malaspinas, A.-S., Vogt, J., Szklarczyk, D., Kelstrup, C. D., Vinther, J., Dolocan, A., Stenderup, J., Velazquez, A. M. V., Cahill, J., Rasmussen, M., Wang, X., Min, J., Zazula, G. D., Seguin-Orlando, A., Mortensen, C., Magnussen, K., Thompson, J. F., Weinstock, J., Gregersen, K., Røed, K. H., Eisenmann, V., Rubin, C. J., Miller, D. C., Antczak, D. F., Bertelsen, M. F., Brunak, S., Al-Rasheid, K. A. S., Ryder, O., Andersson, L., Mundy, J., Krogh, A., Gilbert, M. T. P., Kjær, K., Sicheritz-Ponten, T., Jensen, L. J., Olsen, J. V., Hofreiter, M., Nielsen, R., Shapiro, B., Wang, J., and Willerslev, E. (2013b). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, 499(7456):74–78.

Ounit, R., Wanamaker, S., Close, T. J., and Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, 16(1).

Page, A. J., Taylor, B., Delaney, A. J., Soares, J., Seemann, T., Keane, J. A., and Harris, S. R. (2016). SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial Genomics*, 2(4).

Palkopoulou, E., Dalén, L., Lister, A. M., Vartanyan, S., Sablin, M., Sher, A., Edmark, V. N., Brandström, M. D., Germonpré, M., Barnes, I., and Thomas, J. A. (2013). Holarctic genetic

structure and range dynamics in the woolly mammoth. *Proceedings of the Royal Society B: Biological Sciences*, 280(1770):20131910.

Palkopoulou, E., Lipson, M., Mallick, S., Nielsen, S., Rohland, N., Baleka, S., Karpinski, E., Ivancevic, A. M., To, T.-H., Kortschak, R. D., Raison, J. M., Qu, Z., Chin, T.-J., Alt, K. W., Claesson, S., Dalén, L., MacPhee, R. D. E., Meller, H., Roca, A. L., Ryder, O. A., Heiman, D., Young, S., Breen, M., Williams, C., Aken, B. L., Ruffier, M., Karlsson, E., Johnson, J., Palma, F. D., Alfoldi, J., Adelson, D. L., Mailund, T., Munch, K., Lindblad-Toh, K., Hofreiter, M., Poinar, H., and Reich, D. (2018). A comprehensive genomic history of extinct and living elephants. *Proceedings of the National Academy of Sciences*, 115(11).

Palkopoulou, E., Mallick, S., Skoglund, P., Enk, J., Rohland, N., Li, H., Omrak, A., Vartanyan, S., Poinar, H., Götherström, A., Reich, D., and Dalén, L. (2015). Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Current Biology*, 25(10):1395–1400.

Parducci, L., Alsos, I. G., Unneberg, P., Pedersen, M. W., Han, L., Lammers, Y., Salonen, J. S., Väliranta, M. M., Slotte, T., and Wohlfarth, B. (2019). Shotgun environmental DNA, pollen, and macrofossil analysis of late glacial lake sediments from southern Sweden. *Frontiers in Ecology and Evolution*, 7.

Parducci, L., Jørgensen, T., Tollefsrud, M. M., Elverland, E., Alm, T., Fontana, S. L., Bennett, K. D., Haile, J., Matetovici, I., Suyama, Y., Edwards, M. E., Andersen, K., Rasmussen, M., Boessenkool, S., Coissac, E., Brochmann, C., Taberlet, P., Houmark-Nielsen, M., Larsen, N. K., Orlando, L., Gilbert, M. T. P., Kjær, K. H., Alsos, I. G., and Willerslev, E. (2012). Glacial survival of boreal trees in Northern Scandinavia. *Science*, 335(6072):1083–1086.

Parks, S. L. and Goldman, N. (2014). Maximum likelihood inference of small trees in the presence of long branches. *Systematic Biology*, 63(5):798–811.

Parsons, K. M., Everett, M., Dahlheim, M., and Park, L. (2018). Water, water everywhere: environmental DNA can unlock population structure in elusive marine species. *Royal Society Open Science*, 5(8):180537.

Pasquet, R. S., Peltier, A., Hufford, M. B., Oudin, E., Saulnier, J., Paul, L., Knudsen, J. T., Herren, H. R., and Gepts, P. (2008). Long-distance pollen flow assessment through evaluation of pollinator foraging range suggests transgene escape distances. *PNAS*, 105(36).

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

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, 2(12):e190.

Pečnerová, P., Palkopoulou, E., Wheat, C. W., Skoglund, P., Vartanyan, S., Tikhonov, A., Nikolskiy, P., van der Plicht, J., del Molino, D. D., and Dalén, L. (2017). Mitogenome evolution in the last

surviving woolly mammoth population reveals neutral and functional consequences of small population size. *Evolution Letters*, 1(6):292–303.

Pedersen, M. W., De Sanctis, B., Saremi, N. F., Sikora, M., Puckett, E. E., Gu, Z., Moon, K. L., Kapp, J. D., Vinner, L., Vardanyan, Z., Ardelean, C. F., Arroyo-Cabrales, J., Cahill, J. A., Heintzman, P. D., Zazula, G., MacPhee, R. D. E., Shapiro, B., Durbin, R., and Willerslev, E. (2021). Environmental genomics of Late Pleistocene black bears and giant short-faced bears. *Current Biology*, 31(12):2728–2736.e8.

Pedersen, M. W., Overballe-Petersen, S., Ermini, L., Sarkissian, C. D., Haile, J., Hellstrom, M., Spens, J., Thomsen, P. F., Bohmann, K., Cappellini, E., Schnell, I. B., Wales, N. A., Carøe, C., Campos, P. F., Schmidt, A. M. Z., Gilbert, M. T. P., Hansen, A. J., Orlando, L., and Willerslev, E. (2015). Ancient and modern environmental DNA. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1660):20130383.

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

Pelletier, A., Obbard, M., Mills, K., Howe, E., Burrows, F., White, B., and Kyle, C. (2012). Delineating genetic groupings in continuously distributed species across largely homogeneous landscapes: a study of American black bears (Ursus americanus) in Ontario, Canada. *Canadian Journal of Zoology*, 90(8):999–1014.

Peris, D., Janssen, K., Barthel, H. J., Bierbaum, G., Delclòs, X., Peñalver, E., Solórzano-Kraemer, M. M., Jordal, B. H., and Rust, J. (2020). DNA from resin-embedded organisms: Past, present and future. *PLOS ONE*, 15(9):e0239521.

Peter, B. M. (2016). Admixture, population structure, and f-statistics. *Genetics*, 202(4):1485–1501.

Poinar, H. N., Schwarz, C., Qi, J., Shapiro, B., MacPhee, R. D. E., Buigues, B., Tikhonov, A., Huson, D. H., Tomsho, L. P., Auch, A., Rampp, M., Miller, W., and Schuster, S. C. (2006). Metagenomics to paleogenomics: Large-scale sequencing of mammoth DNA. *Science*, 311(5759):392–394.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., de Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., Siebauer, M., Theunert, C., Tandon, A., Moorjani, P., Pickrell, J., Mullikin, J. C., Vohr, S. H., Green, R. E., Hellmann, I., Johnson, P. L. F., Blanche, H., Cann, H., Kitzman, J. O., Shendure, J., Eichler, E. E., Lein, E. S., Bakken, T. E., Golovanova, L. V., Doronichev, V. B., Shunkov, M. V., Derevianko, A. P., Viola, B.,

Slatkin, M., Reich, D., Kelso, J., and Pääbo, S. (2013). The complete genome sequence of a Neanderthal from the Altai mountains. *Nature*, 505(7481):43–49.

Prüfer, K., Stenzel, U., Hofreiter, M., Pääbo, S., Kelso, J., and Green, R. E. (2010). Computational challenges in the analysis of ancient DNA. *Genome Biology*, 11(5):R47.

Puckett, E. E. and Davis, I. S. (2021). Spatial patterns of genetic diversity in eight bear (ursidae) species. *Ursus*, 2021(32e20).

Puckett, E. E., Etter, P. D., Johnson, E. A., and Eggert, L. S. (2015). Phylogeographic analyses of American black bears (Ursus americanus) suggest four glacial refugia and complex patterns of postglacial admixture. *Molecular Biology and Evolution*, 32(9):2338–2350.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575.

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

Rambaut, A. (2010). Figtree v1.3.1.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in bayesian phylogenetics using tracer 1.7. *Systematic Biology*, 67(5):901–904.

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

Reilly, P. F., Tjahjadi, A., Miller, S. L., Akey, J. M., and Tucci, S. (2022). The contribution of Neanderthal introgression to modern human traits. *Current Biology*, 32(18):R970–R983.

Renaud, G., Hanghøj, K., Willerslev, E., and Orlando, L. (2017). gargammel: a sequence simulator for ancient DNA. 33(4):577–579.

Renaud, G., Schubert, M., Sawyer, S., and Orlando, L. (2019). Authentication and assessment of contamination in ancient DNA. *Ancient DNA*, pages 163–194.

Revell, L. J. (2011). phytools: an r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223.

Reynolds, M. and Klavitter, J. (2006). Translocation of wild Laysan duck Anas laysanensis to establish a population at Midway Atoll National Wildlife Refuge, United States and US Pacific Possession. *Conservation Evidence*, 3:6–8.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The european molecular biology open software suite. *Trends in Genetics*, 16(6):276–277.

Roberts, D. L., Rossman, J. S., and Jarić, I. (2021). Dating first cases of COVID-19. *PLOS Pathogens*, 17(6):e1009620.

Roca, A. L., Ishida, Y., Brandt, A. L., Benjamin, N. R., Zhao, K., and Georgiadis, N. J. (2015). Elephant natural history: A genomic perspective. *Annual Review of Animal Biosciences*, 3(1):139–167.

Rogaev, E. I., Moliaka, Y. K., Malyarchuk, B. A., Kondrashov, F. A., Derenko, M. V., Chumakov, I., and Grigorenko, A. P. (2006). Complete mitochondrial genome and phylogeny of pleistocene mammoth Mammuthus primigenius. *PLoS Biology*, 4(3):e73.

Roure, B., Baurain, D., and Philippe, H. (2012). Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Molecular Biology and Evolution*, 30(1):197–214.

Rozewicki, J., Li, S., Amada, K. M., Standley, D. M., and Katoh, K. (2019). MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Research*.

Ruppert, K. M., Kline, R. J., and Rahman, M. S. (2019). Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation*, 17:e00547.

Rybczynski, N., Gosse, J. C., Harington, C. R., Wogelius, R. A., Hidy, A. J., and Buckley, M. (2013). Mid-Pliocene warm-period deposits in the High Arctic yield insight into camel evolution. *Nature Communications*, 4(1).

Sakata, M. K., Sato, M., Sato, M. O., Watanabe, T., Mitsuishi, H., Hikitsuchi, T., Kobayashi, J., and Minamoto, T. (2022). Detection and persistence of environmental DNA (eDNA) of the different developmental stages of a vector mosquito, culex pipiens pallens. *PLOS One*, 17(8):e0272653.

Sanctis, B. D., Krukov, I., and de Koning, A. P. J. (2017). Allele age under non-classical assumptions is clarified by an exact computational markov chain approach. *Scientific Reports*, 7(1).

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

Sansom, R. S., Choate, P. G., Keating, J. N., and Randle, E. (2018). Parsimony, not bayesian analysis, recovers more stratigraphically congruent phylogenetic trees. *Biology Letters*, 14(6):20180263.

Sauquet, H. (2013). A practical guide to molecular dating. *Comptes Rendus Palevol*, 12(6):355–367.

Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T., McCarthy, S., Montgomery, S. H., Schwalie, P. C., Tang, Y. A., Ward, M. C., Xue, Y., Yngvadottir, B., Alkan, C., Andersen, L. N., Ayub,

Q., Ball, E. V., Beal, K., Bradley, B. J., Chen, Y., Clee, C. M., Fitzgerald, S., Graves, T. A., Gu, Y., Heath, P., Heger, A., Karakoc, E., Kolb-Kokocinski, A., Laird, G. K., Lunter, G., Meader, S., Mort, M., Mullikin, J. C., Munch, K., O'Connor, T. D., Phillips, A. D., Prado-Martinez, J., Rogers, A. S., Sajjadian, S., Schmidt, D., Shaw, K., Simpson, J. T., Stenson, P. D., Turner, D. J., Vigilant, L., Vilella, A. J., Whitener, W., Zhu, B., Cooper, D. N., de Jong, P., Dermitzakis, E. T., Eichler, E. E., Flicek, P., Goldman, N., Mundy, N. I., Ning, Z., Odom, D. T., Ponting, C. P., Quail, M. A., Ryder, O. A., Searle, S. M., Warren, W. C., Wilson, R. K., Schierup, M. H., Rogers, J., Tyler-Smith, C., and Durbin, R. (2012). Insights into hominid evolution from the gorilla genome sequence. *Nature*, 483(7388):169–175.

