Inferences of Historic Population Sizes from DNA Sequence Data

Rishabh Bahl

14/08/2020

Contents

1 Introduction

The remarkable growth of human population from a stable size of few million to several billions took place in only a few thousand years (Max Roser et al. 2013). It is an extremely short period compared to the evolutionary or geological time scales. In the last 150 years, there is an explosive growth in the human population from one to almost five billion that marks a significant change in the population size through time (Smil et al., 1999).

Demography, the study of human population dynamics, is a fundamental part of human ecology. Demographic analysis helps in exploring the change in population size across time through processes of birth, death and migration. It is not limited to human population only, but extends to divergences across various species of the biological world. The demographic analysis is used in a wide variety of contexts such as immigration of individuals to different parts of the world resulting in a huge expansion of the world population, evolutionary dynamics of diseases caused due to certain viruses like human influenza (flu) in 2009 caused due to influenza A virus subtype H1N1 (Smith et al., 2009).

In the context of human biological populations, the analysis is carried out using a diverse sample of human population in order to obtain independent estimates of the population size.
The population size estimates for human population are fundamental for understanding the ecological, behavioural, or genetic processes evolving across time (Ojaveer et al. 2004; Dochtermann and Peacock 2013; Valderrama et al. 2013). The studies in population genetics provide evidences of migrations of modern humans from Africa to different parts of the world. One of the significant movements was along the coast of Asia and reaching Australia by around 50,000–65,000 years ago causing a rapid expansion in population growth of the world (Clarkson C et al., 2017). The phylogenetic analysis involves the application of mitochondrial DNA (mtDNA; Cann et al. 1987), Y-chromosome (Underhill et al. 2000) or the autosomal evidence (Gonser et al. 2000) to draw inferences of the demographic history of the population.
Past population sizes can be estimated from genetic data using coalescent theory (Hudson 1990). For any sample of DNA, the information that can be extracted about population size depends on the accuracy with which the underlying genealogical tree can be resolved and the extent to which the historical demographic processes can be accommodated by the demographic model.

1.1 Background & Context

1.1.1 Genetics

Genetics is a branch of biology concerned with the study of genes, genetic variation, and heredity in organisms. A gene is a sequence of nucleotides in DNA or RNA that encodes the synthesis of a gene product, either RNA or protein. The molecular basis for genes is deoxyribonucleic acid (DNA). DNA is composed of a chain of four types of nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). Genetic information exists in the sequence of these nucleotides, and genes exist as stretches of sequence along the DNA chain. DNA exists as a double-stranded molecule, coiled into the shape of a double helix. Each nucleotide in DNA preferentially pairs with its partner nucleotide on the opposite strand: A pairs with T, and C pairs with G. The structure of DNA is the physical basis for inheritance as the two stranded structure contains all necessary information about an individual.
The genes are transmitted from an organism to the offspring at the time of birth. The offspring inherit traits from the parents. Genes can acquire mutations during the transmission of the gene sequences, leading to

different variants in the population, called alleles. These alleles are slightly different from the parents due to variation in the information carrying proteins causing different phenotypical traits in parents and offspring. Sometimes during the process of DNA replication from the parent to offspring, changes occur in sequence of the DNA. These changes are called mutations and can affect the phenotype of an organism, especially if they occur within the protein coding sequence of a gene. Mutations are changes that occur in the nucleotide sequence of DNA at the time of DNA replication from the parent to offspring. They have other important implications. They create variation within the genes in a population. The DNA sequence consists of two regions: Coding Region and Controlled Region. The coding region of a gene is the portion of a gene's DNA that codes for protein and suffers higher mutation rate than controlled region (ref). Controlled region is the other portion of the DNA sequence and experience extremely variant substitution rates between sites, causing difficulties in the estimation of the genetic distance (Sigurğardóttir S et al. 2000). Mutations in the coding region can have very diverse effects on the phenotype of the organism. There are various advantages of mutations such as resistance to HIV disease or better vision compared to the earlier generations. It may have disadvantages like inherited genetic orders such as Cystic fibrosis, Marfan syndrome etc.
There are various forms of mutations that can occur in coding regions. A substitution mutation is a type of replication error during DNA replication which places another nucleotide or sequence of nucleotides in one position. Other types of mutations include insertions or deletions. Insertions are mutations in which extra base pairs are inserted into a new place in the DNA. Deletions are mutations in which a section of DNA is lost, or deleted. Insertions/Deletions change the length of the DNA sequence.

