UNIVERSITY OF TARTU

FACULTY OF SCIENCE AND TECHNOLOGY

INSTITUTE OF MATHEMATICS AND STATISTICS

Erik Mandel

Reconstructing Ancient Migration Paths from Mitochondrial Genomes

Mathematical Statistics

Bachelor's Thesis (9 ECTS)

Supervisor: Dr Jon Anders Eriksson

TARTU 2023

Reconstructing Ancient Migration Paths from Mitochondrial Genomes

Bachelor Thesis

Erik Mandel

Abstract

This bachelor's thesis aims to identify the timings and trajectories of major pre-historic population movements. The ancient mitochondrial genomes from Allen Ancient DNA Resource are used to construct the phylogenetic relationships with the help of the BEAST software. The first chapter gives the theoretical background for phylogenetic analysis of mitochondrial DNA genome sequences. The second chapter introduces the chosen model with its parameters, the Ancestral Trails Framework principles, and the data preparation for the phylogenetic analysis. The third chapter presents the analysis results performed on the mitochondrial data and models constructed with Ancestral Trails Framework. The author manages to assemble a large dataset with mitochondrial sequences ranging from ancient to present-day samples and estimate the phylogenetic relationships and divergence times between them. The analysis concluded that it is possible to detect ancient migration events and their directions. However, due to the analysed sample's limitations, the human groups' exact timings of splits are uncertain.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics.

Key Words: Mitochondrial genomes, evolution, migrations, Markov models, Bayesian statistics.

Muistsete rändeteede konstrueerimine mitokondriaalsete genoomide põhjal

Bakalaureusetöö

Erik Mandel

Lühikokkuvõte

Käesoleva bakalaureusetöö eesmärk on tuvastada peamised muistsete rändeteede toimumise ajad ning trajektoorid. Muistsete inimeste DNA proove koondavast andmebaasist nimega *Allen Ancient DNA Resource* sorteeriti välja mitokondriaalse DNA järjestused ning analüüsiti andmesubjektide fülogeneetilisi suhteid BEAST tarkvaraga. Esimeses peatükis kirjeldatakse töö teoreetilist tausta, mis on vajalik mitokondriaalse DNA genoomidel fülogeneetilise analüüsi läbiviimiseks. Teises peatükis tutvustatakse analüüsi läbiviimiseks loodud mudelit, eellaste ränderaamistiku (*Ancestral Trails Framework*) põhimõtteid ning andmetöötlust fülogeneetilise analüüsi tarbeks. Kolmandas peatükis esitatakse fülogeneetilise analüüsi tulemused ning eellaste ränderaamistiku põhjal koostatud mudelid. Läbiviidud analüüsist selgus, et muistseid rändeteid ning peamisi trajektoore on võimalik analüüsitud andmete põhjal tuvastada. Analüüsitud andmete vähesuse tõttu ei ole võimalik täpselt määrata inimrühmade lahknemisaegu.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: mitokondriaalsed genoomid, evolutsioon, migratsioon, Markovi mudelid, Bayesi statistika

Table of Contents

Introduction

The studies of ancient migration paths are vital to understanding the cultural history of humans and population genetics, as it provides valuable insights into humanity's expansion across the globe. By analysing the DNA samples found in modern-day humans and ancient remains, it is possible to reconstruct the historical movements and densities of human populations and their interactions for the last 100 000 years. One of the most used biological tools to reconstruct such paths is mitochondrial DNA (mtDNA). The advances in Human Genome Project, especially in DNA sequencing, have allowed scientists to extract high-quality mtDNA sequenced from poorly preserved samples, allowing us to construct the movements of different populations under a large timescale with unprecedented resolution. (Lipson et al., 2018; Soares et al., 2011)

Several recent discoveries have helped to put human history in different regions of the world into a new perspective and changed previous beliefs. To illustrate that fact, analysing the sequences of ancient DNA from the Americas has shed new light on the continent's colonisation. Contradictory to earlier beliefs, it was found that the first humans arrived on the continent much earlier than historians previously thought (Raghavan et al., 2015). Other research found that multiple waves of migration from different parts of the world contributed to the genetic diversity of the Native American populations (Moreno-Mayar et al., 2018). By leveraging open datasets and computational tools, it is now possible to construct comprehensive maps of ancient migration paths based on mtDNA analysis, providing insights into the complex movements of our ancestors throughout history.

This project aims to analyse genetic differences between individuals to uncover populations' joint maternal demographic history worldwide. The author assembles a large, global dataset of mitochondrial genome sequences from present-day and ancient people and uses a spatially explicit framework to reconstruct the locations of ancestors to the individuals in the dataset. The main objectives of the thesis are:

1. Assembly of a global dataset of whole mitochondrial genome sequences (mt genomes) from present-day and ancient populations worldwide.
2. Estimate the phylogenetic relationships and divergence dates between the mt genomes.
3. Identify timings and trajectories of major (pre) historical population movements and human habitats that were key genetic sources to present-day genetic variation.

The project is carried out using computational methods and bioinformatic tools, such as BEAST, and the data is obtained from publicly available resources, such as Allen Ancient DNA Resource.

The thesis consists of three major parts: in the first one, sufficient theoretical and mathematical background about mtDNA, phylogenetic trees and ancestral trails framework is given to the reader; the second part describes the data and methods used to reconstruct phylogenies of mitochondrial genomes and the ancient migration paths. The third part presents the results and conclusions of the study.

The author would like to express his sincere gratitude to this paper's supervisor Professor Anders Eriksson. His patience, enthusiasm towards the topic, and expert academic and mental guidance throughout this thesis's entire work process have proved invaluable. His exceptional character and expertise have shaped the direction of the following paper and the future of the author's education.

1. Theoretical Background

1.1 Mitochondrial DNA

Mitochondrial DNA (mtDNA) is a circular chromosome found in the powerhouse of the cell known as mitochondria. The mitochondrial genome of humans consists of the displacement loop (D-loop) and 37 genes: 13 genes for protein coding, 22 for transfer RNA (tRNA), and two ribosomal RNA (rRNA). The genes in the mtDNA are responsible for producing ATP in the mitochondria. The D-loop initiates mtDNA replication and transcription. Moreover, the D-loop has a significantly higher mutation rate than the other 37 genes, primarily concentrated to three short hyper-variable regions. (Doimo et al., 2020)

Contrary to nuclear DNA, inherited from both parents, the mtDNA passes only from mother to its offspring without recombination. This attribute helps us to trace maternal lineages through generations and map hierarchic relationships between individuals. Furthermore, human cells typically harbour large numbers of mitochondria, each carrying a copy of the mt genome. As a result, it is often possible to recover mt genomes also when nuclear DNA is too degraded to sequence, making it possible to obtain genomic data from a wide temporal and geographic range. (Bandelt et al., 2006; Doimo et al., 2020)

Thanks to these properties, we can use phylogenetic methods to analyse mtDNA sequences and draw valuable insights from the origins and migrations of our ancestors. We can analyse the sequenced mtDNA data by grouping them into haplogroups. A haplogroup is a set of related mtDNA sequences that share a common ancestor and are defined by one or more mutations that distinguish them from other haplogroups. Comparing haplogroup distributions in different populations allows to infer differences in ancestry between populations and movements of individuals over time (Torroni et al., 2020). Additional substitution, molecular clock or tree prior models can further solidify the generated models to draw even more insights from the sequenced mtDNA data (Bandelt et al., 2006).