Schubert, B. W., Hulbert, R. C., Macfadden, B. J., Searle, M., and Searle, S. (2010). Giant short-faced bears (Arctodus simus) in Pleistocene Florida USA, a substantial range extension. *Journal of Paleontology*, 84(1):79–87.

Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J. F., AL-Rasheid, K. A., Willerslev, E., Krogh, A., and Orlando, L. (2012). Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics*, 13(1):178.

Seeber, P., von, H. B., Kauserud, H., Löber, U., Stoof-Leichsenring, K., Herzschuh, U., and Epp, L. (2021). Fungal biodiversity in Arctic paleoecosystems assessed by metabarcoding of lake sedimentary ancient DNA. Technical report.

Shapiro, B., Ho, S. Y. W., Drummond, A. J., Suchard, M. A., Pybus, O. G., and Rambaut, A. (2010). A bayesian phylogenetic method to estimate unknown sequence ages. *Molecular Biology and Evolution*, 28(2):879–887.

Sidow, A., Wilson, A. C., Paabo, S., Hummel, S., Bada, J., Westbroek, P., Hagelberg, E., and Curry, G. (1991). Bacterial DNA in Clarkia fossils. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 333(1268):429–433.

Sigsgaard, E. E., Jensen, M. R., Winkelmann, I. E., Møller, P. R., Hansen, M. M., and Thomsen, P. F. (2020). Population-level inferences from environmental DNA—Current status and future perspectives. *Evolutionary Applications*, 13(2):245–262.

Skoglund, P., Sjödin, P., Skoglund, T., Lascoux, M., and Jakobsson, M. (2014). Investigating population history using temporal genetic differentiation. *Molecular Biology and Evolution*, 31(9):2516–2527.

Slatkin, M. and Racimo, F. (2016). Ancient DNA and human history. *Proceedings of the National Academy of Sciences*, 113(23):6380–6387.

Slon, V., Hopfe, C., Weiß, C. L., Mafessoni, F., Rasilla, M. d. l., Lalueza-Fox, C., Rosas, A., Soressi, M., Knul, M. V., Miller, R., Stewart, J. R., Derevianko, A. P., Jacobs, Z., Li, B., Roberts, R. G., Shunkov, M. V., Lumley, H. d., Perrenoud, C., Gušić, I., Kućan, □., Rudan, P., Aximu-Petri, A., Essel, E., Nagel, S., Nickel, B., Schmidt, A., Prüfer, K., Kelso, J., Burbano, H. A., Pääbo, S.,

and Meyer, M. (2017). Neandertal and Denisovan DNA from Pleistocene sediments. *Science*, 356(6338):605–608.

Slon, V., Mafessoni, F., Vernot, B., de Filippo, C., Grote, S., Viola, B., Hajdinjak, M., Peyrégne, S., Nagel, S., Brown, S., Douka, K., Higham, T., Kozlikin, M. B., Shunkov, M. V., Derevianko, A. P., Kelso, J., Meyer, M., Prüfer, K., and Pääbo, S. (2018). The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature*, 561(7721):113–116.

Speidel, L., Forest, M., Shi, S., and Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321–1329.

Srivastava, A., Sarsani, V. K., Fiddes, I., Sheehan, S. M., Seger, R. L., Barter, M. E., Neptune-Bear, S., Lindqvist, C., and Korstanje, R. (2018). Genome assembly and gene expression in the American black bear provides new insights into the renal response to hibernation. *DNA Research*, 26(1):37–44.

Stuart, A. J. (2014). Late quaternary megafaunal extinctions on the continents: a short review. *Geological Journal*, 50(3):338–363.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1).

Talas, L., Stivrins, N., Veski, S., Tedersoo, L., and Kisand, V. (2021). Sedimentary ancient DNA (sedaDNA) reveals fungal diversity and environmental drivers of community changes throughout the Holocene in the present Boreal lake Lielais Svētiņu (eastern latvia). *Microorganisms*, 9(4):719.

Thakur, I. S. and Roy, D. (2020). Environmental DNA and RNA as records of human exposome, including biotic/abiotic exposures and its implications in the assessment of the role of environment in chronic diseases. *International Journal of Molecular Sciences*, 21(14):4879.

van der Valk, T., Pečnerová, P., del Molino, D. D., Bergström, A., Oppenheimer, J., Hartmann, S., Xenikoudakis, G., Thomas, J. A., Dehasque, M., Sağlıcan, E., Fidan, F. R., Barnes, I., Liu, S., Somel, M., Heintzman, P. D., Nikolskiy, P., Shapiro, B., Skoglund, P., Hofreiter, M., Lister, A. M., Götherström, A., and Dalén, L. (2021). Million-year-old DNA sheds light on the genomic history of mammoths. *Nature*, 591(7849):265–269.

Vandeventer, P. E., Mejia, J., Nadim, A., Johal, M. S., and Niemz, A. (2013). DNA adsorption to and elution from silica surfaces: Influence of amino acid buffers. *The Journal of Physical Chemistry B*, 117(37):10742–10749.

Varas-Nelson, A. C. (2011). *Conservation Genetics of Black Bears in Arizona and Northern Mexico*. PhD thesis, The University of Arizona.

von Hippel, B., Stoof-Leichsenring, K. R., Schulte, L., Seeber, P., Epp, L. S., Biskaborn, B. K., Diekmann, B., Melles, M., Pestryakova, L., and Herzschuh, U. (2022). Long-term fungus–plant covariation from multi-site sedimentary ancient DNA metabarcoding. *Quaternary Science Reviews*, 295:107758.

Wagner, S., Lagane, F., Seguin-Orlando, A., Schubert, M., Leroy, T., Guichoux, E., Chancerel, E., Bech-Hebelstrup, I., Bernard, V., Billard, C., Billaud, Y., Bolliger, M., Croutsch, C., Čufar, K., Eynaud, F., Heussner, K. U., Köninger, J., Langenegger, F., Leroy, F., Lima, C., Martinelli, N., Momber, G., Billamboz, A., Nelle, O., Palomo, A., Piqué, R., Ramstein, M., Schweichel, R., Stäuble, H., Tegel, W., Terradas, X., Verdin, F., Plomion, C., Kremer, A., and Orlando, L. (2018). High-Throughput DNA sequencing of ancient wood. *Molecular Ecology*, 27(5):1138–1154.

Wang, X., Rybczynski, N., Harington, C. R., White, S. C., and Tedford, R. H. (2017). A basal ursine bear (Protarctos abstrusus) from the Pliocene High Arctic reveals Eurasian affinities and a diet rich in fermentable sugars. *Scientific Reports*, 7(1).

Wang, Y., Korneliussen, T. S., Holman, L. E., Manica, A., and Pedersen, M. W. (2022). ngslca: A toolkit for fast and flexible lowest common ancestor inference and taxonomic profiling of metagenomic data. *Methods in Ecology and Evolution*.

Wang, Y., Pedersen, M. W., Alsos, I. G., De Sanctis, B., Racimo, F., Prohaska, A., Coissac, E., Owens, H. L., Merkel, M. K. F., Fernandez-Guerra, A., Rouillard, A., Lammers, Y., Alberti, A., Denoeud, F., Money, D., Ruter, A. H., McColl, H., Larsen, N. K., Cherezova, A. A., Edwards, M. E., Fedorov, G. B., Haile, J., Orlando, L., Vinner, L., Korneliussen, T. S., Beilman, D. W., Bjørk, A. A., Cao, J., Dockter, C., Esdale, J., Gusarova, G., Kjeldsen, K. K., Mangerud, J., Rasic, J. T., Skadhauge, B., Svendsen, J. I., Tikhonov, A., Wincker, P., Xing, Y., Zhang, Y., Froese, D. G., Rahbek, C., Bravo, D. N., Holden, P. B., Edwards, N. R., Durbin, R., Meltzer, D. J., Kjær, K. H., Möller, P., and Willerslev, E. (2021). Late Quaternary dynamics of Arctic biota from ancient environmental genomics. *Nature*, 600(7887):86–92.

Warinner, C., Herbig, A., Mann, A., Fellows Yates, J. A., Weiß, C. L., Burbano, H. A., Orlando, L., and Krause, J. (2017). A robust framework for microbial archaeology. *Annual Review of Genomics and Human Genetics*, 18(1):321–356.

Wilcox, T. M., Zarn, K. E., Piggott, M. P., Young, M. K., McKelvey, K. S., and Schwartz, M. K. (2018). Capture enrichment of aquatic environmental DNA: A first proof of concept. *Molecular Ecology Resources*, 18:1392–1401.

Willerslev, E., Davison, J., Moora, M., Zobel, M., Coissac, E., Edwards, M. E., Lorenzen, E. D., Vestergård, M., Gussarova, G., Haile, J., Craine, J., Gielly, L., Boessenkool, S., Epp, L. S., Pearman, P. B., Cheddadi, R., Murray, D., Bråthen, K. A., Yoccoz, N., Binney, H., Cruaud, C., Wincker, P., Goslar, T., Alsos, I. G., Bellemain, E., Brysting, A. K., Elven, R., Sønstebø, J. H., Murton, J., Sher, A., Rasmussen, M., Rønn, R., Mourier, T., Cooper, A., Austin, J., Möller, P., Froese, D., Zazula, G., Pompanon, F., Rioux, D., Niderkorn, V., Tikhonov, A., Savvinov, G., Roberts, R. G., MacPhee, R. D. E., Gilbert, M. T. P., Kjær, K. H., Orlando, L., Brochmann, C.,

and Taberlet, P. (2014). Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature*, 506(7486):47–51.

Willerslev, E., Hansen, A. J., Binladen, J., Brand, T. B., Gilbert, M. T. P., Shapiro, B., Bunce, M., Wiuf, C., Gilichinsky, D. A., and Cooper, A. (2003). Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science*, 300(5620):791–795.

Williams, J. W., Grimm, E. C., Blois, J. L., Charles, D. F., Davis, E. B., Goring, S. J., Graham, R. W., Smith, A. J., Anderson, M., Arroyo-Cabrales, J., Ashworth, A. C., Betancourt, J. L., Bills, B. W., Booth, R. K., Buckland, P. I., Curry, B. B., Giesecke, T., Jackson, S. T., Latorre, C., Nichols, J., Purdum, T., Roth, R. E., Stryker, M., and Takahara, H. (2018). The Neotoma Paleoecology database, a multiproxy, international, community-curated data resource. *Quaternary Research*, 89(1):156–177.

Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20.

Woodward, S. R., Weyand, N. J., and Bunnell, M. (1994). DNA sequence from Cretaceous period bone fragments. *Science*, 266(5188):1229–1232.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, 16(2):97–159.

Yang, C., Chowdhury, D., Zhang, Z., Cheung, W. K., Lu, A., Bian, Z., and Zhang, L. (2021). A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Computational and Structural Biotechnology Journal*, 19:6301–6314.

Yang, X.-Y., Wang, Z.-F., Luo, W.-C., Guo, X.-Y., Zhang, C.-H., Liu, J.-Q., and Ren, G.-P. (2019). Plastomes of betulaceae and phylogenetic implications. *Journal of Systematics and Evolution*, 57(5):508–518.

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591.

Yi, X. and Latch, E. K. (2021). Nonrandom missing data can bias principal component analysis inference of population genetic structure. *Molecular Ecology Resources*, 22(2):602–611.

Zavala, E. I., Jacobs, Z., Vernot, B., Shunkov, M. V., Kozlikin, M. B., Derevianko, A. P., Essel, E., de Fillipo, C., Nagel, S., Richter, J., Romagné, F., Schmidt, A., Li, B., O'Gorman, K., Slon, V., Kelso, J., Pääbo, S., Roberts, R. G., and Meyer, M. (2021). Pleistocene sediment DNA reveals hominin and faunal turnovers at Denisova cave. *Nature*, 595(7867):399–403.