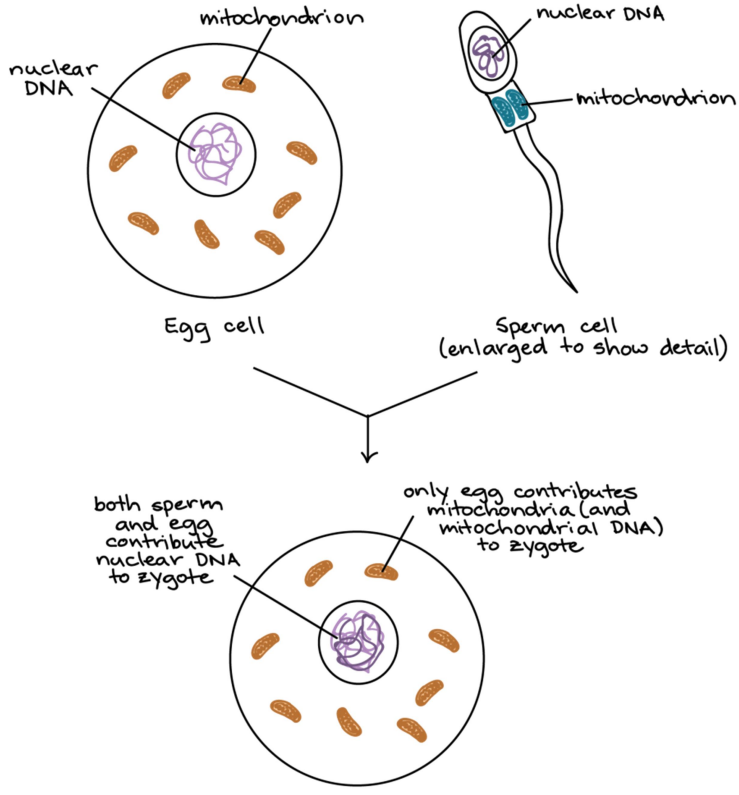


Figure 1.1 The nuclear DNA from both the Egg Cell (Mother) and Sperm Cell (Father) is inherited to the Zygote (Child). The Mitochondrial DNA on the other hand, is only inherited by the Egg Cell (Mother) to the Zygote (Child). (figure from Mitochondrial inheritance and chloroplast DNA, Khan Academy)

1.1.2 Mitochondrial DNA

In medical science, mtDNA sequencing can be used in the identification of genes responsible for hereditary disorders (Schaefer AM et al., 2004). Mitochondrial DNA (mtDNA) is the small circular chromosome found inside mitochondrion of the cell. The mitochondria are organelles found in cells that are the sites of energy production. mtDNA helps in generating energy to power the cell's biochemical reactions by converting chemical energy from food to adenosine triphosphate (ATP). Therefore, Mitochondrion is called the "powerhouse of the cell". In most multicellular organisms, mtDNA is inherited from the mother (maternally inherited) unchanged (R.E. Giles et. al 1980), to all her children, both male and female (see fig. 1). Maternal inheritance enables us to trace maternal lineage far back in time. This is more difficult to achieve using DNA as it is inherited from both the parents and is present in the nucleus of the cell. mtDNA also has a higher mutation rate than nuclear DNA which makes it useful for studying the evolutionary relationships of organisms (Cathy H et al. 2008). mtDNA has been a crucial line of evidence in developing the current understanding of our genetic prehistory. mtDNA's high copy number, absence of recombination (Elson et al. 2001), and rapid substitution rate (Howell et al. 1996), mean that its genealogical tree is well resolved, thus making mtDNA very well suited for coalescent inference of population size.

1.2 Project Aim

The fundamental aim of the project is to draw inferences about the historic population sizes from a sample of DNA sequence data. The inferential analysis involves understanding the sample's evolutionary past through analysis of a present day sample of mtDNA sequence data. Coalescent methods are useful for exploring the demographic history of the sampled dataset (Kingman 1982a). The reconstruction of the genealogical tree for a sampled dataset is used for estimating the population sizes by tracing back to the time to the Most Recent Common Ancestor (Pybus et al. 1999). Conclusions about the historic population sizes require analysing reliable research documents to extract positive evidences for the research.

The dataset chosen for the analysis is the mtDNA sequence dataset for 53 individuals reported in the literature (Ingman et al.2000). It constitutes complete mtDNA sequences for 53 individuals from different geographical locations around the world. This dataset include individuals from African and the Non-African sub continents. The sequence data is collected from 53 individuals living across diverse locations of the world. The diverse nature of the sampled dataset allows studying and making inferences for a large set of population across the world.

Coalescent events (or divergence times) in a genealogical tree represent the common ancestor for a pair of sampled individuals. The past population sizes are estimated at these coalescent events using coalescent based demographic methods. Mutations are changes in the nucleotide sequences that can be responsible for divergences in the genealogical tree. There are different forces of genetic variation that account for differences between the phenotypical traits between individuals (Jobling et al. 2014). However, the thesis will account for genetic variations using nucleotide substitution models. The four nucleotide substitution models namely JC69, HKY85, TN93 and GTR will be used for the analysis. The models differ in substitution rates across different set of nucleotide substitutions. For choosing the best nucleotide substitution model for the dataset, the maximum likelihood function for the data, given the parameters of the demographic model needs to be evaluated. The calculation of the log-likelihood value for the maximum likelihood function will be carried out using the PAML (Phylogenetic Analysis by Maximum Likelihood Version 4.9j by Ziheng Yang) software. *Akaike information criterion (AIC)* is a goodness of fit estimator for evaluating the relative measure for goodness of fit among different statistical models. AIC values will be evaluated using the log-likelihood values from the PAML output to choose the best mutation model.