1.2 Molecular Clock Models

The technique that uses the mutation rate of biomolecules, such as protein sequences or DNA, to estimate the amount of time needed for a certain amount of evolutionary change to occur is often figuratively referred to as the molecular clock. The underlying theory behind such clock models is that most amino acid mutations and substitutions in a nucleotide in a gene occur at a

constant rate, making it possible to count those differences between two separate organisms. The differences themselves mark the passage of time and therefore help scientists to estimate the time since the species last shared a common ancestor. (Hedges & Kumar, 2009; Rodríguez-Trelles et al., 2003)

The main limitations of molecular clocks, in their most simple form, come from the assumption that mutations are uniformly distributed over the genome. However, that is not primarily true for the whole gene. Natural selection also plays a crucial role in real life in favouring some changes over others. In particular, most mutations that change the structure of expressed proteins (non-synonymous mutations) are detrimental and therefore tend to be suppressed by natural selection. Because of the way DNA codes for amino acids, mutations at the first and second codon positions are more often non-synonymous than mutations at the third codon positions, and consequently have lower substitution rates. Meanwhile, most molecular clocks portray genetic differences as selectively neutral, which do not affect the organism's fitness. Despite the criticism, numerous research has proven the minimal overall effect of these problems on the analysis results, declaring the usage of molecular clock models as an invaluable tool for building evolutionary timescales and its service as the null model for testing evolutionary and mutation rates in different species. (Hedges & Kumar, 2009; Kumar, 2005)

One of the most widely used molecular clock models is the strict clock model, with the additional assumption that the constant rate at which mutations occur is the same for all lineages. Mathematically, the clock model is represented as:

$$T = \frac{D}{\mu},$$

where T is the estimated time of divergence between two species, D is the number of genetic differences between two sequences and μ is the estimated rate of genetic mutations per unit of time. (Yang, 2006)

1.3 Nucleotide Substitution Models

Unless stated otherwise, the following paragraph and its subparagraphs are based on the books Yang, 2006 and Drummond & Bouckaert, 2015.

A *nucleotide* is a molecule consisting of a nitrogen-containing base. In DNA the nucleotides are A, C, G and D, which mark adenine, cytosine, guanine, and thymine respectively. The

substitutions between the two pyrimidines (T ↔ C) or between the two purines (A ↔ G) are called *transitions*, while those between a pyrimidine and a purine (T, C ↔ A, G) are called *transversions*. The distance between two sequences is the expected number of nucleotide substitutions per site.

In molecular evolution, nucleotide substitution models (a.k.a. mutation or site models) are Markov-Chain models used in phylogenetic analysis. They help to describe the variability in rates and substitution patterns among the different sites in a nucleotide sequence alignment. These models allow us to calculate pairwise distances between two sites and, using cluster algorithms, convert the distance matrix into a phylogenetic tree. The Markov-process models of nucleotide substitution used in the distance calculation can be used to calculate the likelihood of a given phylogenetic tree, given the observed sequence data and assumptions about the distribution of substitution rates among sites.

Molecular scientists categorise site models into either empirical or theoretical models. Empirical models estimate the rates and patterns of substitution directly from the data. In contrast, theoretical models make assumptions about the underlying process of molecular evolution and use these assumptions to derive a mathematical model of sequence evolution.

The theoretical models, like the Jukes-Cantor Model (JC69) and the Kimura (K80) model, are independent of the sequence; only transition rates are considered. Therefore, when their substitution process reaches equilibrium, the sequence will have equal proportions of four nucleotides. This assumption, however, is primarily unrealistic for every real-world dataset. Therefore, the models which fall into this category will not be used to analyse the topic of this thesis.

The empirical site models, such as Tamura-Nei (TN93) and Felsenstein (F84), accommodate unequal base compositions. Compared to the theoretical substitution models, they allow for different rates for transitions and transversions calculated from the analysed data. They can also incorporate site-specific heterogeneity and accommodate comprehensive options for parametrisation that make the computational process significantly faster. The models that fall into this category are therefore more directly relevant for analysing ancient migration paths.

1.3.1 The HKY model

One widely used empirical site model is the Hasegawa-Kishino-Yano (HKY) model. It was proposed by the authors with the same names in 1985 to better model the substitution process in mtDNA. The HKY model is a simplified version of the more general but computationally complex General Time Reversible (GTR) model. It, therefore, is often used instead of GTR when the computational resources are limited.

The HKY model combines the parameters in the K80 and F81 models to allow for both unequal base frequencies and a transition/transversion bias. The assumption is that the relative rates of different types of substitutions remain constant over time and that the nucleotide frequencies reach a steady-state distribution.

Let $X(t)$ be the random variable representing the state of nucleotide substitution at time t. Assuming that this Markov process is in state $i \in \{T, C, A, G\}$ at time t, the probability that the process transitions to state $j \in \{T, C, A, G\}$ is governed by a substitution rate matrix.

The substitution-rate matrix under the HKY model is

$$Q = (q_{ij}) = \begin{pmatrix} -(\alpha\pi_C + \beta\pi_R) & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & -(\alpha\pi_T + \beta\pi_R) & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & -(\alpha\pi_G + \beta\pi_Y) & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & -(\alpha\pi_A + \beta\pi_Y) \end{pmatrix},$$

where parameters $\pi_T, \pi_C, \pi_A, \pi_G$ are used to specify substitution rates (number of times the nucleotide is observed in a sequence divided by the total number of nucleotides at that position), α and β are the rates of transitions and transversions respectively, $\pi_Y = \pi_T + \pi_C$ and $\pi_R = \pi_A + \pi_G$. The diagonal elements are omitted for clarity so that the rates for a given state should sum to zero across the row. The matrix of transition probabilities over time t, is $P(t) = \exp(Qt)$.

The steady-state distribution under HKY is $\pi = (\pi_T, \pi_C, \pi_A, \pi_G)$, as can be verified directly by checking the identity $\pi Q = (0, 0, 0, 0)$, which implies $\pi P(t) = \pi$ for all t.

1.3.2 Rate variation across sites

All nucleotide substitution models assume that different sites in the DNA sequence evolve at the same rate and in the same way. However, DNA or protein sequences may evolve at different rates among different sites due to various factors such as natural selection, mutation rate variation, and structural or functional constraints. Therefore, allowing rate variation across sites is common because ignoring them may lead to underestimating the sequence distance.

One possibility to consider the rate variation across sites is to use a random variable, for example r, drawn from a statistical distribution. The most used distribution for such cases is the gamma distribution. The final models are represented by a suffix $+\Gamma$, e.g., $\mathrm{HKY} + \Gamma$.

The density function of the gamma distribution is

$$g(r;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} e^{-\beta r} r^{\alpha-1},$$

where $\alpha, \beta, r > 0$ and α and β are the scale parameters. The mean and variance are $E(r) = \alpha/\beta$ and $var(r) = \alpha/\beta^2$. To keep β for adjusting the variance of the distribution, we set $\alpha = \mathrm{E(r)}\beta$.

A common extension of this principle is to model a proportion of the sites to be so strongly constrained by natural selection that their substitution rate is zero. The resulting model is usually denoted $\mathrm{HKY} + \Gamma + \mathrm{I}$:

$$g(r;\alpha,\beta,\kappa) = \kappa\frac{\beta^{\alpha}}{\Gamma(\alpha)} e^{-\beta r} r^{\alpha-1} + (1-\kappa)\delta(\mathrm{r}),$$

where $\delta(\mathrm{r})$ is the delta function.

1.4 Phylogenetic Tree

A *phylogenetic tree* is a mathematical structure called a tree graph representing individuals' shared ancestry. The tree nodes (vertices) represent the common ancestors, while the branches (edges) represent evolutionary (ancestry) relationships between individuals (Yang, 2006). Such trees are widely used in biology with the primary goal of studying evolution. The trees can tell different stories depending on whether the research focuses on the long or short evolutionary timescales.

On long evolutionary timescales, phylogenetic trees can give insights to the evolution of highly conserved genes. For example, phylogenetic trees for genes involved in fundamental cellular processes or metabolism provide insights into the early evolutionary history of life and the relationships between major groups of organisms. (Hedges & Kumar, 2009; Wiley & Lieberman, 2011)

On short evolutionary timescales, the topology and the timing of splits can inform on past demographic processes, such as population expansions or contractions. They can help to understand, for example, the spread of early human populations across the globe. The construction of phylogenetic trees involves the usage of several mathematical methods. All of which require careful consideration of the assumptions and limitations of each method. (Wiley & Lieberman, 2011)

According to Wiley & Lieberman, 2011, the work to construct such trees can be organised in four steps:

1. The molecular data is aligned to identify corresponding (homologous) positions in the sequence for each taxon.
2. The evolution model with the theoretical molecular clock is chosen that describes the evolution of the data.
3. The method to construct the tree is chosen based on various criteria, such as the likelihood of the observed data given to the tree and the complexity of the tree.
4. The resulting tree is evaluated and compared to other possible trees to assess its accuracy and to infer evolutionary relationships among the taxa.