Zeberg, H. and Pääbo, S. (2020). The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature*, 587(7835):610–612.

Zhang, J., Mamlouk, A. M., Martinetz, T., Chang, S., Wang, J., and Hilgenfeld, R. (2011). PhyloMap: an algorithm for visualizing relationships of large sequence data sets and its application to the influenza a virus genome. *BMC Bioinformatics*, 12(1).

Zhang, L., Xi, Z., Wang, M., Guo, X., and Ma, T. (2018). Plastome phylogeny and lineage diversification of Salicaceae with focus on poplars and willows. *Ecology and Evolution*, 8(16):7817–7823.

Zhou, J., Zhang, S., Wang, J., Shen, H., Ai, B., Gao, W., Zhang, C., Fei, Q., Yuan, D., Wu, Z., Tembrock, L. R., Li, S., Gu, C., and Liao, X. (2021). Chloroplast genomes in Populus (Salicaceae): comparisons from an intensively sampled genus reveal dynamic patterns of evolution. *Scientific Reports*, 11(1):9471.

Zhou, Z., Luhmann, N., Alikhan, N.-F., Quince, C., and Achtman, M. (2018). Accurate reconstruction of microbial strains from metagenomic sequencing using representative reference genomes. In *Raphael B. Research in Computational Molecular Biology. RECOMB 2018. Lecture Notes in Computer Science, vol 10812*, pages 225–240.

Metadata for permafrost samples

Table 4: Permafrost aeDNA sample metadata

Lab ID	Age	Site	Region	Latitude	Longitude
tm1 1	18580.00769	BBR9	Central Siberia	73.6481167	102.0177833
tm1 2	17819.74324	BBR9	Central Siberia	73.6481167	102.0177833
tm1 3	16415.69438	BBR9	Central Siberia	73.6481167	102.0177833
tm1 4	16904.1216	BBR9	Central Siberia	73.6481167	102.0177833
tm1 5	16710.39937	BBR9	Central Siberia	73.6481167	102.0177833
tm2 1	16710.39937	BBR9	Central Siberia	73.6481167	102.0177833
tm2 2	16935.11297	BBR9	Central Siberia	73.6481167	102.0177833
tm2 3	16935.11297	BBR9	Central Siberia	73.6481167	102.0177833
tm1 6	16291.255	BBR9	Central Siberia	73.6481167	102.0177833
tm3 1	15808.06817	BBR9	Central Siberia	73.6481167	102.0177833
tm3 2	14609.91573	BBR9	Central Siberia	73.6481167	102.0177833
tm1 7	14296.73693	BBR9	Central Siberia	73.6481167	102.0177833
tm3 3	14296.73693	BBR9	Central Siberia	73.6481167	102.0177833
tm3 4	10529.03091	BBR9	Central Siberia	73.6481167	102.0177833
tm3 5	10529.03091	BBR9	Central Siberia	73.6481167	102.0177833
tm2 4	8587.052	BBR9	Central Siberia	73.6481167	102.0177833
tm2 5	8587.052	BBR9	Central Siberia	73.6481167	102.0177833
tm2 6	7160.329998	BBR9	Central Siberia	73.6481167	102.0177833
tm2 7	7160.329998	BBR9	Central Siberia	73.6481167	102.0177833
tm8 1	modern	BBR9	Central Siberia	73.6481167	102.0177833
tm8 6	modern	BBR9	Central Siberia	73.6481167	102.0177833
tm4 1	7934.855088	BBR7	Central Siberia	73.5168	101.0088667
tm4 2	7854.574	BBR7	Central Siberia	73.5168	101.0088667
tm4 3	7787.983	BBR7	Central Siberia	73.5168	101.0088667
tm4 4	7755.901	BBR7	Central Siberia	73.5168	101.0088667
tm4 5	7555.689804	BBR7	Central Siberia	73.5168	101.0088667
tm4 6	7584.965893	BBR7	Central Siberia	73.5168	101.0088667
tm4 7	7404.011668	BBR7	Central Siberia	73.5168	101.0088667
tm8 4	modern	BBR7	Central Siberia	73.5168	101.0088667

Lab ID	Age	Site	Region	Latitude	Longitude
tm7 14	modern	BBR7	Central Siberia	73.5168	101.0088667
tm7 15	modern	BBR7	Central Siberia	73.5168	101.0088667
tm4 8	3931.113702	LUR10	Central Siberia	73.15645	93.40715
tm4 9	4700	LUR10	Central Siberia	73.15645	93.40715
tm4 10	5200	LUR10	Central Siberia	73.15645	93.40715
tm8 3	modern	LUR10	Central Siberia	73.15645	93.40715
tm4 11	18316.46175	BBS5	Central Siberia	73.65275	102.1207
ar1 1	44980	BBS5	Central Siberia	73.65275	102.1207
cr2 1	44203.24697	BBS5	Central Siberia	73.65275	102.1207
tm4 12	49970	BBS5	Central Siberia	73.65275	102.1207
ar2 10	56810	BBS5	Central Siberia	73.65275	102.1207
ar1 2	44982	BBS5	Central Siberia	73.65275	102.1207
ar1 3	49970	BBS5	Central Siberia	73.65275	102.1207
tm7 4	41546.14679	BBS5	Central Siberia	73.65275	102.1207
ar1 4	56810	BBS5	Central Siberia	73.65275	102.1207
cr2 2	2153.332511	BBS6	Central Siberia	73.698889	102.196944
cr2 3	31314.01658	BBS6	Central Siberia	73.698889	102.196944
cr2 4	42549.31598	BBS6	Central Siberia	73.698889	102.196944
tm7 8	contaminated	BBR1	Central Siberia	72.5397333	100.4312667
tm7 9	contaminated	BBR1	Central Siberia	72.5397333	100.4312667
tm7 10	contaminated	BBR1	Central Siberia	72.5397333	100.4312667
tm7 11	contaminated	BBR1	Central Siberia	72.5397333	100.4312667
tm7 12	contaminated	BBR1	Central Siberia	72.5397333	100.4312667
tm7 13	contaminated	BBR1	Central Siberia	72.5397333	100.4312667
tm8 5	contaminated	BBR1	Central Siberia	72.5397333	100.4312667
ar1 5	contaminated	BBR1	Central Siberia	72.5397333	100.4312667
tm8 2	contaminated	BBR1	Central Siberia	72.5397333	100.4312667
tm4 13	50000	BBR6	Central Siberia	73.5261667	101.0085
tm4 14	17503.88481	BBR10	Central Siberia	73.64811	102.1078
tm4 15	15315.45772	BBR10	Central Siberia	73.64811	102.1078
tm5 1	45296.59	LoR3D	Central Siberia	73.3504167	96.9746333
tm5 2	42716.335	LoR3D	Central Siberia	73.3504167	96.9746333

Lab ID	Age	Site	Region	Latitude	Longitude
tm5 3	42868.855	LoR3D	Central Siberia	73.3504167	96.9746333
tm5 6	44201.23968	LoR3D	Central Siberia	73.3504167	96.9746333
tm5 4	50000	LoR3D	Central Siberia	73.3504167	96.9746333
ar1 10	13316.51511	FI	Central Siberia	74.6225	100.828056
tm6 6	12754.03189	FI	Central Siberia	74.6225	100.828056
ar1 7	12742.7942	FI	Central Siberia	74.6225	100.828056
ar1 6	12557.48106	FI	Central Siberia	74.6225	100.828056
ar1 9	12497.90394	FI	Central Siberia	74.6225	100.828056
tm6 1	21761.46634	FI	Central Siberia	74.6225	100.828056
tm6 7	12328.21792	FI	Central Siberia	74.6225	100.828056
ar1 11	27565.8326	UTRD4	Central Siberia	74.2664	99.8264
ar1 12	27229.12334	UTRD4	Central Siberia	74.2664	99.8264
ar1 13	26911.17558	UTRD4	Central Siberia	74.2664	99.8264
ar1 14	26500.43973	UTRD4	Central Siberia	74.2664	99.8264
ar1 15	26421.31839	UTRD4	Central Siberia	74.2664	99.8264
ar1 17	26217.88252	UTRD4	Central Siberia	74.2664	99.8264
tm6 2	21329.07041	UTRD4	Central Siberia	74.2664	99.8264
ar1 18	21272.60704	UTRD4	Central Siberia	74.2664	99.8264
tm6 3	20902.26849	UTRD4	Central Siberia	74.2664	99.8264
ar1 19	20741.86655	UTRD4	Central Siberia	74.2664	99.8264
ar1 20	20441.41092	UTRD4	Central Siberia	74.2664	99.8264
ar1 21	4691.164348	TLH1	Central Siberia	74.64083333	100.7311111
tm6 11	5474.329587	TLH1	Central Siberia	74.64083333	100.7311111
ar1 22	4851.769057	TLH1	Central Siberia	74.64083333	100.7311111
ar1 23	5487.828494	TLH1	Central Siberia	74.64083333	100.7311111
tm6 9	7078.371351	TLH1	Central Siberia	74.64083333	100.7311111
tm8 7	3320	KS1	Central Siberia	72.09666667	102.3280556
cr8 1	3785.246	KS1	Central Siberia	72.09666667	102.3280556
cr8 2	4235.697	KS1	Central Siberia	72.09666667	102.3280556
cr8 3	4986.433	KS1	Central Siberia	72.09666667	102.3280556
tm6 8	6700	KS1	Central Siberia	72.09666667	102.3280556
ar6 19	contaminated	KS2	Central Siberia	72.08861111	102.2872222

Lab ID	Age	Site	Region	Latitude	Longitude
ar6 18	1220	KS2	Central Siberia	72.08861111	102.2872222
ar1 29	33967.82032	BAP	Central Siberia	74.49361111	101.2761111
ar1 30	32382.61115	BAP	Central Siberia	74.49361111	101.2761111
cr2 5	30904.05511	BAP	Central Siberia	74.49361111	101.2761111
ar1 31	30378.94041	BAP	Central Siberia	74.49361111	101.2761111
ar2 1	38133.73431	BAP	Central Siberia	74.49361111	101.2761111
ar2 2	33766.97751	BAP	Central Siberia	74.49361111	101.2761111
ar1 25	28487.92305	BAP	Central Siberia	74.49361111	101.2761111
ar1 26	28607.20444	BAP	Central Siberia	74.49361111	101.2761111
ar1 27	31452.55615	BAP	Central Siberia	74.49361111	101.2761111
ar1 28	27506.73663	BAP	Central Siberia	74.49361111	101.2761111
ar2 3	24257.6038	BAP	Central Siberia	74.49361111	101.2761111
ar2 4	23875.5164	BAP	Central Siberia	74.49361111	101.2761111
ar2 5	24103.69552	BAP	Central Siberia	74.49361111	101.2761111
ar2 6	32125.56106	BAP	Central Siberia	74.49361111	101.2761111
ar2 7	30281.70217	BAP	Central Siberia	74.49361111	101.2761111
ar2 9	23021.94737	BAP	Central Siberia	74.49361111	101.2761111
cr2 6	47589.19527	OVR	Central Siberia	74.1464	100.1264
cr2 7	48364.03846	OVR	Central Siberia	74.1464	100.1264
cr2 9	45382.90548	OVR	Central Siberia	74.1464	100.1264
cr2 10	46266.46808	OVR	Central Siberia	74.1464	100.1264
cr2 11	45736.55589	OVR	Central Siberia	74.1464	100.1264
cr2 12	35907.00031	CS1	Central Siberia	74.54766667	100.5358333
ar4 1	5339.01384	BK1	NE Siberia	71.90617	132.78635
tm9 1	10518.64428	BK1	NE Siberia	71.90617	132.78635
tm8 9	5459.366061	BK1	NE Siberia	71.90617	132.78635
tm9 2	15282.58733	BK1	NE Siberia	71.90617	132.78635
cr2 13	7878.84	BK1	NE Siberia	71.90617	132.78635
cr2 14	11015.73	BK1	NE Siberia	71.90617	132.78635
cr2 15	8809.476	BK1	NE Siberia	71.90617	132.78635
ar4 2	10518.64428	BK1	NE Siberia	71.90617	132.78635
cr2 17	15282.58733	BK1	NE Siberia	71.90617	132.78635