For further analysis, a demographic model is required for reconstructing the genealogical tree and estimating the population size through time for each coalescent interval of the tree. In this thesis, I will use the Bayesian inference for estimating the population size through time by sampling the population sizes from the posterior distribution along with the demographic parameters, given the gene sequence data. The Bayesian Skyline Plot (BSP) (Drummond et al. 2005) model is a method for estimating the past population dynamics through time from a sample of sequences. It uses a standard Markov Chain Monte Carlo (MCMC) (Metropolis et al. 1953) sampling algorithm to simultaneously estimate a posterior probability distribution for the ancestral genealogy, branch lengths, substitution model parameters, and population parameters through time, given a set of gene sequences. The posterior means for population sizes and tree height will be useful for making inferences about the past population dynamics of the sampled dataset. The resulting BSP represents a credibility interval for inferred population size that incorporates uncertainty in parameters such as the underlying ancestral gene tree, branch lengths, and rate parameters. The different sites of an mtDNA sequence can evolve at different substitution rates. There are various site heterogeneity models accounting for variable substitution rates. These models include Constant Rates Model, Invariant Sites Model (I), Gamma Model (Γ) and Gamma + Invariant Sites model (Γ + I). A sensitivity analysis will be carried out using BEAST models in order to choose an efficient site heterogeneity model along with the other parameters of the model. The chain convergence diagnostics such as Trace Plots, Autocorrelation Plots, Effective Sample Size and Gelman Rubin Statistic (Gelman and Rubin 1992) will be used for estimating the goodness of fit of the coalescent model.

2 Methodology

2.1 Coalescent Theory

Coalescent theory is a model of how gene variants sampled from a population may have originated from a common ancestor. Individuals in a population genetic sample are not independent. They are related due to co-ancestry. The coalescent is the model that describes the relationships within a sample from the present individuals back to the Most Recent Common Ancestor (MRCA). The coalescent process can relate a population's demographic history and genealogy of small segments of the ancestral tree which will be discussed in this section. The variations in the DNA sequences allows us to reconstruct the ancestral tree and infer the population size going back in time along the small segments of the ancestral tree (Kingman 1982a, 1982b). Tracing back through the tree to the MRCA, there are coalescent events occurring along the lineages in the genealogical tree. Figure 2.1 shows an example of a genealogy of sample size n = 5. The times (t_2, t_3, t_4, t_5) are defined as the times at which four coalescent events occurred. These five individuals are connected to their respective ancestor by branches in the tree. The time between two coalescent events is defined as the waiting time, $w_k = t_k - t_{k+1}$(where k = 2,…, n and $t_{n+1} = 0$ is the present time) . The length of the tree branches is determined using waiting times.

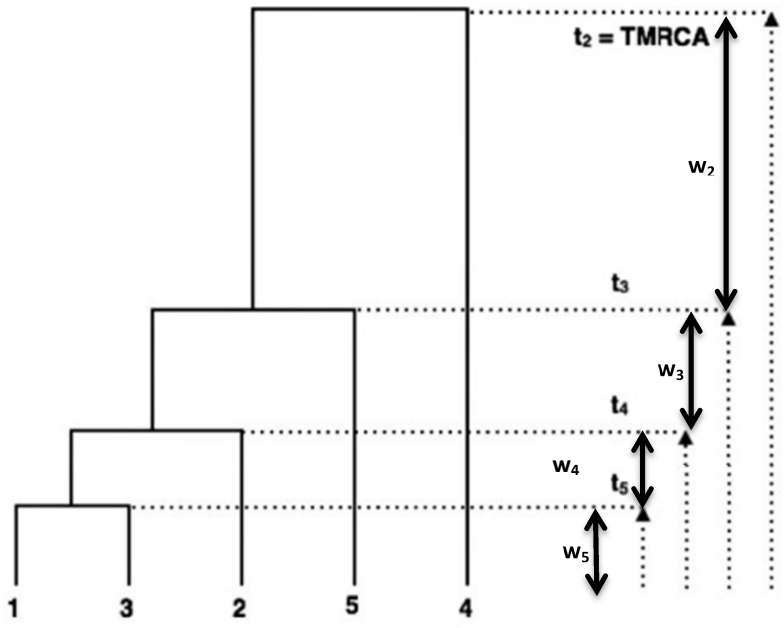


Figure 2.1 The genealogy for a sample of five individuals sampled contemporaneously. Coalescent events occurring in the tree are tracing back to the time of the Most Recent Common Ancestor for five individuals.

Strimmer and Pybus (2001) illustrated that the coalescent events among lineages occur according to a nonhomogeneous Poisson process, and the rate of coalescence (λ) is given by:-

$$\lambda_k = \binom{k}{2}$$

where, k is the number of lineages before the coalescence and ranges from sample size *n* down to 2.

When t is measured in units of generations,

$$\lambda_k = \frac{\binom{k}{2}}{N_0 g}$$

where N_0 is the initial population size and g is the number of years per generation.

The waiting times w_k, between coalescent events are independent and exponentially distributed,

$$w_k \sim \text{Exp}(\lambda_k)$$

The relationship describing the rate of coalescence between the coalescent intervals will be exploited in this thesis, by estimating waiting times in a genealogical tree from a sample of DNA sequences. The population size will be estimated at the coalescent intervals and the waiting times will be used for analysing the past population dynamics of the sample.

2.2 Nucleotide Substitution Models

DNA sequences help to understand the evolutionary past and explore demographic history of the present day sample. There are different forces of genetic variations responsible for the variations observed in the sampled sequences (Jobling et al. 2014) but this thesis will focus on the nucleotide substitutions in the sequence data. Continuous time Markov Chain methods are used to estimate the number of nucleotide substitutions. The nucleotide sites in the sequence are assumed to be evolving independently of each other. The four nucleotides (T, C, A, G) are the states of the Markov Chain. There are different models for nucleotide substitutions according to the substitution rates. They are used to estimate the evolutionary distance between sequences from the observed differences between the sequences. The expected number of nucleotide substitutions per site gives the estimated distance between the sequences.