The following subsection briefly introduces the most well-known methods for constructing phylogenetic trees.

1.4.1 Methods to Construct the Phylogenetic Trees

Several methods to construct phylogenetic trees have been developed to find the tree that best fits the molecular data and represents the evolutionary relationships among the studied organisms. The four most well-known approaches are, in approximate order of increasing sophistication, Neighbour Joining, Maximum Parsimony, Maximum Likelihood, and Bayesian Inference.

The Neighbour-Joining (NJ) method constructs a tree based on the distance matrix of pairwise differences between molecular sequences. It starts by joining pairs of taxa to a star tree, where all the taxa are connected to the central node and then iteratively joins pairs closest in the distance until the result tree is obtained. The Neighbour-Joining method is often used when the relationships among data are complex, and the molecular data is highly divergent. Examples of software programs that use this approach are the MEGA and SplitsTree programs. (Huson & Bryant, 2006; Kumar et al., 2018; Saitou & Nei, 1987)

The Maximum Parsimony (MP) method aims to find the tree that requires the fewest changes in character states, like nucleotides or amino acids, to explain the observed data. This method is based on Occam's razor principle, which states that the simplest explanation is usually the best. The maximum Parsimony method is used when the relationships among the taxa are relatively simple. The PAUP* or TNT programs are examples of software programs that use this method. (Felsenstein, 2004; Wiley & Lieberman, 2011)

The Maximum Likelihood (ML) method seeks a tree that maximizes the likelihood of the observed molecular data, given a particular model of molecular evolution. This method is based on Bayesian statistics and assumes that the probability of observing the data is proportional to the likelihood of the tree under consideration. The Maximum Likelihood method is often used when the amount of molecular data is large, and the relationships among the data are complex. Examples of software programs that use this approach are the RAxML and PAML programs. (Stamatakis, 2014; Wiley & Lieberman, 2011; Yang, 2007)

The Bayesian Inference (BI) method aims to estimate the posterior probability distribution of the trees given the observed molecular data and a prior distribution of the trees. This method uses Bayes' theorem to update the prior probability distribution of the trees with the help of observed data resulting in a set of trees with associated probabilities. The Bayesian Inference method is often used when the amount of molecular data is moderate, and the relationships among the taxa are uncertain. The MrBayes and BEAST are examples of software programs

that use this method. (Drummond & Bouckaert, 2015; Felsenstein, 2004; Wiley & Lieberman, 2011)

All the above methods can be used to estimate the topology of the phylogenetic tree and estimate branch lengths, but to estimate the age of divergences between genetic sequences (in years or generations) it is necessary to calibrate the rate of evolution against reference points. There are two main approaches for this. The first, called *node calibration*, uses information about the age of an internal node to determine the evolutionary rate for the tree (often using archaeological information about the divergence of species as estimated by characteristic anatomical features). The second approach, called *tip-dating*, that has only recently become available with DNA from archaeological remains, is to use a panel of sequences from different, known, time points. Of the methods mentioned above, only the BI methods support this inference. Recent research suggest that tip-dating offers better accuracy for recently diverged sequences, such as when constructing phylogenies of sequences from different human populations (Rieux et al., 2014).

2. Materials and Methods

The key steps used to analyse ancient migration paths are the following:

1. Assembly and data quality analyses of mitochondrial sequences from human archaeological remains with matching geographic and temporal information.
2. Reconstruction of phylogenetic trees and estimation of split times informed by the dates of the archaeological samples.
3. Based on the phylogenetic trees, reconstruction of the geographic locations of the ancestors of the samples.

This chapter gives a theoretical overview of the used datasets and programs. It describes the initial steps of the analysis to construct phylogenetic trees and ancient migration paths based on such trees and the beginning steps of the analysis to construct such trees.

2.1 Mitochondrial Genomes from Ancient Individuals

The Allen Ancient DNA Resource (AADR) is a public database that provides a central repository for ancient DNA data. It was founded in 2014 by David Reich. Currently, the database contains genomes from more than 800 species and over 10,000 ancient DNA samples. (Mallick et al., 2023)

The dataset of mitochondrial genomes from ancient individuals used in this research is also from the AADR repository. Because the database is frequently updated, we use version 52.2 of the dataset, published on the 22nd of August 2022 (David Reich Lab, 2022). It consists of two files:

1. A file in fasta format (.fa) that consists of the genetic ID of the sample and the mitochondrial genome sequence.
2. The *Anno* file holds each sample's legend and background information, like latitude and longitude coordinates, sample age, mtDNA haplogroup and country of origin.

The dataset consists of 3,877 samples, all of which are present in the *Anno* file. Unfortunately, the file has duplicates of some samples since some of the mtDNA sequences have been sequenced multiple times, or the samples have been published in multiple research papers. After removing the duplicate samples (126), we can conveniently say the dataset has 3,751 unique samples.

Another aspect to consider is the quality of the data. Human genomes (e.g. mtDNA) tend to decay over time, and therefore the data sequences might not be complete. To ensure the high quality of the data and conduct reliable phylogenetic inference, we excluded mtDNA sequences with more than 10% loss outside of the D-loop (14,534 nucleotide positions, see Appendix 1 for beginning and end index of the D-loop) from the dataset. 157 samples were removed, resulting in a dataset of 3594 samples. The summary statistics of the age of each sample are represented in the following table (Table 1).

Table 1: Summary Statistics for the age of mtDNA samples in years

Mean	3 593
Median	3 407
Standard deviation	2 339
Min	60
Max	30 950

The table concludes that this data should be sufficient to perform the phylogenetic analysis of the samples and investigate ancient migration paths.

Another issue to be considered is how the samples are spread across the globe. The particular interest in the problem comes directly from the research question of investigating ancient migration paths. If the sample had missing coordinates in the *Anno* file, we tracked down the original research and successfully replaced the missing values with actual coordinates. The samples should be spread over the globe to get the most information on possible migrations. To justify the case, we plot the sample coordinates on the world map using statistical software R. As we can see from the world map, although the data is most abundant in Europe and Central Asia; Africa, Oceania, and the Americas are also represented, enabling a global analysis of possible migration paths of ancient people (Figure 1).

Figure 1: Geographical Locations of mtDNA samples.

2.2 Phylogenetic Reconstructions

2.2.1 The BEAST Program

We used the Bayesian Evolutionary Analysis Sampling Trees (BEAST) software program for the inference of phylogenetic trees. The main reason is that the program uses Bayesian inference, which suits well for moderate data in which taxon relationships are unknown. Additional property that justifies the approach is the possibility of using tip-date sampling. This approach makes it possible to include the ages of mtDNA sequence samples in the samples' ID.

The following sections of this paragraph are based on the Drummond & Bouckaert, 2015 book “Bayesian Evolutionary analysis with BEAST”.

Bayesian Evolutionary Analysis Sampling Trees (BEAST) is an open-source Markov chain Monte Carlo (MCMC) sampler for phylogenetic models. The program was first released in 2002 and developed by Alexei Drummond at the University of Auckland. The newest version of the program, BEAST 2, was released in 2014. It is supported by the BEAGLE library, which helps to increase computation power with optimised usage of multi-core processors.

The critical innovation of BEAST is its ability to simultaneously estimate the tree topology (i.e. the branching pattern of the evolutionary relationships), branch lengths (i.e. the amount of

evolutionary change along each branch), and the parameters of a molecular clock model (i.e. the rate of evolution and the age of the common ancestor) using an MCMC algorithm.

One of the strengths of BEAST is its ability to generate posterior distributions of parameters of interest, such as the age of the common ancestor or the rate of molecular evolution. This quantifies uncertainty in the estimated values, which is vital for robust evolutionary biology inference.

However, BEAST can be computationally intensive, especially for large datasets or complex models. It also requires expertise in Bayesian statistics and molecular evolution to interpret the results correctly.