Lab ID	Age	Site	Region	Latitude	Longitude
cr2 18	5459.366061	BK1	NE Siberia	71.90617	132.78635
cr2 19	5339.01384	BK1	NE Siberia	71.90617	132.78635
tm9 3	547.0201465	BK2	NE Siberia	72.002778	132.833611
tm9 4	8174.969785	BK2	NE Siberia	72.002778	132.833611
tm9 5	11158.46873	BK2	NE Siberia	72.002778	132.833611
tm9 6	10548.84783	BK2	NE Siberia	72.002778	132.833611
cr2 20	8174.969785	BK2	NE Siberia	72.002778	132.833611
cr2 21	50000	BK2	NE Siberia	72.002778	132.833611
cr2 22	11158.46873	BK2	NE Siberia	72.002778	132.833611
cr2 23	10548.84783	BK2	NE Siberia	72.002778	132.833611
cr2 25	6740.825941	BK3	NE Siberia	71.9056	132.7853
ar2 11	6740.825941	BK3	NE Siberia	71.9056	132.7853
cr2 26	7332.410758	BK3	NE Siberia	71.9056	132.7853
cr2 27	7564.712031	BK3	NE Siberia	71.9056	132.7853
tm9 7	7564.712031	BK3	NE Siberia	71.9056	132.7853
tm9 9	7668.134304	BK3	NE Siberia	71.9056	132.7853
cr2 28	51940.149	YUB	NW Siberia	60.600889	71.926278
cr2 29	47180.99561	YUB	NW Siberia	60.600889	71.926278
cr2 30	35402.989	YUB	NW Siberia	60.600889	71.926278
cr2 31	24464.38648	YUB	NW Siberia	60.600889	71.926278
cr3 1	22043.98	YUB	NW Siberia	60.600889	71.926278
cr3 2	18245.06728	YUB	NW Siberia	60.600889	71.926278
cr3 3	14204.928	YUB	NW Siberia	60.600889	71.926278
cr3 4	11041.344	YUB	NW Siberia	60.600889	71.926278
cr3 5	8827.953	YUB	NW Siberia	60.600889	71.926278
ar6 1	31506.323	Mar-01	NW Siberia	68.655667	71.922528
ar6 2	26647.48877	Mar-01	NW Siberia	68.655667	71.922528
ar6 3	28718.307	Mar-01	NW Siberia	68.655667	71.922528
ar6 4	27396.81661	Mar-01	NW Siberia	68.655667	71.922528
ar6 5	26062.888	Mar-01	NW Siberia	68.655667	71.922528
ar6 6	23755.27461	Mar-01	NW Siberia	68.655667	71.922528
ar6 7	41886.33961	Mar-02	NW Siberia	68.655667	71.922528

Lab ID	Age	Site	Region	Latitude	Longitude
ar6 9	27689.92542	Mar-02	NW Siberia	68.655667	71.922528
ar6 10	33955.12	Mar-02	NW Siberia	68.655667	71.922528
ar6 11	30811.89951	Mar-02	NW Siberia	68.655667	71.922528
ar6 12	29726.45159	Mar-02	NW Siberia	68.655667	71.922528
cr9 1	3920	DO	NE Siberia	71.866667	127.066667
cr9 4	4893.547	DO	NE Siberia	71.866667	127.066667
cr9 5	5618.547	DO	NE Siberia	71.866667	127.066667
cr9 6	6332.905	DO	NE Siberia	71.866667	127.066667
cr9 7	7029.332	DO	NE Siberia	71.866667	127.066667
cr9 8	7772.886	DO	NE Siberia	71.866667	127.066667
cr9 2	8490	DO	NE Siberia	71.866667	127.066667
cr9 3	10367.169	DO	NE Siberia	71.866667	127.066667
cr9 9	10500	DO	NE Siberia	71.866667	127.066667
cr9 10	10616.76	DO	NE Siberia	71.866667	127.066667
cr1 26	3526.654	LT	Central Siberia	79.2452778	101.8152778
cr1 27	4106.081	LT	Central Siberia	79.2452778	101.8152778
cr1 28	5098.266	LT	Central Siberia	79.2452778	101.8152778
cr1 29	5889.311	LT	Central Siberia	79.2452778	101.8152778
cr1 30	7492.559	LT	Central Siberia	79.2452778	101.8152778
cr1 31	8489.499	LT	Central Siberia	79.2452778	101.8152778
cr1 1	9441.406	LT	Central Siberia	79.2452778	101.8152778
cr1 2	10913.35532	LT	Central Siberia	79.2452778	101.8152778
cr1 3	15818.585	LT	Central Siberia	79.2452778	101.8152778
cr1 4	19344.811	LT	Central Siberia	79.2452778	101.8152778
cr1 5	24211.25132	LT	Central Siberia	79.2452778	101.8152778
cr3 11	50000	MK2	NW Siberia	69.7396944	84.8181111
cr3 12	50000	MK2	NW Siberia	69.7396944	84.8181111
cr3 13	25870.33	IH4	NW Siberia	66.758037	86.680415
cr3 14	41425.95	PO2	NW Siberia	66.726667	86.639134
cr8 6	45361.1	PO2	NW Siberia	66.726667	86.639134
cr3 15	47000	PO2	NW Siberia	66.726667	86.639134
cr3 17	41749.49	PO2	NW Siberia	66.726667	86.639134

Lab ID	Age	Site	Region	Latitude	Longitude
cr3 6	34000	ZAS	North Europe	58.15	56.9333333
cr3 7	50000	ZAS	North Europe	58.15	56.9333333
cr3 9	37500	ZAS	North Europe	58.15	56.9333333
cr3 10	50000	ZAS	North Europe	58.15	56.9333333
cr3 18	54000	PO1	NW Siberia	66.8719	86.6269
cr8 7	32000	PO1	NW Siberia	66.8719	86.6269
cr3 19	37178.32185	DY	NE Siberia	68.66667	159.08333
cr3 20	47457.38265	DY	NE Siberia	68.66667	159.08333
cr3 21	46907.76269	DY	NE Siberia	68.66667	159.08333
cr6 7	47819.82824	DY	NE Siberia	68.66667	159.08333
cr3 22	45296.59383	DY	NE Siberia	68.66667	159.08333
cr3 23	46907.76269	DY	NE Siberia	68.66667	159.08333
cr6 9	46907.76269	DY	NE Siberia	68.66667	159.08333
cr3 25	35875.85049	DY	NE Siberia	68.66667	159.08333
cr3 26	37151.73181	DY	NE Siberia	68.66667	159.08333
cr3 27	38284.08745	DY	NE Siberia	68.66667	159.08333
cr3 28	37751.25251	DY	NE Siberia	68.66667	159.08333
cr3 29	34570.68346	DY	NE Siberia	68.66667	159.08333
cr6 10	35941.58173	DY	NE Siberia	68.66667	159.08333
cr6 11	35450.63995	DY	NE Siberia	68.66667	159.08333
cr6 12	35145.845	DY	NE Siberia	68.66667	159.08333
cr3 30	34927.5699	DY	NE Siberia	68.66667	159.08333
cr6 13	38075.93343	DY	NE Siberia	68.66667	159.08333
ar4 9	23812.04131	DY	NE Siberia	68.66667	159.08333
ar4 10	24462.02564	DY	NE Siberia	68.66667	159.08333
ar4 11	24887.16889	DY	NE Siberia	68.66667	159.08333
ar4 12	25715.17225	DY	NE Siberia	68.66667	159.08333
ar4 13	25873.47552	DY	NE Siberia	68.66667	159.08333
ar4 14	27228.40683	DY	NE Siberia	68.66667	159.08333
ar4 15	27690.36188	DY	NE Siberia	68.66667	159.08333
ar4 17	27751.3355	DY	NE Siberia	68.66667	159.08333
ar4 18	28458.47363	DY	NE Siberia	68.66667	159.08333

Lab ID	Age	Site	Region	Latitude	Longitude
ar4 19	29089.58136	DY	NE Siberia	68.66667	159.08333
ar4 20	29416.02084	DY	NE Siberia	68.66667	159.08333
ar4 21	30950.983	DY	NE Siberia	68.66667	159.08333
ar4 22	20323.97348	DY	NE Siberia	68.66667	159.08333
ar4 23	21555.88984	DY	NE Siberia	68.66667	159.08333
ar4 25	120.4234828	DY	NE Siberia	68.66667	159.08333
ar4 26	134.2790122	DY	NE Siberia	68.66667	159.08333
ar4 27	142.0864024	DY	NE Siberia	68.66667	159.08333
ar4 28	746.3078758	DY	NE Siberia	68.66667	159.08333
ar4 29	556.2402409	DY	NE Siberia	68.66667	159.08333
cr3 31	36832.53389	DY	NE Siberia	68.66667	159.08333
cr4 30	35313.64903	DY	NE Siberia	68.66667	159.08333
cr4 31	37642.01878	DY	NE Siberia	68.66667	159.08333
cr6 14	38365.61307	DY	NE Siberia	68.66667	159.08333
cr5 21	41247.59039	MR1	NE Siberia	64.283333	171.25
cr6 2	42316.14131	MR1	NE Siberia	64.283333	171.25
cr6 3	44227.20632	MR1	NE Siberia	64.283333	171.25
cr6 4	37830.24412	MR1	NE Siberia	64.283333	171.25
cr5 22	43399.68193	MR2	NE Siberia	64.283333	171.25
cr5 23	35245.55779	MR2	NE Siberia	64.283333	171.25
cr5 25	34599.01518	MR2	NE Siberia	64.283333	171.25
cr5 26	34777.70077	MR2	NE Siberia	64.283333	171.25
cr5 27	33853.61073	MR2	NE Siberia	64.283333	171.25
cr5 28	37919.66326	MR2	NE Siberia	64.283333	171.25
ar5 2	27866.51123	MR2	NE Siberia	64.283333	171.25
ar5 3	28293.59517	MR2	NE Siberia	64.283333	171.25
cr5 29	32128.60664	MR2	NE Siberia	64.283333	171.25
ar5 4	31990.84878	MR2	NE Siberia	64.283333	171.25
ar5 5	29510.72962	MR2	NE Siberia	64.283333	171.25
cr5 30	30797.54083	MR2	NE Siberia	64.283333	171.25
ar5 6	31178.91593	MR2	NE Siberia	64.283333	171.25
ar5 7	27069.26392	MR2	NE Siberia	64.283333	171.25

Lab ID	Age	Site	Region	Latitude	Longitude
ar5 9	26055.15078	MR2	NE Siberia	64.283333	171.25
ar5 10	25804.01505	MR2	NE Siberia	64.283333	171.25
ar5 11	25400.9824	MR2	NE Siberia	64.283333	171.25
ar5 12	26805.87931	MR2	NE Siberia	64.283333	171.25
ar5 13	32698.79286	MR2	NE Siberia	64.283333	171.25
ar5 14	23880.66114	MR2	NE Siberia	64.283333	171.25
ar4 30	23838.18706	MR3	NE Siberia	64.283333	171.25
ar4 31	24667.04648	MR3	NE Siberia	64.283333	171.25
ar5 1	24539.22387	MR3	NE Siberia	64.283333	171.25
ar5 15	24941.42558	MR3	NE Siberia	64.283333	171.25
ar5 31	24958.93246	MR3	NE Siberia	64.283333	171.25
ar5 17	24187.9181	MR3	NE Siberia	64.283333	171.25
cr5 31	24077.19738	MR3	NE Siberia	64.283333	171.25
ar5 18	25218.07806	MR3	NE Siberia	64.283333	171.25
cr5 20	36448.34256	MR4	NE Siberia	64.283333	171.25
cr6 1	44443.4602	MR5	NE Siberia	64.283333	171.25
ar5 19	20115.1609	MR6	NE Siberia	64.283333	171.25
ar5 20	19054.37987	MR6	NE Siberia	64.283333	171.25
tm7 6	8054.121596	AC	NE Siberia	64.735176	177.30732
ar2 12	10267.07189	AC	NE Siberia	64.735176	177.30732
ar2 13	16890.68591	AC	NE Siberia	64.735176	177.30732
tm9 12	10355.41751	AC	NE Siberia	64.735176	177.30732
ar4 3	modern	AC	NE Siberia	64.735176	177.30732
cr5 2	22836.68	CAB	NE Siberia	71.666667	129.5
cr5 4	22836.68	CAB	NE Siberia	71.666667	129.5
cr5 11	22836.68	CAB	NE Siberia	71.666667	129.5
cr5 3	12305.07	KK	NE Siberia	69.383333	158.466667
cr5 6	12305.07	KK	NE Siberia	69.383333	158.466667
cr5 9	12305.07	KK	NE Siberia	69.383333	158.466667
cr5 10	12305.07	KK	NE Siberia	69.383333	158.466667
cr5 13	12305.07	KK	NE Siberia	69.383333	158.466667
cr5 17	12305.07	KK	NE Siberia	69.383333	158.466667