2.2.1 The Jukes – Cantor Model

The Jukes and Cantor (1969) model is the simplest substitution model. It assumes constant mutation rate λ for every nucleotide changing to any other nucleotide.
The Markov Chain can be characterised using a transition probability over any time t > 0. The probability that a given nucleotide i will become j time t later is called the transition probability $p_{ij}(t)$ with i, j = T, C, A, or G. Thus the transition-rate matrix is:-

$$\mathbf{P} = \begin{array}{c} \\ \text{T} \\ \text{C} \\ \text{A} \\ \text{G} \end{array} \begin{array}{c} \begin{array}{cccc} \text{T} & \text{C} & \text{A} & \text{G} \end{array} \\ \begin{bmatrix} 1-3\lambda & \lambda & \lambda & \lambda \\ \lambda & 1-3\lambda & \lambda & \lambda \\ \lambda & \lambda & 1-3\lambda & \lambda \\ \lambda & \lambda & \lambda & 1-3\lambda \end{bmatrix} \end{array}$$

Each row of the matrix sums to 1. The probability that the nucleotide remains unchanged after the evolution is 1 - 3λ. When there are a large number of substitutions occurred at every site, the model attains an equilibrium position where the probability for each nucleotide in the chain is ¼ i.e. when $t \to \infty$, $p_{ij}(t) = 0.25$, for all i and j. The model assumes equal base frequencies for the four nucleotides. It is denoted by $\underline{\boldsymbol{\pi}}$ = (¼, ¼, ¼, ¼)

2.2.2 The Hasegawa – Kishino – Yano Model

The JC69 model assumes constant mutation rates for substitutions across all the four nucleotides. This is not a realistic assumption as the mutation rates differ for different substitutions across the four nucleotides. 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. Transitions occur at higher rates than transversions (Ebersberg I et al. 2002).

The Hasegawa - Kishino - Yano Model (1984, 1985) model distinguishes between the transition and transversion rates.
The transition matrix in the HKY85 model is:-

$$\mathbf{P} = \begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{bmatrix} 1-(\alpha\pi_C+\beta\pi_R) & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & 1-(\alpha\pi_T+\beta\pi_R) & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & 1-(\alpha\pi_G+\beta\pi_Y) & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & 1-(\alpha\pi_A+\beta\pi_Y) \end{bmatrix} \end{array}$$

where π_T, π_C, π_A, π_G are respective nucleotide frequencies in the sequence and $\pi_Y = \pi_T + \pi_C$ and $\pi_R = \pi_A + \pi_G$ are the frequencies of pyrimidines and purines, respectively
The base frequencies for JC69 model are assumed to be equal ($\pi_i = 0.25$, where i = T, C, A and G). This is an unrealistic assumption in real dataset. The HKY85 model allows for unequal base compositions. The transition/transversion rate ratio is given by $\kappa = \alpha/2\beta$.

2.2.3 The Tamura - Nei Model

Unlike the HKY85 model, the Tamura and Nei (1993) model distinguishes between the two different types of transitions; i.e. (A ↔ G) has a different rate to (T ↔ C). Transversions are all assumed to occur at the same rate. The TN93 model also allows for unequal base compositions. The transition-rate matrix under the TN93 model is:-

$$\mathbf{P} = \begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{bmatrix} 1-(\alpha_1\pi_C+\beta\pi_R) & \alpha_1\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha_1\pi_T & 1-(\alpha_1\pi_T+\beta\pi_R) & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & 1-(\alpha_2\pi_G+\beta\pi_Y) & \alpha_2\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha_2\pi_A & 1-(\alpha_2\pi_A+\beta\pi_Y) \end{bmatrix} \end{array}$$

where π_T, π_C, π_A, π_G are respective nucleotide frequencies in the sequence and $\pi_Y = \pi_T + \pi_C$ and $\pi_R = \pi_A + \pi_G$ are the frequencies of pyrimidines and purines, respectively. The transition-transversion ratio 1 (κ_1) and transition-transversion ratio 2 (κ_2) are given by $\kappa_1 = \alpha_1/\beta$ and $\kappa_2 = \alpha_2/\beta$ respectively.

2.2.4 GTR model (Tavaré 1986)

GTR, the Generalised time-reversible model of Tavaré 1986, is the most general, neutral, independent, finite-sites, time-reversible model possible. In addition to the unequal base frequencies from TN93 model, GTR model also allows for different substitution rates for different type of transitions as well as transversions. The GTR model is the most complex model of all the models shown. It has the maximum number of parameters compared to the other models.

The transition-rate matrix under the GTR model is:-

$$\mathbf{P} = \begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{bmatrix} 1-(a\pi_C + \mathrm{b}\pi_A + \mathrm{c}\pi_G) & a\pi_C & \mathrm{b}\pi_A & \mathrm{c}\pi_G \\ a\pi_T & 1-(a\pi_T + \mathrm{d}\pi_A + e\pi_G) & \mathrm{d}\pi_A & \mathrm{e}\pi_G \\ b\pi_T & \mathrm{d}\pi_C & 1-(b\pi_T + \mathrm{d}\pi_C + \mathrm{f}\pi_G) & \mathrm{f}\pi_G \\ c\pi_T & \mathrm{e}\pi_C & f\pi_A & 1-(c\pi_T + \mathrm{e}\pi_C + \mathrm{f}\pi_A) \end{bmatrix} \end{array}$$

where a, b, c, d, e and f represent the GTR rate parameters for the six different pairs of substitutions across the four nucleotides and π_T, π_C, π_A, π_G are respective nucleotide frequencies in the sequence.