The following list shows the contents of the BEAST software package and additional software tools to analyse BEAST output:

- **BEAST** – this package contains the BEAST program and a set of tools supporting it:
    - **BEAUTi** – program to generate BEAST .xml files specifying the data, its parameters and models used in the phylogenetic analysis.
    - **TreeAnnotator** – program to generate a consensus phylogenetic tree from BEAST output that represents the posterior distribution of the trees.
    - **Logcombiner –** program to combine log files produced by MCMC analysis.
- **Tracer** – program used to explore the output of BEAST. It summarises both graphically and quantitatively the distributions of continuous parameters and provides diagnostic information.
- **Figtree –** application for displaying and printing molecular phylogenies obtained using BEAST.

2.2.2 Reconstructing phylogenetic trees with BEAST

A *consensus tree* is a representation of a set of phylogenetic trees on the same set of taxa (Degnan et al., 2009). The *burn-in* is the number of samples it took for the MC calculations to approach equilibrium (*TreeAnnotator | BEAST Documentation*, n.d.).

To begin the reconstruction of phylogenetic trees, we first had to generate the BEAUTi file in Nexus format, which allows us to define parameters for models and distributions used in the analysis along with the sequence data to be analysed.

To use the tip-date sampling possibilities available in BEAST, the MasterID of the sample was modified to the format *MasterID@Date*. Using tip dates in the sample name allowed us to specify the sample's age later in BEAST without constructing taxon sets for the data.

We used the Analyses of Phylogenetics and Evolution (ape) library for statistical software R to write the nexus file from the data with new IDs with function write.nexus(). We wrote the data in sequential DNA format. In the mtDNA sequences, we marked the gaps with the "-" symbol and missing values with "n".

Since different genes in an individual's mtDNA genome have different mutation rates, we first partitioned the genome into groups of sites with similar expected mutation rates (Appendix 1). We excluded the D-loop from the analysis due to its high mutation rate.

In phylogenetics, *partitions* refer to the division of a dataset into non-overlapping subsets of characters (e.g., genes or codons) that are assumed to evolve independently or have different evolutionary models applied to them. The partitions are used to account for differences in substitution patterns among sites. (Kainer & Lanfear, 2015)

The protein encodings (CDS) had to be read in three partitions (starting from first, second and third codon position with a step of three), while complement proteins must follow the same rule but be read from back to front. The resulting protein partitions starting from indices 1 and 2 were merged. A control algorithm was constructed to avoid overlaps, where protein positions starting from 1 and 2 take priority over 3, and protein positions starting from 3 take priority over RNA positions. The RNA positions of the mtDNA were read from starting index to the end with a step of one. The generated partitions were appended to the Nexus file.

After importing the prepared data in the Nexus file format to BEAUTi, the model parameters to construct the phylogeny of the data needed to be specified. I used the tip-date sampling possibility and parsed the dates from the ID-s thanks to the format they were rewritten in. Additional specification before the present were used to mark each sample's age correctly.

No previously defined taxon sets were used for any model. The decision not to use any taxon sets was made because no prior knowledge about the different relationships between two random samples in the dataset was known.

As the previously defined protein and RNA partitions have different biological roles and functional constraints, they may evolve at different rates and their substitution and clock models where therefore unlinked. The tree models were kept unliked because our goal was to get one consensus tree. As suggested in previous work, we used the HKY+Γ+I site model for all partitions (the number of gamma categories was kept at the default 4), and assumed a strict molecular clock. The tree prior was chosen to be Bayesian Skyline with then groups, allowing for changes in the effective populations size in the history of sample.

The only prior distribution that was changed from the standard was the root height of the tree model. The prior distribution was changed to uniform, with the lower bound being close to the oldest sample in the dataset and the upper bound 500,000 to avoid the root height parameter from reaching infinity. In short, the root height of the tree was set to U(30,682; 500,000).

The length of the MCMC chain was increased ten times compared to the standard length for the calculations to reach equilibrium distribution. The final length of the chain produced was 100 million, with a parameter sampling step of 1,000. Lastly, the BEAST input file (in XML format) was generated using these model specifications.

The author of this thesis used the generated XML file with previously described model specifications to run BEAST. The BEAGLE package was used to increase the computational efficiency of BEAST. The BEAST run was done in the Institute of Genomics HPC cluster and project author's personal computer (Lenovo P52 with 8th generation Intel Core i7 processor).

The BEAST run produced the output file for the trees and log files. Due to the size of the file that holds the states of the tree and its inability to be processed by treeAnnotator to find a consensus tree, we used logcombiner to resample the tree states at a lower frequency.

The treeAnnotator was used to get one consensus tree inferred from multiple phylogenetic trees. The author specified the burnin as the number of states. The target tree was maximum clade credibility tree with median hights.

2.3 Ancestral Trails Framework

The following paragraph and its subparagraphs are based on (Eriksson et al., 2012) or seminars that the author of the thesis had with Dr Anders Eriksson during the period of October 2022 to May 2023.

The ancestral trails (AnTs) framework is a tool for studying species' evolutionary history by reconstructing their ancestors' ancestral geographic locations, developed by the supervisor, Dr Anders Eriksson.

One key aspect of the AnTs framework is the concept of movement through space as a random process. The movement through space is often modelled using a diffusion model, which assumes organisms move through space their environment randomly and unpredictably. This model can be applied to simple cases, such as homogeneous diffusion on a sphere or plane, or more complex scenarios incorporating geography, terrain, and climate.

For example, by considering factors such as land vs sea/water, mountain chains, deserts and arctic regions, researchers can build more sophisticated models of how species might have moved through their environment over time. These models can shed light on the geographic distribution of species that can be overlooked by phylogenetic analysis alone.

Bayesian inference is one way to combine these models with the phylogenetic tree. This approach uses probabilistic models to estimate a given species' most likely ancestral locations based on its evolutionary history and environmental factors, which might have influenced its movement.

The usage of ancestral trails framework and Bayesian inference can help researchers to better understand the movement of species and their evolution. This helps to shed new light on various research areas such as the species' adaptation to changing environments and the maintenance and generation of biodiversity.

2.3.1 Estimating Past Geographical Locations

For each lineage in the tree, starting from the present and moving backwards in time, AnTs calculate the probability of finding a lineage in a given location at a given time based on the modelled background migration process. When pairs of lineages (A and B) have a most recent common ancestor, the probability for this ancestor to be in hexagon i is

$$P_i^{AB} = \frac{1}{N_i} P_i^A P_i^B / Z \,,$$

where Z is a normalisation such that the sum over all locations is unity. The calculation of the probability distribution of the joint lineage then proceeds from this distribution. This way, the probability distributions of locations are calculated for every internal node in the tree.

To sample spatiotemporal trajectories consistent with these distributions (and the tree), AnTs starts from the root and picks a location according to the distribution at the root. Conditional on this, the trajectories for each descendant lineage become statistically independent. To sample the locations of an immediate descendant A of the root, AnTs draws the location i randomly with a probability proportional to $P_i^A M_{ij}$ where j is the location of the root and M_{ij} is the probability of migrating from j to i in one generation.

This procedure is only correct if there is no interaction between lineages until they coalesce (according to the observed tree). To correct for this, AnTs offers a rejection sampling approach, based on the probability of no coalescence given the sampled trajectories. The product of $1 - 1/K_i$ gives this probability over all generations in which the lineages are in the same hexagon at the same time. If more than two lineages are present in the same hexagon the calculation is adjusted accordingly. These probabilities are then combined across the tree to yield the joint probability of observing the tree (i.e., not having any extra coalescent events), which are used for rejection sampling and to calculate weighted averages.

2.3.2 Spatially Explicit Coalescent Background Model

The spatial demographic model is based on (Eriksson et al., 2012) and further developed in (Raghavan et al., 2015). The world is divided into hexagons ~100 km wide, represented as a graph in which hexagons are nodes and links between neighbouring hexagons are edges. Each cell can be either land or uninhabitable (sea or ice), which can change over time, informed by bathymetry and reconstructions of past sea levels and ice sheets during glacial periods.