Lab ID	Age	Site	Region	Latitude	Longitude
cr5 7	25000	CHR	NE Siberia	69.483333	156.983333
cr5 14	25000	CHR	NE Siberia	69.483333	156.983333
cr5 18	25000	CHR	NE Siberia	69.483333	156.983333
cr5 19	25000	CHR	NE Siberia	69.483333	156.983333
cr5 5	modern	PJ	NE Siberia	68.666667	160.833333
cr5 1	50000	CAS	North Europe	68.147817	39.758698
cr5 12	50000	CAS	North Europe	68.147817	39.758698
cr5 15	50000	CAS	North Europe	68.147817	39.758698
cr8 27	6080.666009	PP	NE Siberia	68.499291	162.4068
cr8 28	7512.219	PP	NE Siberia	68.499291	162.4068
cr8 29	9520.931554	PP	NE Siberia	68.499291	162.4068
cr8 30	9706.209	PP	NE Siberia	68.499291	162.4068
cr8 31	9845.447	PP	NE Siberia	68.499291	162.4068
cr8 33	9910.618661	PP	NE Siberia	68.499291	162.4068
cr8 34	18526.263	PP	NE Siberia	68.499291	162.4068
cr8 35	32396.823	PP	NE Siberia	68.499291	162.4068
cr8 36	40727.79815	PP	NE Siberia	68.499291	162.4068
cr8 37	41817.446	PP	NE Siberia	68.499291	162.4068
cr8 38	42555.509	PP	NE Siberia	68.499291	162.4068
cr8 39	43598.10169	PP	NE Siberia	68.499291	162.4068
tm9 13	11377.98751	SV2	North America	65.983333	-148.95
tm9 14	7918.702862	SV2	North America	65.983333	-148.95
cr4 1	7904.682123	SV2	North America	65.983333	-148.95
tm9 15	10868.81827	SV2	North America	65.983333	-148.95
ar2 14	31581.83287	SV2	North America	65.983333	-148.95
ar2 15	10975.85652	SV2	North America	65.983333	-148.95
ar2 17	11038.88826	SV2	North America	65.983333	-148.95
ar2 18	11747.75075	SV2	North America	65.983333	-148.95
cr4 2	modern	SV1	North America	65.983333	-148.95
cr4 3	modern	SV1	North America	65.983333	-148.95
cr4 4	modern	SV1	North America	65.983333	-148.95
ar3 25	11668.477	SV1	North America	65.983333	-148.95

Lab ID	Age	Site	Region	Latitude	Longitude
cr9 11	11313.843	SV1	North America	65.983333	-148.95
cr4 5	11116.66347	SV1	North America	65.983333	-148.95
ar5 29	11040.032	SV1	North America	65.983333	-148.95
ar5 22	11040.032	SV1	North America	65.983333	-148.95
cr4 6	10932.948	SV1	North America	65.983333	-148.95
cr9 12	10818.21967	SV1	North America	65.983333	-148.95
ar5 21	11081.821	SV1	North America	65.983333	-148.95
cr4 7	10818.21967	SV1	North America	65.983333	-148.95
cr4 9	9254.033499	SV1	North America	65.983333	-148.95
ar3 27	10357.99184	SV1	North America	65.983333	-148.95
ar5 23	8481.202	SV1	North America	65.983333	-148.95
ar3 26	7928.890956	SV1	North America	65.983333	-148.95
cr4 10	7788.25114	SV1	North America	65.983333	-148.95
ar3 4	10600	PS	North America	66.233333	-148.266667
ar3 5	12334.16709	PS	North America	66.233333	-148.266667
ar3 6	25725.12334	PS	North America	66.233333	-148.266667
ar3 7	23474.18758	PS	North America	66.233333	-148.266667
ar3 9	23759.51016	PS	North America	66.233333	-148.266667
ar3 17	31169.31119	PS	North America	66.233333	-148.266667
ar3 15	15578.85781	PS	North America	66.233333	-148.266667
ar3 10	18282.7706	PS	North America	66.233333	-148.266667
ar3 11	12311.81855	PS	North America	66.233333	-148.266667
ar3 12	8373	PS	North America	66.233333	-148.266667
ar3 13	21057.5863	PS	North America	66.233333	-148.266667
ar3 14	12721.08504	PS	North America	66.233333	-148.266667
ar3 18	12582.19345	PS	North America	66.233333	-148.266667
ar6 13	12954.56863	TH	North America	68.1933861	-162.5803667
ar6 14	12995.41353	TH	North America	68.1933861	-162.5803667
ar6 15	13036.96812	TH	North America	68.1933861	-162.5803667
ar6 17	13026.43423	TH	North America	68.1933861	-162.5803667
tm6 16	13393	ZL	North America	63.471037	-162.053181
tm6 15	7965	ZL	North America	63.471037	-162.053181

Lab ID	Age	Site	Region	Latitude	Longitude
tm6 14	4696	ZL	North America	63.471037	-162.053181
cr9 14	31842	ZL	North America	63.471037	-162.053181
tm6 17	17352	ZL	North America	63.471037	-162.053181
tm6 18	20387	ZL	North America	63.471037	-162.053181
cr9 15	30078	ZL	North America	63.471037	-162.053181
cr9 16	24974	ZL	North America	63.471037	-162.053181
cr9 17	32763	ZL	North America	63.471037	-162.053181
cr9 18	28574	ZL	North America	63.471037	-162.053181
cr9 19	29514	ZL	North America	63.471037	-162.053181
ar6 20	9059.271	RBS	North America	68.3534889	-158.8874361
ar6 21	9066.242	RBS	North America	68.3534889	-158.8874361
cr4 11	29294.21358	QC	North America	64.896755	-164.310837
cr9 21	30154.34896	QC	North America	64.896755	-164.310837
ar4 5	29731.11565	QC	North America	64.896755	-164.310837
ar3 19	33342.48321	QC	North America	64.896755	-164.310837
ar3 20	30175.04582	QC	North America	64.896755	-164.310837
ar3 21	29725.21352	QC	North America	64.896755	-164.310837
ar3 22	29629.23039	QC	North America	64.896755	-164.310837
cr4 12	35875.85049	QC	North America	64.896755	-164.310837
cr9 26	23735.55592	QC	North America	64.896755	-164.310837
cr9 27	28857.22227	QC	North America	64.896755	-164.310837
ar4 4	50000	QC	North America	64.896755	-164.310837
cr8 13	5707	AMR	North America	67.74383164	-156.1921285
cr8 14	6070	AMR	North America	67.74383164	-156.1921285
cr8 15	modern	AMR	North America	67.74383164	-156.1921285
ar2 19	30071.49519	GS	North America	63.933333	-138.966667
ar2 20	29237.08217	GS	North America	63.933333	-138.966667
ar2 21	28641.88384	GS	North America	63.933333	-138.966667
ar2 22	30218.34006	GS	North America	63.933333	-138.966667
ar2 23	29818.5444	GS	North America	63.933333	-138.966667
ar2 25	27019.00063	GS	North America	63.933333	-138.966667
ar2 26	24473.28649	GS	North America	63.933333	-138.966667

Lab ID	Age	Site	Region	Latitude	Longitude
ar2 27	24518.03425	GS	North America	63.933333	-138.966667
ar2 28	24194.3038	GS	North America	63.933333	-138.966667
cr9 22	29787.32941	GS	North America	63.933333	-138.966667
cr9 23	46686.01317	GS	North America	63.933333	-138.966667
ar2 29	24585.93241	GS	North America	63.933333	-138.966667
ar2 30	23103.87112	GS	North America	63.933333	-138.966667
ar2 31	27208.28593	GS	North America	63.933333	-138.966667
ar3 1	23528.0138	GS	North America	63.933333	-138.966667
ar3 2	33499.75836	GS	North America	63.933333	-138.966667
ar3 3	31071.67078	GS	North America	63.933333	-138.966667
ar3 23	16653.43	RS	North America	63.69	-138.58
ar3 28	15475.46894	RS	North America	63.69	-138.58
ar3 29	9439.738672	RS	North America	63.69	-138.58
cr9 24	9496.325294	RS	North America	63.69	-138.58
ar3 30	9359.57	RS	North America	63.69	-138.58
cr4 13	9359.57	RS	North America	63.69	-138.58
cr9 25	9386.930089	RS	North America	63.69	-138.58
ar3 31	8861.750026	RS	North America	63.69	-138.58
ar4 6	6505.759	RS	North America	63.69	-138.58
ar4 7	4922.716	RS	North America	63.69	-138.58
cr4 14	47457.38265	CM	North America	63.67	-138.64245
ar5 25	24842.03584	CM	North America	63.67	-138.64245
cr4 15	45753.00818	CM	North America	63.67	-138.64245
cr9 28	46686.01317	CM	North America	63.67	-138.64245
cr9 29	41197.35694	CM	North America	63.67	-138.64245
ar5 26	29097.79502	CM	North America	63.67	-138.64245
cr4 17	42271.27652	CM	North America	63.67	-138.64245
ar5 30	40039.29378	CM	North America	63.67	-138.64245
cr4 18	20833.87113	CM	North America	63.67	-138.64245
ar5 27	150.9937635	CM	North America	63.67	-138.64245
cr4 19	46181.86744	CM	North America	63.67	-138.64245
cr4 20	47819.82824	CM	North America	63.67	-138.64245

Lab ID	Age	Site	Region	Latitude	Longitude
cr4 21	46181.86744	CM	North America	63.67	-138.64245
cr4 22	40158.03844	CM	North America	63.67	-138.64245
cr4 23	45556.62164	CM	North America	63.67	-138.64245
cr9 30	128.6921975	CM	North America	63.67	-138.64245
cr9 31	175.125219	CM	North America	63.67	-138.64245
cr9 32	142.554935	CM	North America	63.67	-138.64245
ar5 28	14173.76	TC	North America	63.097244	-139.538727
cr4 25	30794.46	GR	North America	63.683333	-138.6
cr8 4	modern	GR	North America	63.683333	-138.6
cr8 5	modern	GR	North America	63.683333	-138.6
cr4 26	modern	NP	North America	60.578887	-139.005478
cr4 27	modern	NP	North America	60.578887	-139.005478
cr4 28	modern	NP	North America	60.578887	-139.005478
cr4 29	modern	NP	North America	60.578887	-139.005478
cr9 33	1452	BS	North Atlantic Islands	67.609221	-76.245117
cr9 34	2548	BS	North Atlantic Islands	67.609221	-76.245117
cr9 35	9482	BS	North Atlantic Islands	67.609221	-76.245117
cr9 36	9684	BS	North Atlantic Islands	67.609221	-76.245117
cr6 15	modern	09C1	North Atlantic Islands	78.04860472	15.09092799
cr6 17	modern	09C1	North Atlantic Islands	78.04860472	15.09092799
cr6 18	modern	09C1	North Atlantic Islands	78.04860472	15.09092799
cr6 19	modern	09C1	North Atlantic Islands	78.04860472	15.09092799
cr6 20	modern	09C1	North Atlantic Islands	78.04860472	15.09092799
cr6 21	modern	09C1	North Atlantic Islands	78.04860472	15.09092799
cr6 22	3404.829289	09C2	North Atlantic Islands	78.04761404	15.09239121
cr6 23	4294.749	09C2	North Atlantic Islands	78.04761404	15.09239121
cr6 25	5144.008137	09C2	North Atlantic Islands	78.04761404	15.09239121
cr6 26	5424.356498	09C2	North Atlantic Islands	78.04761404	15.09239121
cr9 49	109.4092	ES	North Atlantic Islands	78.032864	15.113404
cr9 50	152.5916	ES	North Atlantic Islands	78.032864	15.113404
cr9 51	122.2708	ES	North Atlantic Islands	78.032864	15.113404
cr9 37	modern	CL10	North Atlantic Islands	78.0925	14.9787