2.3 Site Heterogeneity Models

The assumption that different sites in the sequence data evolve at the same rate is not realistic for real population data. Every nucleotide site in the DNA sequence does not evolve at the same rate. The substitution rates are variable for different sites in the sequence. Constant rates lead to underestimation of the sequence distance and impact the phylogenetic analysis (Tateno *et al.* 1994; Huelsenbeck 1995a; Yang 1996c; Sullivan and Swofford 2001). To account for different rates for different sites, using a rate parameter for every site may not be feasible as there will be too many parameters to be estimated. This will lead to a complex model which will not be computationally efficient. There are four different methods available for exploring the site heterogeneity of the mtDNA sequences. The methods are as follows:-

1. Constant Rates Model

A constant rate model is a basic model with no variation in rates among different sites in the sequence. The substitution rate for the coding region is estimated to be 1.26×10^{-8} substitutions per site per year (Mishmar et. al 2003). This rate will be used for analysing the past population dynamics of the sampled dataset. As the constant rate model is clearly an underestimation of the true mutation among the individuals (Tateno *et al.* 1994; Huelsenbeck 1995a; Yang 1996c; Sullivan and Swofford 2001), some other statistically proven better models are explored.

2. Gamma Distributed Rates Model (Γ)

A second approach is to use a Gamma distribution model (denoted by Γ) to account for variable rates in the sequence (Yang 1994a). There are few sites in the sequence which evolve at a very high substitution rate while other sites are conserved and remain unchanged. The rate parameter r for any site is a random variable drawn from a gamma distribution with shape and scale parameters as α and β. When $\beta = \alpha$, so that the mean $\alpha/\beta = 1$, the rates depend on values of α.For lower values ($\alpha \leq 1$), there are very high rates for some sites in the sequence and invariable rates for most of the other sites (Fig. 2.2). The continuous gamma distribution is complex involving large number of terms and is computationally difficult. Discrete gamma model uses several categories to approximate the continuous gamma distribution. Rates over sites are taken as random variables drawn from a discrete gamma distribution. This is an efficient method as it involves only one single parameter α (shape). The sites are divided into K (the analysis present in the thesis used K = 10) equal probability rate classes to approximate continuous distribution (Yang 1994a).

3. Invariable Sites Model (I)

The Invariable sites model (denoted by I) is a simple model where proportion of all the sites in sequence is invariable (p_{inv}) while others are changing at a same rate. It takes into account that there are few sites that are varying at a same rate and most of the other sites remain unchanged. The invariant sites model is useful for datasets with invariant sites and the evolutionary process is not stationary, reversible or homogeneous. It finds its applications in various biological researches such as studies on viruses, bacteria etc. (Vivek Jayaswal et al. 2007)

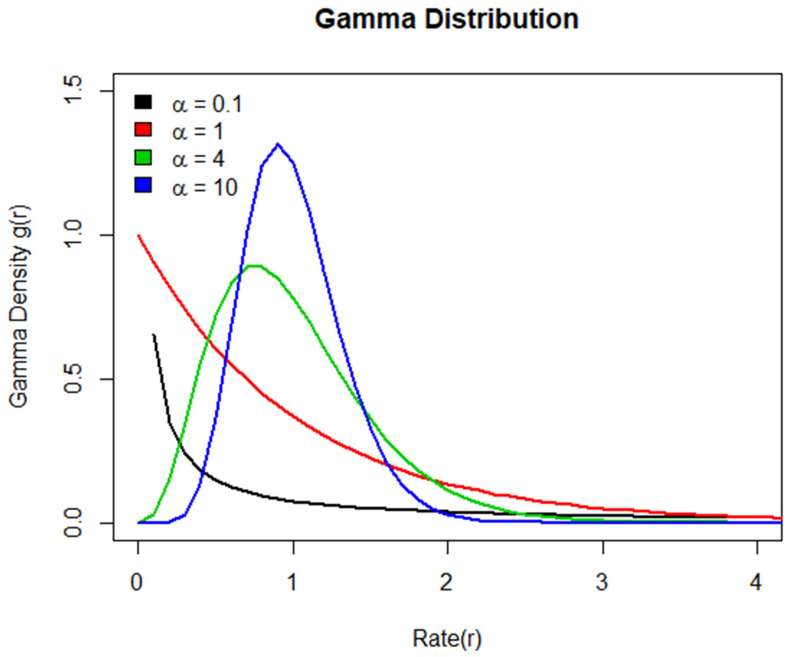


Figure 2.2 Probability density function of the gamma distribution for variable rates among sites. The rate parameter follows a Gaussian distribution as the value for the shape parameter increases keeping the scale parameter fixed and mean of distribution to be 1.

4. Gamma + Invariable Sites Model (Γ+I)

The Gamma + Invariant Sites mixture model (denoted by Γ+I) is an extension to the Gamma distribution site heterogeneity model (Γ). It has an additional proportion of invariant sites parameter (p_{inv}) added to the Γ model (Gu *et al.* 1995). The Γ+I model consists of a proportion of sites (p_{inv}) that are invariant while the other sites (with proportion $p_1 = 1 - p_{inv}$) can have constant rates or rates drawn from Gamma distribution.