Land hexagons have a maximum population size K, linked to net primary productivity (NPP). NPP is a measure of total plant and animal productivity estimated as a function of annual mean temperature and precipitation. The population size is zero below a lower threshold $\mathrm{NPP}_{\mathrm{min}}$, grows linearly so that it reaches K at the upper threshold $\mathrm{NPP}_{\mathrm{max}}$, and stays equal to K for larger values of NPP. The values for K, $\mathrm{NPP}_{\mathrm{min}}$ and $\mathrm{NPP}_{\mathrm{max}}$ were taken from Eriksson et al.

(2012), where NPP values were reconstructed from global circulation models of past climate for the last 120 000 years. The parameters were fitted to genetic data from diverse ancestry panel of populations around the world (the HGDP panel).

The model assumes local random mating within each land hexagon and random migration to neighbouring hexagons in each generation. Neighbouring land hexagons (A and B) exchange migrants with the following rate:

$$\text{migrant exchange rate} = m\frac{K_A K_B}{K_A + K_B},$$

where the parameter m sets the scale of migration. In each generation, lineages in the same population can have a common ancestor with a probability of $1/K$. The value of K varies across space and time, depending on NPP.

3 Results

3.1 Reconstruction of Dated Phylogenetic Tree

The author explored different models of sequence evolution suggested by the literature. It involved changing the parameters of the previously defined model (increasing and decreasing of different group sizes), trying different substitution models (e.g., TN93), experimenting with different clock types (e.g., relaxed clock) and using different tree priors (e.g., Bayesian Skygrid and constant size) However, the problems with the convergence of the model remained. Additionally, the running times of the program increased with no significant improvement in the output quality.

Considering these problems, it was decided to reduce the model complexity and dataset size. The size of the dataset was fixed at 400 samples. To create a dataset balanced in temporal and spatial distribution, all samples from Africa (143 samples) and those over 7500 years old (198 samples) were included in the analysis. The remaining samples were picked from the remaining samples with simple random sampling without replacement (59 samples). The geographical locations of the samples are represented on a world map (Figure 2). Compared to the first model represented in paragraph 2.2.2, the number of groups of the tree prior Bayesian skyline was reduced to 5.

Figure 2. Geographical Locations of mtDNA samples in the Final Dataset

After 5.4 days of running time in the Molecular Institutes cluster and 4.8 days in the author's personal computer, the program finished its calculations. The calculations were finished with tree and log files.

Comparing the log file's clock rates with the work of Rieux et al., 2014, which used 350 ancient and modern human mtDNA genomes to investigate reliable rate estimates for DNA substitution rate, shows that the substitution rates of the named analysis are lower than us (Table 2). The difference is also statistically significant, since 95% confidence HPD intervals do not overlap. This means additional rescaling to the spatial reconstruction of ancient migration paths should be done to get a better overview of possible results.

Table 2. Comparison of Molecular Clock Rate Parameter Estimates by Partition between Mandel and Rieux et al., 2014

		Substitution Rates (μ/Site/Year) (Units of 10^{-8})					
		Mandel			Rieux et al., 2014		
		PC1+PC2	PC3	RNA	PC1+PC2	PC3	RNA
Best Estimate		1.576	5.066	1.935	0.756	3.323	1.007
95% HPD	Lower	1.275	4.167	1.523	2.568	0.757	0.757
	Upper	1.893	6.012	2.381	4.074	1.266	1.266

Due to the size of the tree file and the treeAnnotators inability to find a consensus tree, we used log-combiner to resample the tree states at a lower frequency. The author increased the parameter sampling step from 1000 to 2000 while keeping the original burn-in of the data.

We then used treeAnnotator with a specified burn-in of 6 million states to get the consensus tree. Our target tree was a maximum clade credibility tree with median heights. The consensus tree file was then modified to add the samples' geographical locations to the phylogenetic tree leaves. This allowed us to inspect the consistency of the tree.

Examining the tree led us to discover two inconsistencies in the tree. The tree's topology was consistent with official haplotype assignments, with few exceptions. Two samples found in nowadays Russia (genetical IDs I1193 and I7455), with haplogroup C5b, were nested in a clade consisting of samples assigned to the sister haplogroup C1b, all of which are found in South America. This phenomenon could be caused by the problem within the dataset regarding these samples or by some program-related problem responsible for the tree's reconstruction. Due to

the inconsistency between the official and placement mtDNA haplogroup, the author removed these two samples using the APE package in the statistical software R.

The resulting final consensus tree had 398 tips. At closer inspection, this tree was found to have nine internal nodes with negative branch lengths (negative branch lengths ranging 29-1,000 years). According to Bagley, 2016 it is a problem that people encounter from time to time while conducting phylogenetic analysis, which substantially affects the analysis made on that tree. An approach suggested by Bagley, 2016 is to use the R software's APE package that allows resolving these problems using functions used to solve the problem in this thesis. However, any approach mentioned in the article comes with the same cost; the tip-to-node distance is shortened for the adjacent branch. Luckily, this should be fine for our migration path reconstruction.

After fixing negative branch lengths, the author plotted the final consensus tree with a program called Figtree (Appendix 2). The tree's tip labels (leaves) are formatted as the country where the sample was found. For example, Sudan, written in red, refers to this specific sample in the tree found in Sudan and red to Africa, the region where that country is located. The branch lengths are in years, specified as the time before the present. The internal nodes are theoretical maternal ancestors of the species up to the root node, which should be the maternal ancestor for all the samples in the dataset.

3.2. Spatial Reconstruction

To visualise ancient migration paths, the phylogenetic tree was plotted in 3D space (consisting of longitude, latitude, and time).

Two such trees were generated. The first used the original parameters of the dataset (Figure 3). The second was constructed by the result that the molecular clock rates in this thesis are higher than those proposed by Rieux et al., 2014 (for justification, see Table 2 in par. 3.1). Therefore, divergence times in the original phylogeny were rescaled (by factor of 1.45) to approximately match the lower evolutionary rate of Rieux et al., 2014, and AnTs was used to estimate the ancestral paths based on the rescaled phylogenetic tree. The resulting plot is a stretched-out version of the original one, with changes in the scale of the *y*-axis, some moving trajectories, and split locations of some migrating groups (Figure 4).

The samples were plotted on a world map where the x-axis represents the latitude and y-axis longitude coordinates. The z-axis represents the node's age in the phylogenetic tree in years. The years follow the *time before the present* logic. The only fixed nodes in the tree are the leaf nodes of which we know the background information. The internal nodes and the root node locations are assumed considering the principles of the AnTs framework (e.g., the climate of the time and sea levels).

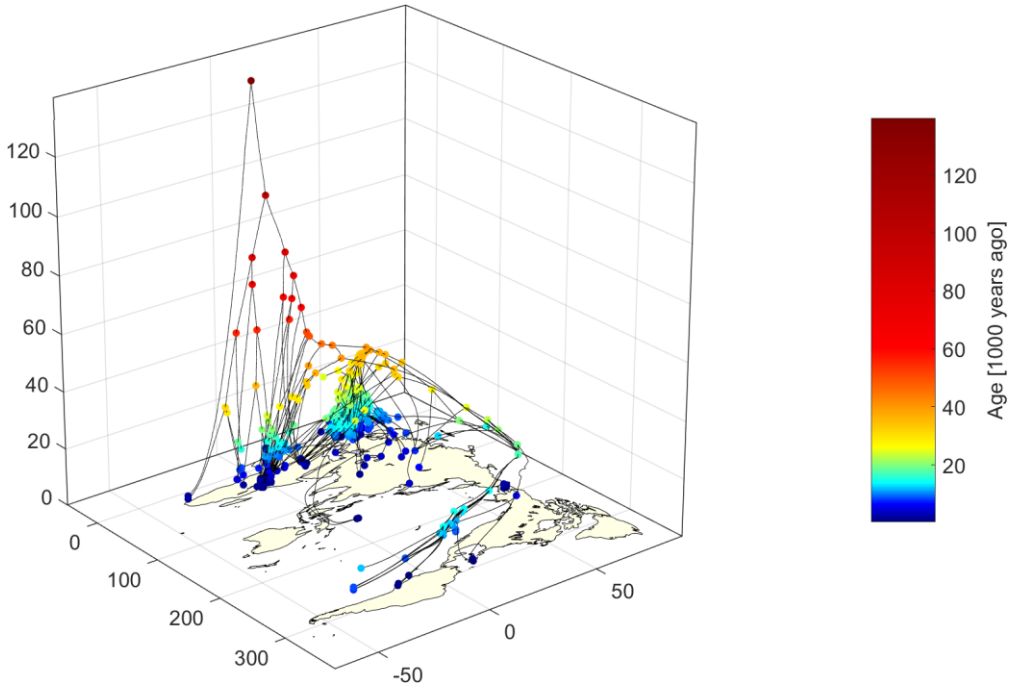


Figure 3. The Original Phylogenetic Tree in 3D Space. Vertical axis, time before present (t) in units of 1,000 years. Filled circles show the location of internal nodes of the phylogenetic tree (coloured by age, see colour bar to the right of the plot), and black lines are the edges in the tree. For reference, land contours are shown For reference, land contours are shown in light yellow in the plane of $t = 0$.