Lab ID	Age	Site	Region	Latitude	Longitude
cr9 38	modern	CL10	North Atlantic Islands	78.0925	14.9787
cr9 39	133.1143913	CL10	North Atlantic Islands	78.0925	14.9787
cr9 40	124.5749249	CL10	North Atlantic Islands	78.0925	14.9787
cr9 41	130.2091692	CL10	North Atlantic Islands	78.0925	14.9787
cr9 42	130.1949009	CL10	North Atlantic Islands	78.0925	14.9787
cr9 43	300.4478317	CL10	North Atlantic Islands	78.0925	14.9787
cr9 44	314.2736978	CL10	North Atlantic Islands	78.0925	14.9787
cr9 45	391.7450447	CL10	North Atlantic Islands	78.0925	14.9787
cr9 46	1360.520736	CL10	North Atlantic Islands	78.0925	14.9787
cr9 47	1329.243442	CL10	North Atlantic Islands	78.0925	14.9787
cr9 48	1407.08741	CL10	North Atlantic Islands	78.0925	14.9787
cr6 27	6154.645	DA	North Atlantic Islands	79.7215833	10.94705
cr6 28	4955.559925	DA	North Atlantic Islands	79.7215833	10.94705
cr6 29	4728.522	DA	North Atlantic Islands	79.7215833	10.94705
cr6 30	5759.982426	DA	North Atlantic Islands	79.7215833	10.94705
cr6 31	4636.085	DA	North Atlantic Islands	79.7215833	10.94705
cr6 5	4563.194323	DA	North Atlantic Islands	79.7215833	10.94705
cr7 1	modern	RS1	North Atlantic Islands	78.470996	16.215293
cr7 2	326.857	RS1	North Atlantic Islands	78.470996	16.215293
cr7 3	320.2160859	RS1	North Atlantic Islands	78.470996	16.215293
cr7 4	10777.964	RS1	North Atlantic Islands	78.470996	16.215293
cr6 6	16952.52909	RS1	North Atlantic Islands	78.470996	16.215293
cr7 5	modern	RS2	North Atlantic Islands	78.558424	16.434787
cr7 6	2540.718	RS2	North Atlantic Islands	78.558424	16.434787
cr7 7	20692.37627	RS2	North Atlantic Islands	78.558424	16.434787
cr7 9	24032.04051	RS2	North Atlantic Islands	78.558424	16.434787
cr7 10	11282.911	RS2	North Atlantic Islands	78.558424	16.434787
cr7 11	13305.38383	RS2	North Atlantic Islands	78.558424	16.434787
cr7 12	16643.241	RS2	North Atlantic Islands	78.558424	16.434787
cr7 13	2521.523117	RS2	North Atlantic Islands	78.558424	16.434787
ar6 22	406.24	LI	North Atlantic Islands	64.398201	-50.201302
ar6 23	526.743	LI	North Atlantic Islands	64.398201	-50.201302

Lab ID	Age	Site	Region	Latitude	Longitude
ar6 25	2426.283	LI	North Atlantic Islands	64.398201	-50.201302
ar6 26	3517.167	LI	North Atlantic Islands	64.398201	-50.201302
ar6 27	4429.823	LI	North Atlantic Islands	64.398201	-50.201302
ar6 28	7262.996	LI	North Atlantic Islands	64.398201	-50.201302
ar6 29	9199.098	LI	North Atlantic Islands	64.398201	-50.201302
ar6 30	10156.769	LI	North Atlantic Islands	64.398201	-50.201302
ar6 31	10580.101	LI	North Atlantic Islands	64.398201	-50.201302
cr1 6	3964.614	K608	North Atlantic Islands	64.60217	-50.5013
cr1 7	4080.588	K608	North Atlantic Islands	64.60217	-50.5013
cr7 14	4129.353	K608	North Atlantic Islands	64.60217	-50.5013
cr1 9	4547.276	K608	North Atlantic Islands	64.60217	-50.5013
cr7 15	6140.674	K608	North Atlantic Islands	64.60217	-50.5013
cr1 10	6598.213	K608	North Atlantic Islands	64.60217	-50.5013
cr7 17	6915.254	K608	North Atlantic Islands	64.60217	-50.5013
cr7 18	7828.568	K608	North Atlantic Islands	64.60217	-50.5013
cr1 11	8637.691	K608	North Atlantic Islands	64.60217	-50.5013
cr7 19	8900.048	K608	North Atlantic Islands	64.60217	-50.5013
cr1 12	9337.331686	K608	North Atlantic Islands	64.60217	-50.5013
cr7 20	9814.964	K608	North Atlantic Islands	64.60217	-50.5013
cr1 13	6857.085578	LC	North Atlantic Islands	61.1399	-45.5347
cr1 14	8478.048842	LC	North Atlantic Islands	61.1399	-45.5347
cr1 15	9388.987068	LC	North Atlantic Islands	61.1399	-45.5347
cr1 17	7782.086035	LC	North Atlantic Islands	61.1399	-45.5347
cr1 18	1753.520876	LC	North Atlantic Islands	61.1399	-45.5347
cr1 19	333.374	LS	North Atlantic Islands	65.6833333	-37.9166667
cr7 21	479.086	LS	North Atlantic Islands	65.6833333	-37.9166667
cr7 22	1317.903	LS	North Atlantic Islands	65.6833333	-37.9166667
cr1 20	1466.966047	LS	North Atlantic Islands	65.6833333	-37.9166667
cr7 23	2056.584	LS	North Atlantic Islands	65.6833333	-37.9166667
cr1 21	2383.557	LS	North Atlantic Islands	65.6833333	-37.9166667
cr7 25	3156.372	LS	North Atlantic Islands	65.6833333	-37.9166667
cr1 22	4506.222	LS	North Atlantic Islands	65.6833333	-37.9166667

Lab ID	Age	Site	Region	Latitude	Longitude
cr7 26	3604.301	LS	North Atlantic Islands	65.6833333	-37.9166667
cr1 23	4113.28	LS	North Atlantic Islands	65.6833333	-37.9166667
cr7 27	4992.207	LS	North Atlantic Islands	65.6833333	-37.9166667
cr7 28	6023.171	LS	North Atlantic Islands	65.6833333	-37.9166667
cr1 25	6890.828726	LS	North Atlantic Islands	65.6833333	-37.9166667
cr7 29	7453.342	LS	North Atlantic Islands	65.6833333	-37.9166667
cr7 30	7904.480978	LS	North Atlantic Islands	65.6833333	-37.9166667
cr7 31	8162.49	LS	North Atlantic Islands	65.6833333	-37.9166667
cr9 52	8431.226	ANL	North Europe	69.254384	16.06003
cr9 53	8821.868	ANL	North Europe	69.254384	16.06003
cr9 54	9664.871	ANL	North Europe	69.254384	16.06003
cr9 55	11952.746	ANL	North Europe	69.254384	16.06003
cr9 56	13659.33	ANL	North Europe	69.254384	16.06003
cr9 57	14679.709	ANL	North Europe	69.254384	16.06003
cr9 58	15478.439	ANL	North Europe	69.254384	16.06003
cr9 59	16968.755	ANL	North Europe	69.254384	16.06003
cr9 60	17264.582	ANL	North Europe	69.254384	16.06003
cr9 61	17526.994	ANL	North Europe	69.254384	16.06003
cr9 62	18608.359	ANL	North Europe	69.254384	16.06003
cr9 63	19398.505	ANL	North Europe	69.254384	16.06003
cr9 64	20640.875	ANL	North Europe	69.254384	16.06003
cr9 65	21412.308	ANL	North Europe	69.254384	16.06003
cr9 66	21810.745	ANL	North Europe	69.254384	16.06003
cr9 67	22075.281	ANL	North Europe	69.254384	16.06003
cr9 68	122.2708	VA	North Europe	70.31666667	30.01666667
cr9 69	113.2701	VA	North Europe	70.31666667	30.01666667
cr9 70	122.2708	VA	North Europe	70.31666667	30.01666667
cr9 73	135.7463	VA	North Europe	70.31666667	30.01666667
cr8 17	191.451	ELS	North Atlantic Islands	63.647254	-18.254364
cr8 18	325	ELS	North Atlantic Islands	63.647254	-18.254364
cr8 19	1016	ELS	North Atlantic Islands	63.647254	-18.254364
cr8 20	1079	ELS	North Atlantic Islands	63.647254	-18.254364

Lab ID	Age	Site	Region	Latitude	Longitude
cr8 21	1091.207	ELS	North Atlantic Islands	63.647254	-18.254364
cr8 22	1134.043	ELS	North Atlantic Islands	63.647254	-18.254364
cr8 23	1190	ELS	North Atlantic Islands	63.647254	-18.254364
cr8 25	1356.223	ELS	North Atlantic Islands	63.647254	-18.254364
cr8 26	1402.882	ELS	North Atlantic Islands	63.647254	-18.254364
cr8 9	6876.707	06D1	North Europe	69.969472	23.902103
cr8 10	12050.15	06D1	North Europe	69.969472	23.902103
cr8 11	14078.435	06D1	North Europe	69.969472	23.902103
cr8 12	15823.33	06D1	North Europe	69.969472	23.902103

Appendix: Larger versions of selected figures

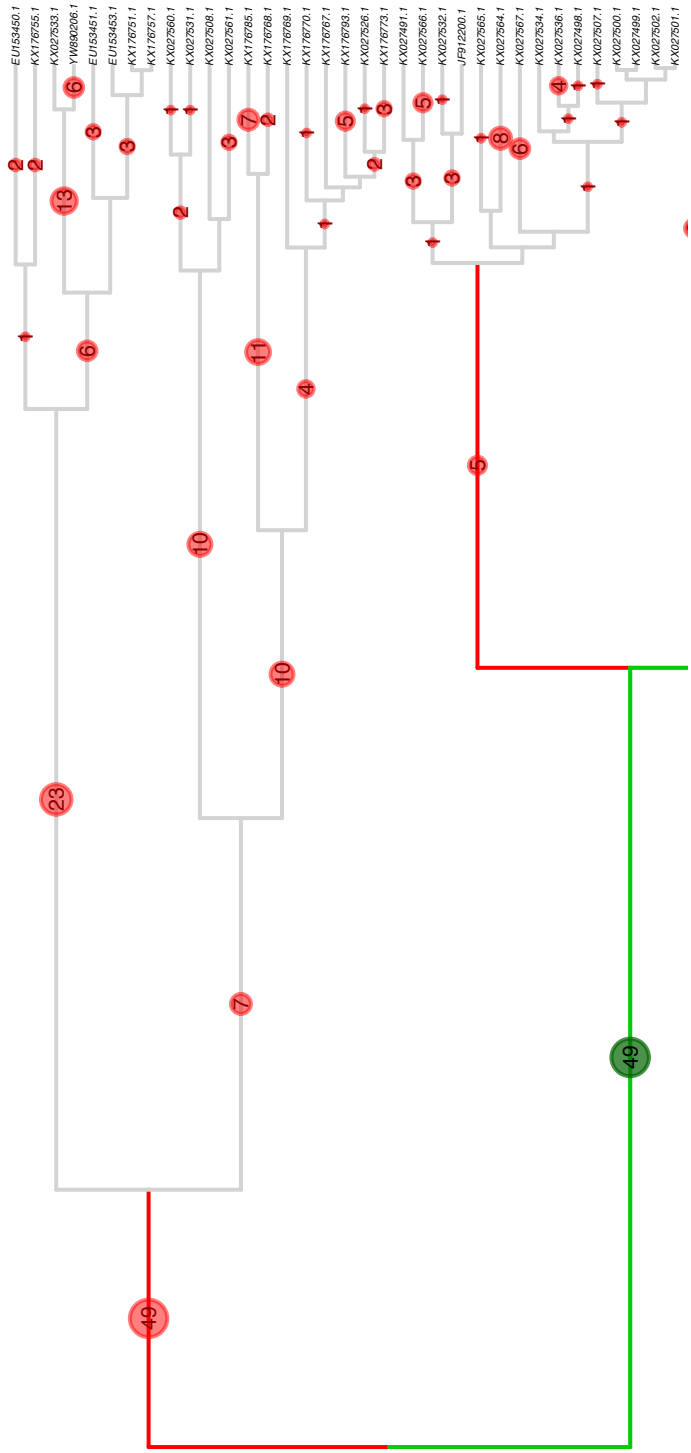


Figure 37: A zoomed in, rotated copy of the top half of Figure 12, here for legibility.