2.4 Bayesian Inference

Bayes' theorem (Thomas Bayes 1761) is a mathematical formula for determining the conditional probability of an outcome occurring based on a previous outcome occurring. Mathematically, Bayes' theorem is given as: -

$$\boldsymbol{P(A|B)} = \frac{\boldsymbol{P(B|A)} \times \boldsymbol{P(A)}}{\boldsymbol{P(B)}}$$

where:
P(A|B) is the conditional probability that event A occurs given that event B has already occurred
P(A) and P(B) are the marginal probabilities of event A and event B occurring respectively.

Bayesian statistics uses an interpretation of Bayes' Theorem. The demographic model parameters required in the phylogenetic analysis for the sampled dataset can be probabilistically modelled using the Bayesian statistics. The parameters and data are denoted by **θ** and **D** respectively.

The Bayes' theorem is now given by: -

$$\boldsymbol{p(\theta|D) = \frac{p(D|\theta) \times p(D)}{p(D)}}$$

Similar to the expression for the Bayes' theorem above, the elements of this expression are considered to be a probability distribution rather than the probability of an event.
The probability distribution for the data i.e. p(D) is difficult to calculate, so the theorem can be simplified to a proportionality expression given by: -

$$\boldsymbol{p(\theta|D) \propto p(D|\theta) \times p(\theta)}$$

where:
$p(\theta|D)$ is the probability distribution of the parameters given the data and is known as posterior distribution
$p(D|\theta)$ is the probability distribution of the data given the parameters and is known as the likelihood
$p(\theta)$ is the probability distribution of the parameters θ and is known as the prior distribution
The prior distribution is the marginal distribution of θ and captures existing knowledge about the θ before any data is observed.

2.4.1 The Bayesian Skyline Plot Model

Monte Carlo Markov Chain (MCMC) algorithm is a computer-driven sampling method (Gamerman and Lopes 2006; Gilks et al. 1996) and allows to characterize a distribution without knowing all of the properties of the probability distribution. The properties of the probability distribution are estimated using random samples drawn from the distribution itself, as it allows examining a large number of samples, which is much easier than calculating the properties directly from the distribution's equations. MCMC is based on the Markov property such that the random samples are drawn sequentially. The distribution of the current draw is dependent only on the previous value drawn.

MCMC is a very efficient and useful algorithm for drawing samples from the posterior distribution, based on the values of the parameters from approximate prior distributions. The samples are drawn in such a way that the stationary distribution of the Markov chain is similar to that of the posterior distribution. The MCMC is run for a large number of iterations to obtain independent samples from the posterior distribution. The phylogenetic analysis requires drawing samples of the demographic parameters from the posterior distribution, in order to reconstruct the genealogy and explore the demographic history of the sample mtDNA sequences.

The demographic analysis requires a mathematical model to describe the evolutionary events that occurred in the past and the change in population size through time. The reconstruction of the genealogical tree and past population sizes associated with each coalescent event in the past are determined by estimating the demographic parameters, typically by maximum likelihood or Bayesian methods (Kuhner, Yamato, and Felsenstein 1998; Pybus, Rambaut, and Harvey 2000; Drummond et al. 2002). These coalescent based model parameters are estimated directly from the gene sequence data under a variety of scenarios including recombination (Griffiths and Marjoram 1996; Kuhner, Yamato, and Felsenstein 2000; Fearnhead and Donnelly 2001), population subdivision(Bahlo and Griffiths 2000; Beerli and Felsenstein 2001), and variable population size (Kuhner, Yamato, and Felsenstein 1998; Beaumont 1999; Drummond et al. 2002).

The Bayesian Skyline Plot (BSP) (Drummond and Rambaut 2003) model is used for the demographic analysis in this thesis. It infers past population dynamics using the sampled gene sequences rather than from an estimated genealogy as in the case of Classic Skyline Plot (Pybus et al. 2000) or The Generalised Skyline Plot (Strimmer and Pybus 2001). The BSP model is useful as it takes into account the errors associated with the

genealogical reconstruction which other models were unable to do. The model uses standard MCMC sampling procedure (Metropolis et al. 1953; Hastings 1970) to estimate a posterior distribution of effective population size through time, given the demographic parameters associated with the nucleotide substitution models.

The coalescent time (or divergence time) is the time elapsed since a common ancestor separated into two descending individuals. The lineages of the input genealogy go back through time to the most recent common ancestor (MRCA). There are n - 1 times at which coalescent events occur, denoted by $\mathbf{t} = \{t_1, t_2, \ldots, t_{n-1}\}$, where n is the sample size (Fig. 2.3). The times **t** are measured from the tips, such that $t_0 = 0$ at tips (or the present time when the sequences are sampled).

The waiting time between coalescent events (coalescent intervals) is defined as the time elapsed between two coalescent events of a genealogical tree (Fig 2.3). It is given by $w_i = t_i - t_{i-1}$.
The number of lineages present between each coalescent interval corresponds to a series of decreasing values in a piecewise constant model. The number of lineages is denoted by $\mathbf{k} = \{k_1, k_2, \ldots, k_{n-1}\}$ (Fig. 2.3).