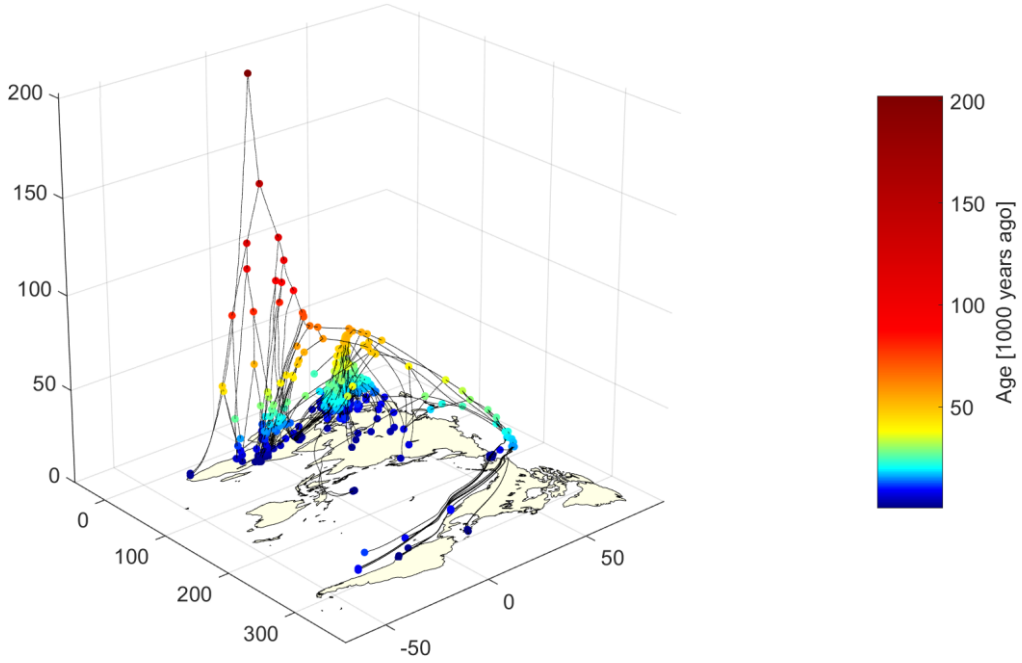


Figure 4. The Rescaled Phylogenetic Tree in 3D Space, as in Figure 3. Not the longer time scale.

From both 3D plots, we can see the main historical migration events of the human species. For example, the emigration from Africa to Eurasia is clearly visible (the orange node on top of Europe). It can be observed that the same group of people then diverged into subgroups. One of the groups settled in Europe, and the other started their expansion towards Eastern Eurasia and the Americas. Both figures illustrate well how the new founders, benefitting from the low sea levels at the time, crossed the land bridge between Eurasia and North America to colonise the new land from the top of the continent to the bottom of it.

The similarities between the original and rescaled phylogenetic trees continue as the author observes more simplistic migration paths. It is visible that Japan was colonised from Eastern Eurasia by crossing the sea between them. Another example is the migration to Micronesia, where ancient humans crossed islands and seas in Oceania. Both plots suggest that most people living in Africa are the descendants of people who have lived on the continent for tens of thousands of years and never left Africa.

The main differences between these models arise when we look at them in greater detail, trying to explore the scenarios and timings of such expansions. The original tree dates the expansion out of Africa 40,000-50,000 years back, but the rescaled tree states that it happened around 60,000-70,000 years ago. The same difference is related to expansion to America, which

according to the original tree, started 15,000 years ago, whereas rescaled tree dates the start of the journey 20,000 years to the past.

Another distinguishable difference is the scenario when and where the group of people who colonised the Americas split. The original model locates the group's split to Central America (10,000 years ago). Meanwhile, while exploring the rescaled plot, the split is positioned to Canada and Alaska region, dating it approximately 19,000 years ago, right after the colonisation of the North-West part of the continent.

The parameters and assumptions of the AnTs framework justify such a geographical shift in the split place. Since most of Northern America was covered by ice until 15,000-16,000 years ago (Raghavan et al., 2015), people could not continue their journey towards Central America. The phylogenetic analysis, however, has determined that the split of these people has happened genetically; therefore, it is placed in a location populated by the same group of people. Small shifts between the trajectories of the plots can also be seen elsewhere, so they go hand in hand with the climate at the time.

To investigate whether the author of the thesis has managed to rediscover previous findings of other academic papers, a small comparison between the findings of this thesis and academically published papers was made. It should be remembered that the timings of such significant events are of great debate.

Comparing the results with Fu et al., 2013, we see that the separation of non-Africans from the sub-Saharan African mitochondrial DNAs happened less than 62,000 to 95,000 years ago. The separation dates from the analysis of this thesis range from 40,000 to 70,000 years ago, which overlap the lower end of the range found by Fu et al., 2013 and are therefore broadly consistent.

According to Moreno-Mayar et al., 2018, emigration to the Americas began approximately 21,000 years ago using the land bridge between Eurasia and Northern America, after which the split in the founders' group happened in Northern America around 15,500 years ago. The author understands that the close resemblance of the dates should be taken with a grain of salt, but the fact that the split and expansion existed in both models, somewhat approves the existence of such a migration path.

Jinam et al., 2012, and Liu et al., 2022, support the migration paths to Japan from Eastern Asia and to Micronesia from Japan. The study by Schuster et al., 2010 on Khoi and Bantu-speaking populations in Africa agrees that some African populations have genetic markers indicating long-standing ancestry in the region.

Conclusion

The main objective of this thesis was to assemble a global dataset of the whole mitochondrial genome sequences (mt genomes) from present-day and ancient populations. Then estimate the phylogenetic relationships and divergence dates between the mt genomes and finally identify the timings and trajectories of major (pre) historical population movements and human habitats that were key genetic sources to present-day populations.

To reach these objectives, the author used Alen Ancient DNA Resource to obtain the original data of ancient mtDNA sequences, statistical software R to prepare and modify the dataset of 400 samples and the programme BEAST to generate the phylogenetic relationships between individuals. The author of this thesis used the strict clock model with the Bayesian Skyline tree prior. The first phylogenetic tree was constructed with the help of FigTree. The population models, both in 3D and video format, were generated from MATLAB codes by Dr Anders Eriksson as part of the AnTs framework.

The thesis' main results were that the trees, and mutation model parameters obtained were quite robust and broadly consistent with previously published analyses. The uncertainty lays in the clock rates, which were higher than previous studies have suggested. Nevertheless, the author of this thesis achieved most of his goals. It was discovered that the uncertainty in time does not affect spatial reconstructions but the timings of the events. Even though the dates of the events were unlikely compared to previously published academic papers, we could still inspect the emigration out of Africa and the colonisation of the Americas.

Additionally, smaller migrations like colonising Japan or Micronesia were distinguishable events. Moreover, we could see the evidence that sampled lineages constrain the inferred ancestral locations. However, lowering the uncertainty in the future would allow us to inspect more accurate timings (e.g., out-of-Africa emigration, expansion to America) and geographic locations of the splits (e.g., Northern American immigrants' split).