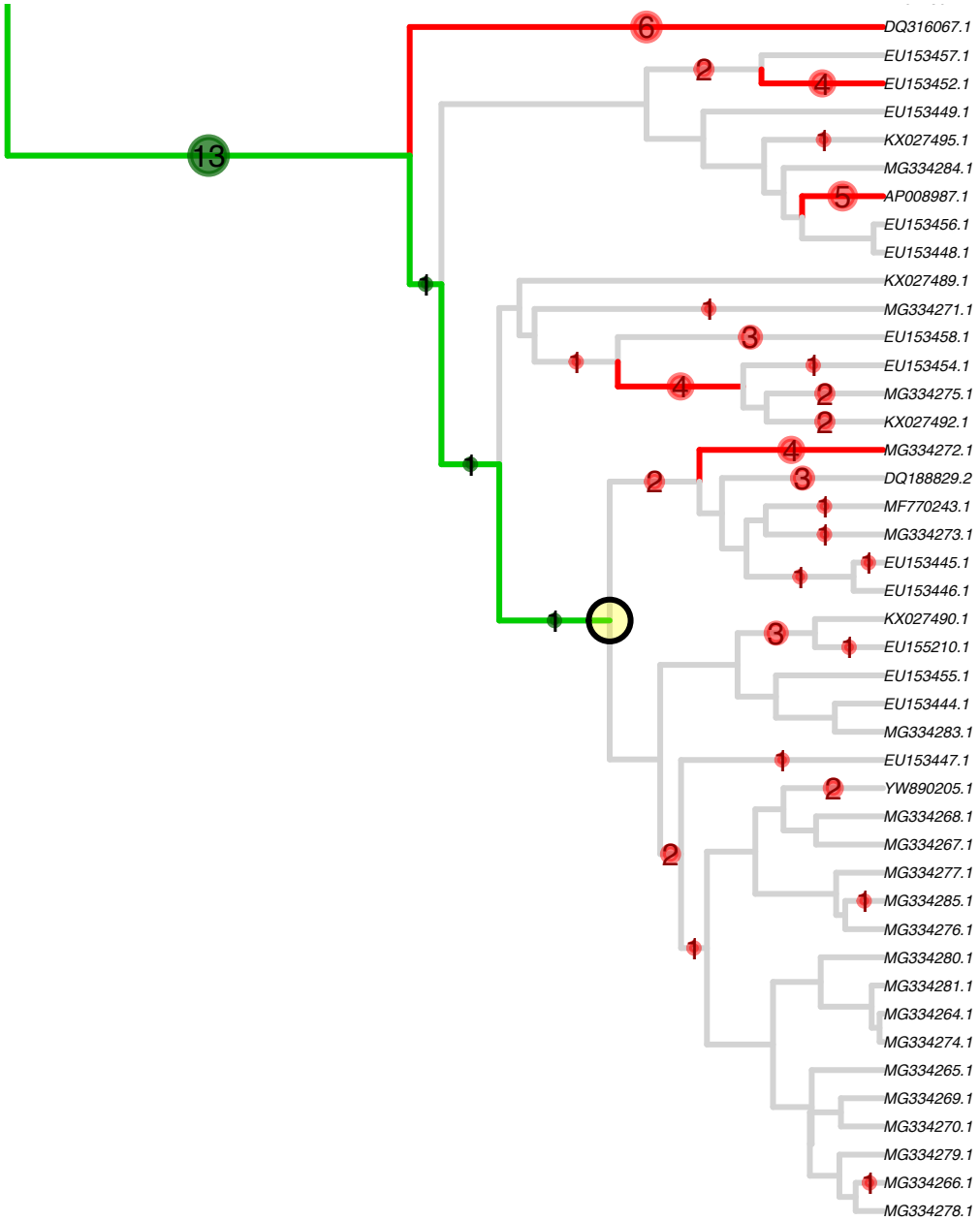


Figure 38: A zoomed in copy of the bottom half of Figure 12, here for legibility.

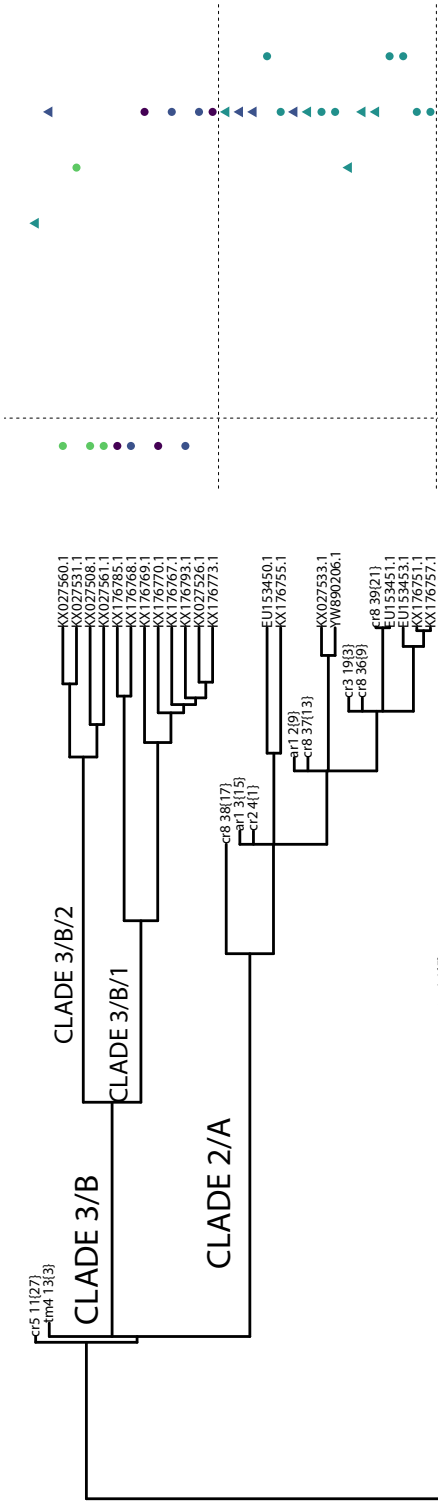


Figure 39: A zoomed in, rotated copy of the first quarter of Figure 13, here for legibility.

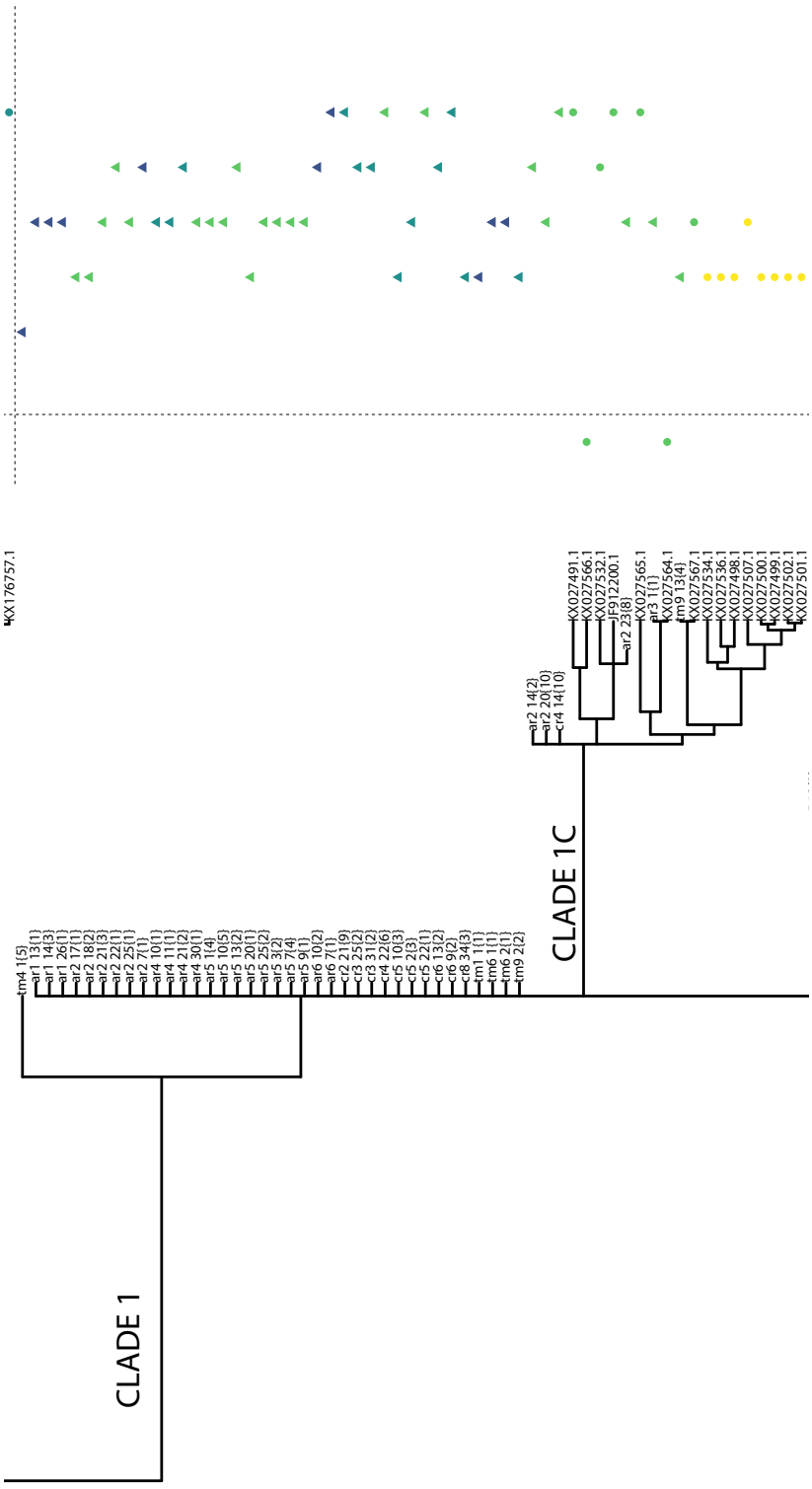


Figure 40: A zoomed in, rotated copy of the second quarter of Figure 13, here for legibility.

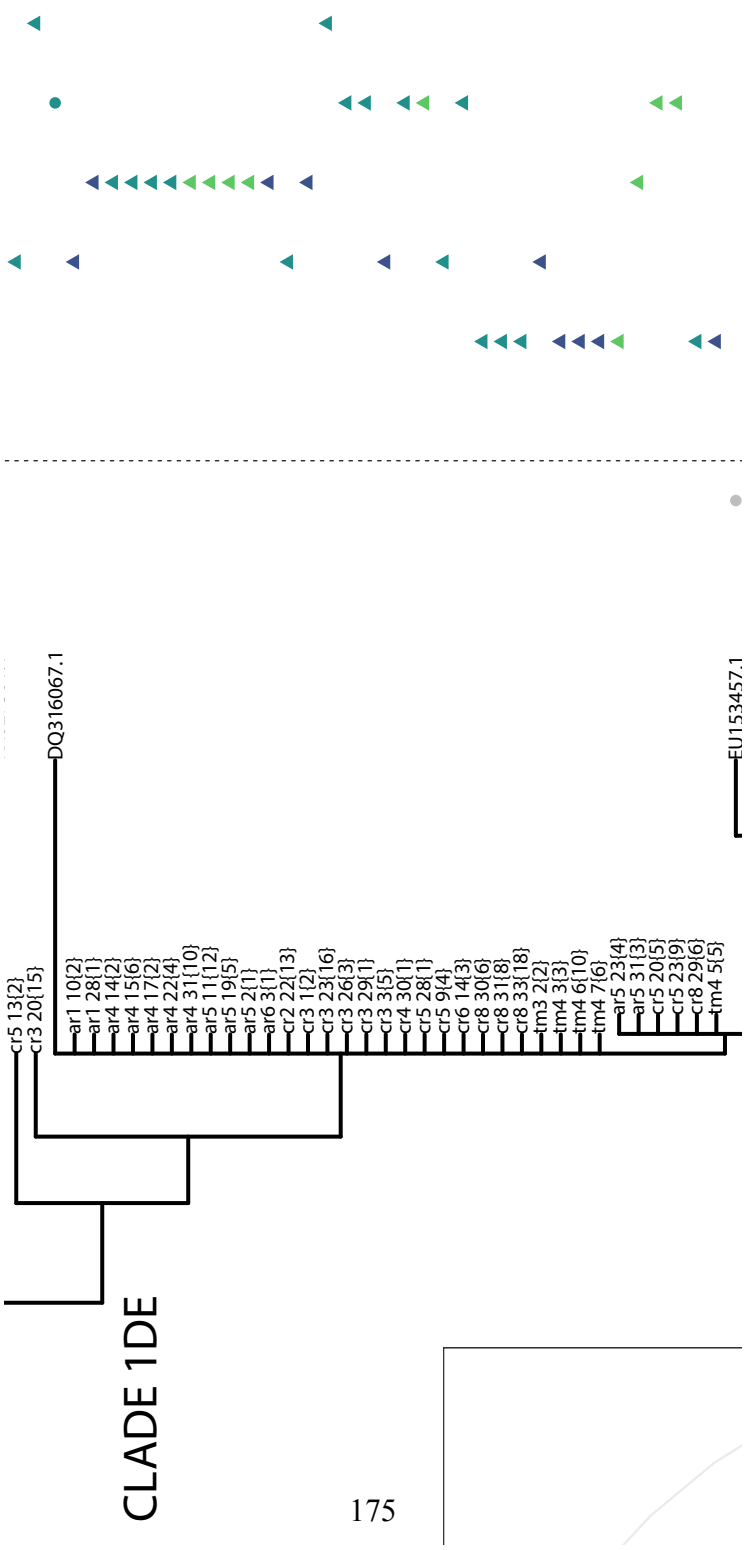


Figure 41: A zoomed in, rotated copy of the third quarter of Figure 13, here for legibility.