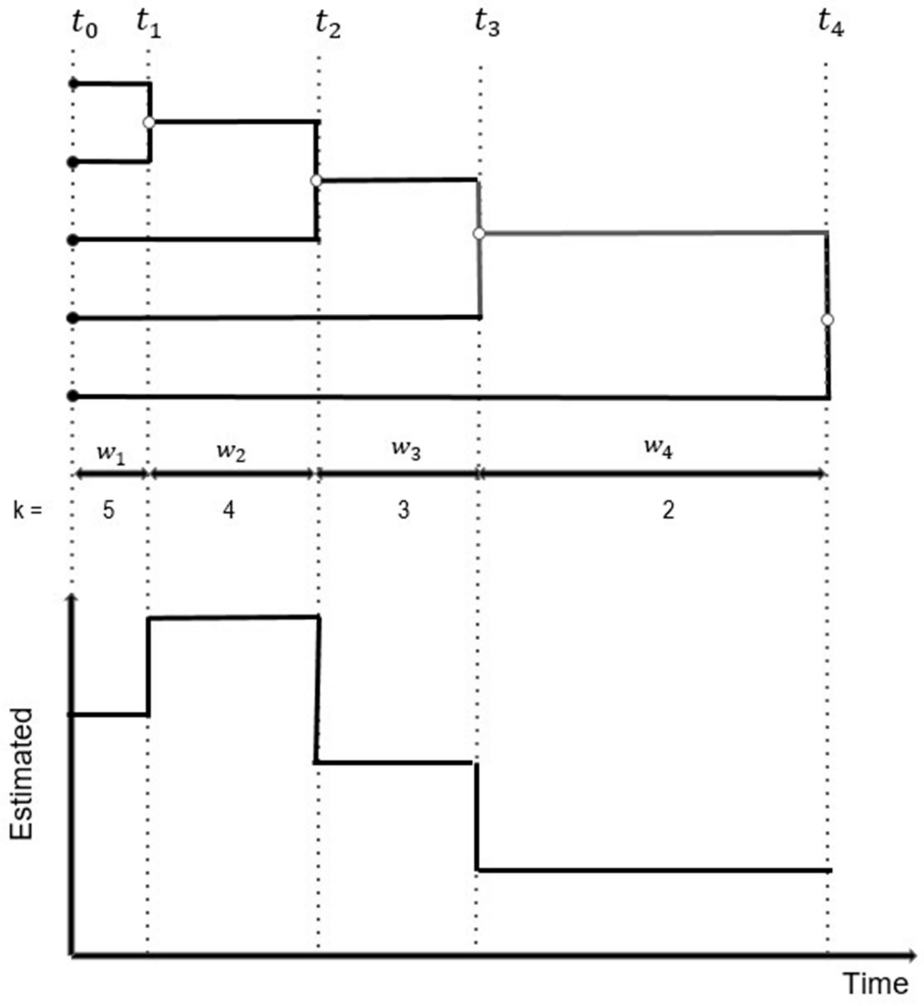


Figure 2.3 Top - A genealogical tree illustrating coalescence going back through time to the MRCA for five individuals sampled contemporaneously. There are n - 1 coalescent events, where n is the number of individuals. **Bottom -** The change in group population size at coalescent events.

The Generalised Skyline Plot (GSP) model was introduced by Strimmer and Pybus (2001), where they demonstrated that a piecewise model is used for estimating the effective population size through time (fig 3.1). The piecewise constant model assumes that the population size is constant between coalescent events, but may change at coalescent event times, **t**. The GSP allows grouping of adjacent coalescent intervals which have the same effective population size and this idea is continued in the BSP model. The number of coalescent events in each grouped interval is represented using an ordered subset of group sizes, denoted by **S** = {s_1, s_2, …, s_m} where $s_i > 0$ (Fig 3.1) and m is the number of grouped intervals$(1 \leq m \leq n-1)$.The population sizes estimated for each group interval is denoted by vector **Φ** = {ϕ_1, ϕ_2, …,ϕ_m}. The data used for the demographic analysis is an estimated genealogy with specified branch lengths, and is denoted by g. The reconstruction of genealogical tree was illustrated by Drummond et al. 2000.

The expression for the log-likelihood of the piecewise constant model was given by: -

$$\log f(g|\Phi, S) = \sum_{i=1}^{n-1} \left(\log \frac{k_i(k_i-1)}{2\varphi_{h(i)}} \right) - \frac{k_i(k_i-1)w_i}{2\varphi_{h(i)}}$$

The function h() provides a mapping from the indices of **t** to the indices of **v** and is defined:-

$$h(i) = \begin{cases} 1, & if\ i \leq s_1 \\ j, & if \sum_{k=1}^{j-1} s_k < i < \sum_{k=1}^{j} s_k \end{cases}$$

The posterior distribution for n contemporaneous sequences is given by: -

$$f(\mathbf{\Phi}, \mathrm{S}, \mathbf{\Psi}, \mathrm{g} \mid \mathrm{D}, \mu) = \frac{1}{Z} \Pr\{D|\mu, g, \Psi\} f_G(g \mid \Phi, \mathrm{S})\, f_\Phi(\Phi) f_S(\mathrm{S}) f_\Psi(\Psi)$$

The parameters from the posterior distribution of the Bayesian Skyline Plot (BSP) model are sampled by using the previously published MCMC algorithm (Drummond et al. 2002). The samples are drawn for demographic parameters, population sizes and group sizes. The hyperparameter m is not sampled as it gives consistent posterior distribution function for a range of priori values. To reduce the complexity of the model, a fixed value for number of groups (m = 10) is assigned for all the MCMC runs. μ is the substitution rate and its estimated value is 1.26×10^{-8} substitutions per site per year (Mishmar D et al. 2003). The vector **Ψ** contains the parameters of the substitution model such as transition/transversion ratio (κ), shape parameter (α) for gamma-distributed rates across sites, proportion of invariant sites (p_{inv}), base frequencies (π_i, where i = A, T, C and G) for the four nucleotides, etc. We can now find the parameter posterior distribution for parameters like population size, branch length etc. using reasonable prior distributions for the parameters.