The author's suggestion to the people willing to conduct similar analyses would be to find a way to scale the size of the dataset up to the entire AADR mtDNA database and combine it with available modern samples. To increase the accuracy of the parameters, longer computational times are a must. Checking that the parameters converge helps to have a more reliable analysis in the future. Additionally, comparing the spatial patterns inferred from different non-combining markers (e.g., Y chromosome) would give new insights into how

people might have migrated in the past. If this could be achieved, the researcher could have a much higher resolution for the spatial analysis and a tighter tree to investigate even more.

References

Bagley, J. (2016, March 1). *Dealing with negative phylogenetic branch lengths in BEAST starting trees—Blog*. https://justinbagley.rbind.io/2016/03/01/dealing-negative-phylogenetic-branch-lengths-beast-starting-trees/ (06.04.2023).

Bandelt, H.-J., Richards, M., & Macaulay, V. (2006). Mitochondrial DNA in Homo Sapiens. In *Human Mitochondrial DNA and the Evolution of Homo Sapiens* (1st ed, pp. 6–8). Springer.

David Reich Lab. (2022). *Downloadable genotypes of present-day and ancient DNA data.* (52.2) [Data set]. Allen Ancient DNA Resource. https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data. (10.11.2022).

Degnan, J. H., DeGiorgio, M., Bryant, D., & Rosenberg, N. A. (2009). Properties of Consensus Methods for Inferring Species Trees from Gene Trees. *Systematic Biology*, ***58*(1)**, 35–54. https://doi.org/10.1093/sysbio/syp008 (01.04.2023).

Doimo, M., Pfeiffer, A., Wanrooij, P. H., & Wanrooij, S. (2020). MtDNA Replication, Maintenance, and Nucleoid Organization. In G. Gasparre & A. M. Borcelli, *The Human Mitochondrial Genome From Basic Biology to Disease* (pp. 3–22). Elsevier.

Drummond, A. J., & Bouckaert, R. R. (2015). *Bayesian Evolutionary Analysis with BEAST* (1st ed.). Cambridge University Press.

Eriksson, A., Betti, L., Friend, A. D., Lycett, S. J., Singarayer, J. S., von Cramon-Taubadel, N., Valdes, P. J., Balloux, F., & Manica, A. (2012). Late Pleistocene climate change and the global expansion of anatomically modern humans. *Proceedings of the National Academy of Sciences*, ***109*(40)**, 16089–16094. https://doi.org/10.1073/pnas.1209494109 (22.04.2023).

Felsenstein, J. (2004). *Inferring Phylogenies* (2nd ed.). Sinauer Associates.

Fu, Q., Mittnik, A., Johnson, P. L. F., Bos, K., Lari, M., Bollongino, R., Sun, C., Giemsch, L., Schmitz, R., Burger, J., Ronchitelli, A. M., Martini, F., Cremonesi, R. G., Svoboda, J., Bauer, P., Caramelli, D., Castellano, S., Reich, D., Pääbo, S., & Krause, J. (2013). A revised timescale for human evolution based on ancient mitochondrial genomes. *Current Biology : CB*, ***23*(7)**, 553–559. https://doi.org/10.1016/j.cub.2013.02.044 (15.04.2023)

Hedges, S. B., & Kumar, S. (2009). Discovering the Timetree of Life. In S. B. Hedges & S. Kumar, *The Timetree of Life* (1st ed., pp. 3–18). Oxford University Press.

Homo sapiens mitochondrion, complete genome (251831106). (2023). [Data set]. NCBI Nucleotide Database. http://www.ncbi.nlm.nih.gov/nuccore/NC_012920.1 (15.11.2022).

Huson, D. H., & Bryant, D. (2006). Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution*, ***23*(2)**, 254–267. https://doi.org/10.1093/molbev/msj030 (14.03.2023)

Jinam, T., Nishida, N., Hirai, M., Kawamura, S., Oota, H., Umetsu, K., Kimura, R., Ohashi, J., Tajima, A., Yamamoto, T., Tanabe, H., Mano, S., Suto, Y., Kaname, T., Naritomi, K., Yanagi, K., Niikawa, N., Omoto, K., Tokunaga, K., … Japanese Archipelago Human Population Genetics Consortium. (2012). The history of human populations in the Japanese Archipelago inferred from genome-wide SNP data with a special reference to the Ainu and the Ryukyuan populations. *Journal of Human Genetics*, ***57*(12)**, Article 12. https://doi.org/10.1038/jhg.2012.114 (23.04.2023).

Kainer, D., & Lanfear, R. (2015). The Effects of Partitioning on Phylogenetic Inference. *Molecular Biology and Evolution*, ***32*(6)**, 1611–1627. https://doi.org/10.1093/molbev/msv026 (23.04.2023).

Kumar, S. (2005). Molecular clocks: Four decades of evolution. *Nature Reviews. Genetics*, ***6*(8)**, 654–662. https://doi.org/10.1038/nrg1659. (22.04.2023)

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution*, ***35*(6)**, 1547–1549. https://doi.org/10.1093/molbev/msy096 (21.03.2023)

Lipson, M., Cheronet, O., Mallick, S., Rohland, N., Oxenham, M., Pietrusewsky, M., Pryce, T. O., Willis, A., Matsumura, H., Buckley, H., Domett, K., Hai, N. G., Hiep, T. H., Kyaw, A. A., Win, T. T., Pradier, B., Broomandkhoshbacht, N., Candilio, F., Changmai, P., … Reich, D. (2018). Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science (New York, N.Y.)*, ***361*(6397)**, 92. https://doi.org/10.1126/science.aat3188 (21.03.2023)

Liu, Y.-C., Hunter-Anderson, R., Cheronet, O., Eakin, J., Camacho, F., Pietrusewsky, M., Rohland, N., Ioannidis, A., Athens, J. S., Douglas, M. T., Ikehara-Quebral, R. M.,

Bernardos, R., Culleton, B. J., Mah, M., Adamski, N., Broomandkhoshbacht, N., Callan, K., Lawson, A. M., Mandl, K., … Reich, D. (2022). Ancient DNA Reveals Five Streams of Migration into Micronesia and Matrilocality in Early Pacific Seafarers. *Science (New York, N.Y.)*, ***377*(6601)**, 72–79. https://doi.org/10.1126/science.abm6536 (22.03.2023).

Mallick, S., Micco, A., Mah, M., Ringbauer, H., Lazaridis, I., Olalde, I., Patterson, N., & Reich, D. (2023). The Allen Ancient DNA Resource (AADR): A curated compendium of ancient human genomes. *BioRxiv*, 2023.04.06.535797. https://doi.org/10.1101/2023.04.06.535797 (07.04.2023).

Moreno-Mayar, J. V., Vinner, L., de Barros Damgaard, P., de la Fuente, C., Chan, J., Spence, J. P., Allentoft, M. E., Vimala, T., Racimo, F., Pinotti, T., Rasmussen, S., Margaryan, A., Iraeta Orbegozo, M., Mylopotamitaki, D., Wooller, M., Bataille, C., Becerra-Valdivia, L., Chivall, D., Comeskey, D., … Willerslev, E. (2018). Early human dispersals within the Americas. *Science (New York, N.Y.)*, ***362*(6419)**, eaav2621. https://doi.org/10.1126/science.aav2621 (07.04.2023).

Raghavan, M., Steinrücken, M., Harris, K., Schiffels, S., Rasmussen, S., DeGiorgio, M., Albrechtsen, A., Valdiosera, C., Ávila-Arcos, M. C., Malaspinas, A.-S., Eriksson, A., Moltke, I., Metspalu, M., Homburger, J. R., Wall, J., Cornejo, O. E., Moreno-Mayar, J. V., Korneliussen, T. S., Pierre, T., … Willerslev, E. (2015). Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science (New York, N.Y.)*, ***349*(6250)**, aab3884. https://doi.org/10.1126/science.aab3884 (22.04.2023).

Rieux, A., Eriksson, A., Li, M., Sobkowiak, B., Weinert, L. A., Warmuth, V., Ruiz-Linares, A., Manica, A., & Balloux, F. (2014). Improved Calibration of the Human Mitochondrial Clock Using Ancient Genomes. *Molecular Biology and Evolution*, ***31*(10)**, 2780–2792. https://doi.org/10.1093/molbev/msu222 (02.05.2023).