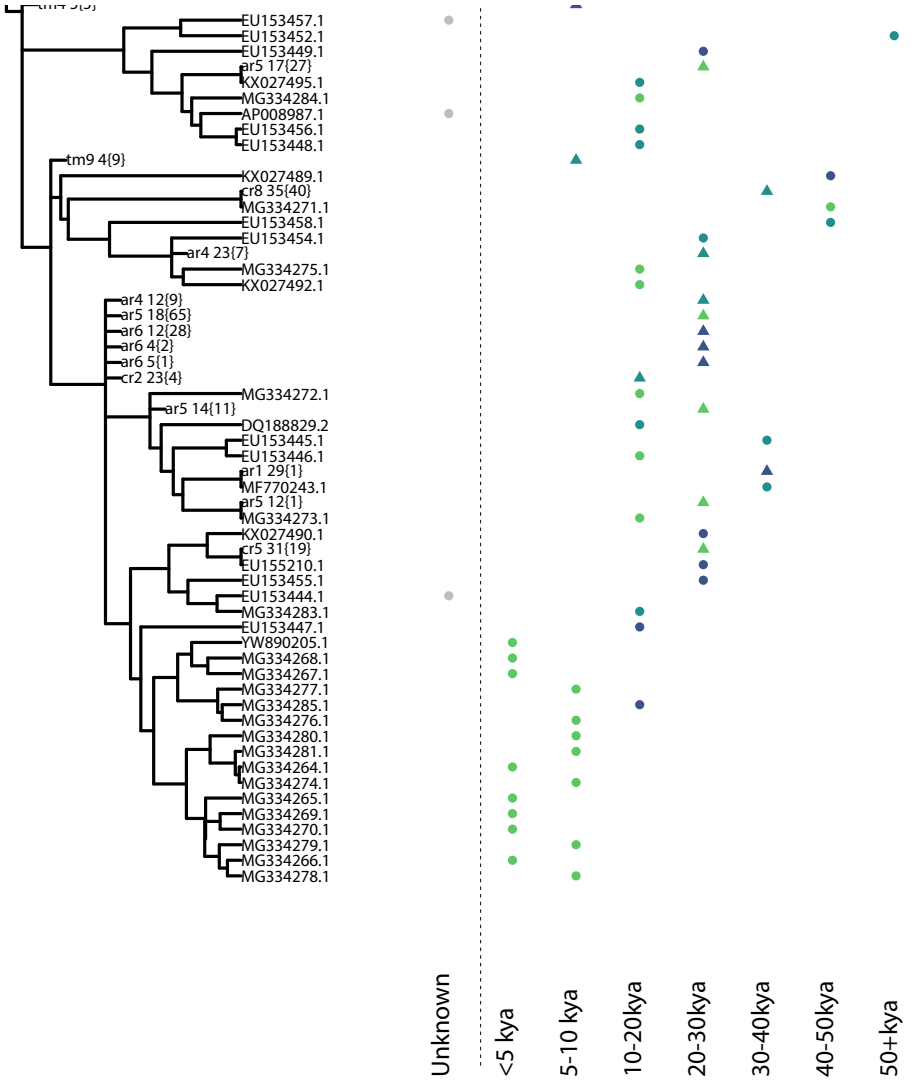


Figure 42: A zoomed in copy of the last quarter of Figure 13, here for legibility.

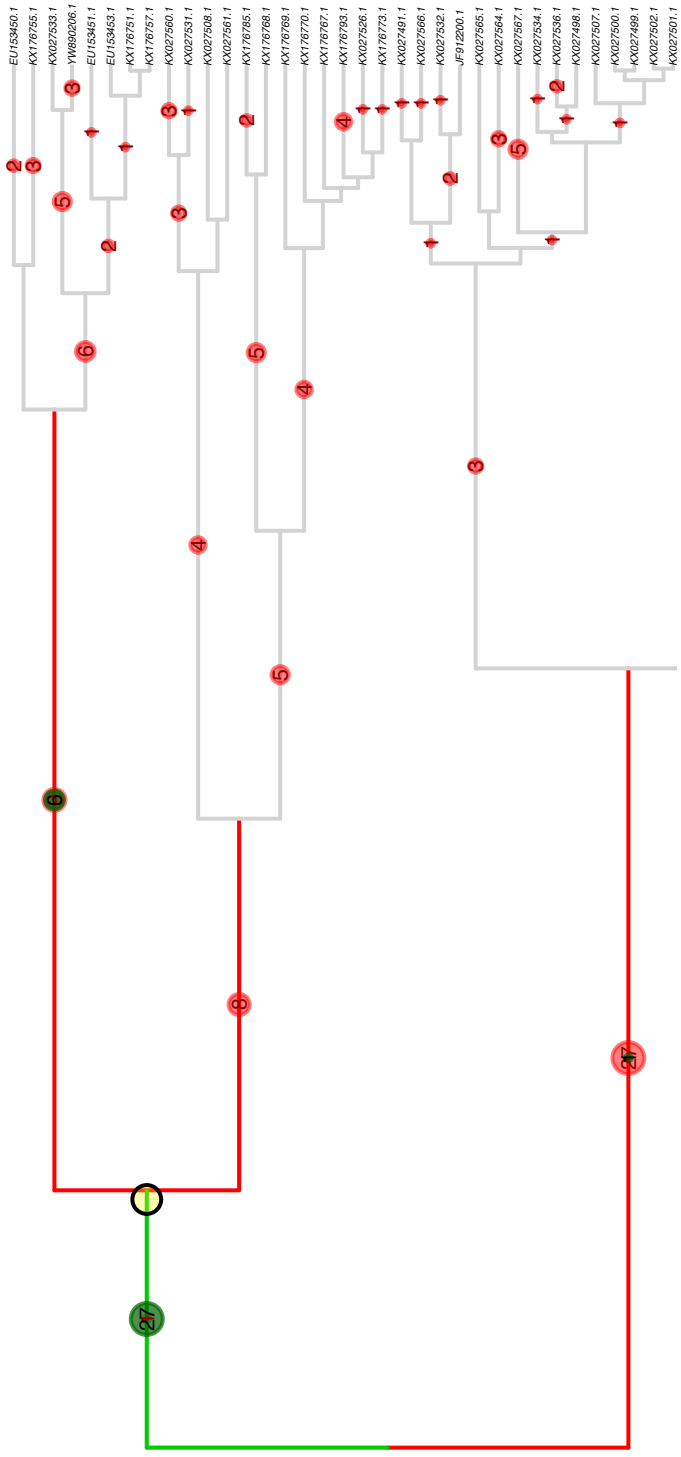


Figure 43: A zoomed in, rotated copy of the top half of Figure 14, here for legibility.

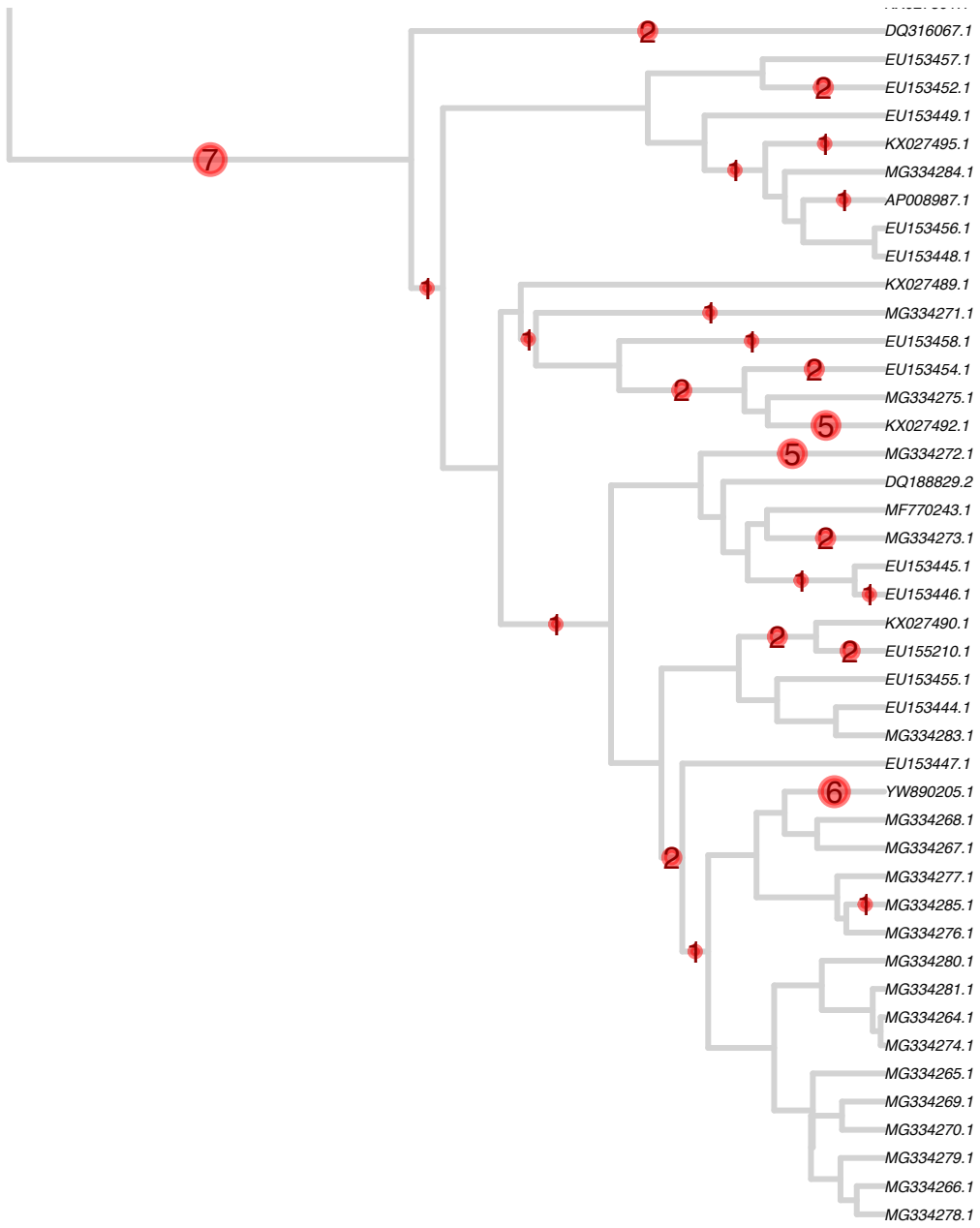


Figure 44: A zoomed in copy of the bottom half of Figure 14, here for legibility.