Rodríguez-Trelles, F., Tarrío, R., & J.Ayala, F. (2003). Molecular clocks: Whence and whither? In P. C. J. Donoghue & M. P. Smith (Eds.), *Telling the Evolutionary Time. Molecular Clocks and the Fossil Record.* (1st ed., pp. 5–26). CRC Press.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, ***4*(4)**, 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454 (29.04.2022)

Soares, P., Rito, T., Trejaut, J., Mormina, M., Hill, C., Tinkler-Hundal, E., Braid, M., Clarke, D. J., Loo, J.-H., Thomson, N., Denham, T., Donohue, M., Macaulay, V., Lin, M., Oppenheimer, S., & Richards, M. B. (2011). Ancient Voyaging and Polynesian Origins. *American Journal of Human Genetics*, ***88*(2)**, 239–247. https://doi.org/10.1016/j.ajhg.2011.01.009. (19.03.2023).

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, ***30*(9)**, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033 (08.03.2023).

Torroni, A., Achilli, A., Olivieri, A., & Semino, O. (2020). Haplogroups and the History of Human Evolution Through mtDNA. In G. Gasparre & A. M. Borcelli, *The Human Mitochondrial Genome From Basic Biology to Disease* (pp. 111–130). Elsevier.

TreeAnnotator | BEAST Documentation. (n.d.). https://beast.community/treeannotator. (09.04.2023).

Wiley, E. O., & Lieberman, B. S. (2011). *Phylogenetics: Theory and Practice of Phylogenetic Systematics* (2nd ed.). Wiley-Blackwell.

Yang, Z. (2006). *Computational Molecular Evolution*. Oxford University Press.

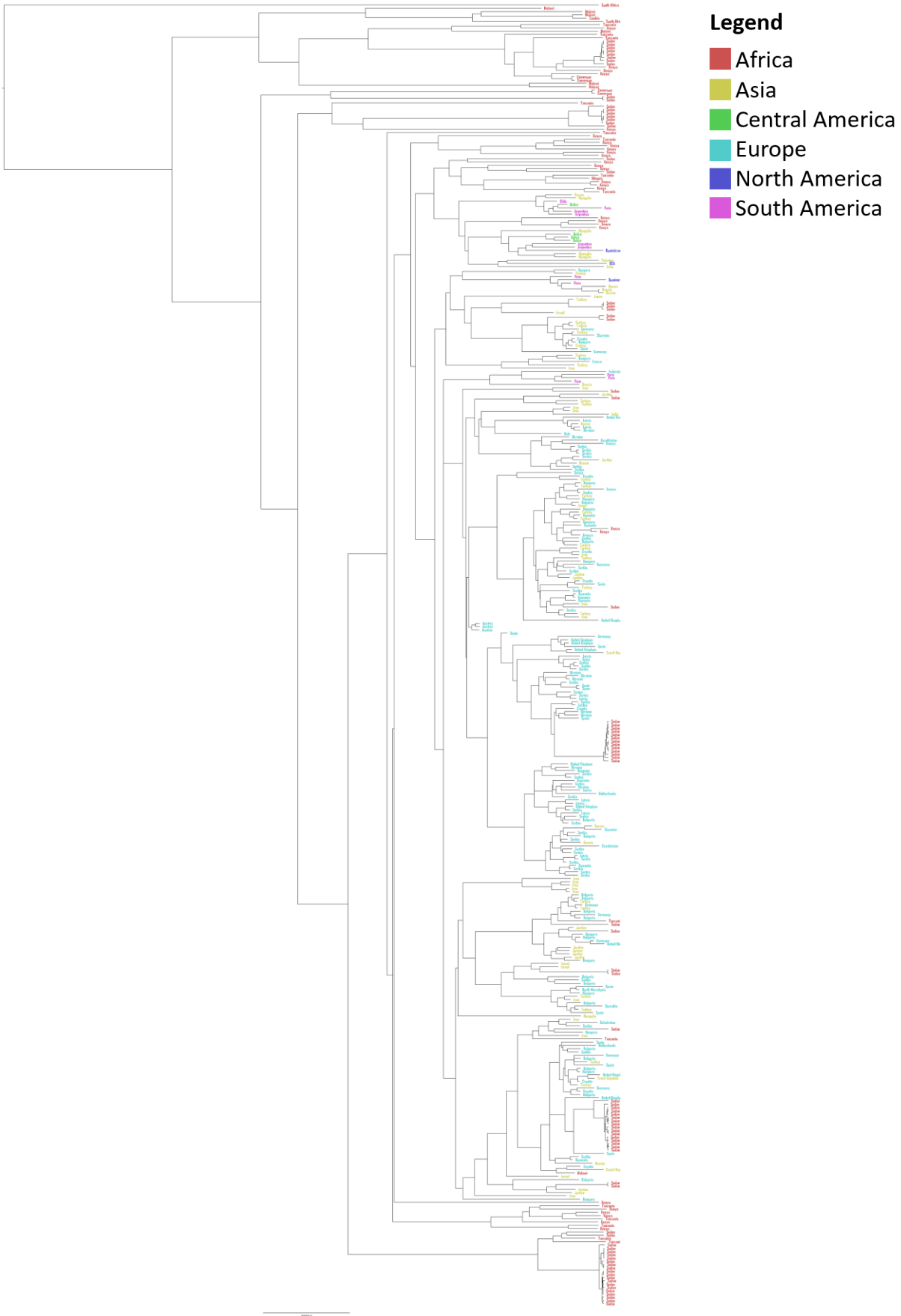
Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, ***24*(8)**, 1586–1591. https://doi.org/10.1093/molbev/msm088 (05.05.2023)

Appendix 1. Complete Genome of Homo Sapiens Mitochondrion

The following table is based on the source (*Homo Sapiens Mitochondrion, Complete Genome*, 2023)

Name	Type	Starting index	End index	Complement
-	D-loop	1	576	+
TRNF	tRNA	577	647	-
RNR1	rRNA	648	1 601	-
TRNV	tRNA	1 602	1 670	-
RNR2	rRNA	1 671	3 229	-
TRNL1	tRNA	3 230	3 304	-
ND1	CDS	3 307	4 262	-
TRNI	tRNA	4 263	4 331	-
TRNQ	tRNA	4 329	4 400	+
TRNM	tRNA	4 402	4 469	-
ND2	CDS	4 470	5 511	-
TRNW	tRNA	5 512	5 579	-
TRNA	tRNA	5 587	5 655	+
TRNN	tRNA	5 657	5 729	+
TRNC	tRNA	5 761	5 826	+
TRNY	tRNA	5 826	5 891	+
COX1	CDS	5 904	7 445	-
TRNS1	tRNA	7 446	7 514	+
TRND	tRNA	7 518	7 585	-
COX2	CDS	7 586	8 269	-
TRNK	tRNA	8 295	8 364	-
ATP8	CDS	8 366	8 572	-
ATP6	CDS	8 527	9 207	-
COX3	CDS	9 207	9 990	-
TRNG	tRNA	9 991	10 058	-
ND3	CDS	10 059	10 404	-
TRNR	tRNA	10 405	10 469	-
ND4L	CDS	10 470	10 766	-
ND4	CDS	10 760	12 137	-
TRNH	tRNA	12 138	12 206	-
TRNS2	tRNA	12 207	12 265	-
TRNL2	tRNA	12 266	12 336	-
ND5	CDS	12 337	14 148	-
ND6	CDS	14 149	14 673	+
TRNE	tRNA	14 674	14 742	+
CYTB	CDS	14 747	15 887	-
TRNT	tRNA	15 888	15 953	-
TRNP	tRNA	15 956	16 023	+
-	D-loop	16 024	16 569	+

Appendix 2. Phylogenetic Tree



Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Erik Mandel,

1. Annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose ”Reconstructing Ancient Migration Paths from Mitochondrial Genomes”, mille juhendaja on Jon Anders Eriksson, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.
2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.
3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.
4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Erik Mandel

09.05.2023