3 Implementing Models of Population Size

3.1 Maximum Likelihood Estimation

3.1.1 Phylogenetic Analysis by Maximum Likelihood (PAML)

The log-likelihood value and the maximum likelihood estimates for the demographic model parameters such as branch lengths, transition-transversion ratio (κ), and shape parameter (α) for gamma distributed variable rates across sites, etc. are computed using PAML (Phylogenetic Analysis by Maximum Likelihood Version 4.9j by Ziheng Yang) software package (available from http://abacus.gene.ucl.ac.uk/software/PAML.html). PAML is a user-friendly interface application and allows setting up a model by choosing the components of the model such as nucleotide substitution method, molecular clock, and shape parameter (α) etc. in order to calculate the maximum likelihood estimates for the particular model. PAML model allows various nucleotide substitution models including the four models (JC69, HKY85, TN93 and GTR) useful for my analysis. Running the PAML program yields the output having the log-likelihood value and the maximum likelihood estimates of the mutation model parameters. For the TN93 model, parameter estimates for branch lengths $\mathbf{p} = \{p_1, p_2, \ldots, p_{n-1}\}$, transition-transversion ratio 1 (κ_1), transition-transversion ratio 2 (κ_2), base frequencies (π_i, where i = A, T, C and G) and shape parameter (α) for gamma-distributed rates across sites is given in the output. PAML output also contains standard errors (S.E.) for the estimated model parameters. The log-likelihood value and the number of parameters value provided in the output is useful for evaluating the goodness of fit for the mutation model.

3.1.2 Comparing Models: AIC

Akaike information criterion (AIC) is a goodness of fit estimator for comparing the relative quality of the statistical models for a given dataset. The basic formula is defined as:-

$$AIC = -2(\text{log-likelihood}) + 2n_p$$

where:
n_p is the number of model parameters

AIC values are computed using the log-likelihood values given from the PAML output for each nucleotide substitution model. AIC is useful for evaluating the goodness of fit of the mutation model. As it is a relative measure, lower AIC values indicate a more parsimonious model. The best model will be chosen by comparing the AIC values and the model with the lowest AIC value will be chosen as the best nucleotide substitution model for the dataset.

3.2 Bayesian Analysis Sampling Trees (BEAST)

The MCMC sampling from the posterior distribution is carried out using the BEAST (Bayesian Analysis Sampling Trees) software (Drummond AJ et al. 2012). Beast allows reconstructing phylogenies of the sample mtDNA sequences. It is also useful for Bayesian analysis of the molecular sequences as it

draws large number of samples from the posterior distribution. In this thesis, we will use BEAST to analyze the mtDNA sequence dataset and the prior distributions for the demographic parameters. The model used in BEAST for the sampling procedure is constructed using BEAUti (Bayesian Evolutionary Analysis Utility). BEAUti is a graphical user interface (GUI) application for generating the model files which are XML (Extensible Markup Language) files that are compatible with the BEAST software. Firstly, the dataset is imported to the BEAUti software. The dataset comprises of mtDNA sequences for 53 genes (individuals) and each gene with the length of 15446 (the coding region of the mtDNA sequence data) nucleotide sites.The next step for constructing a model is to choose the nucleotide substitution model. BEAUti gives four options for choosing the substitution model namely JC69, HKY85, TN93 and GTR. The site heterogeneity or the variation in rate for different nucleotide sites in the sequence can be modelled using Gamma (Γ), Gamma + Invariant (Γ+I), Invariant (I) and Constant rate sites model.

For the site heterogeneity models Γ and Γ+I, the number of discrete Gamma categories (K) are chosen to approximate the continuous Gamma Distributed site variable rates. The value for K depends on the size of the datasets. For small datasets, it is suggested to use as few as K = 4 categories to provide good approximation to the continuous model (Yang, 1994). In large datasets, higher number of categories would be more beneficial (Mayrose et al. 2005). BEAUti allows selecting K, ranging from 4 to 10. Since, the size of the dataset is large, comprising of 53 individuals; the maximum number of Gamma categories i.e. K = 10 is selected.

The base frequencies for the four nucleotides is equal for JC69 model but models like HKY85, TN93 and GTR allow for unequal base frequencies (π_i, where i = A,T, C and G) for the four nucleotides. The base frequencies are sampled for these models by assigning a prior distribution to the base frequencies. The coalescent events in the phylogenetic analysis are due to the mutations that occurred in the past. These events occurred at times that are dated using an estimated molecular clock. The molecular clock assumption states that the evolutionary rate is constant over time among lineages of the genealogical tree (Zuckerlandl & Pauling 1965) . A strict clock model assumes that every branch in the genealogical tree evolves with the same evolutionary rate. There are a variety of relaxed molecular clocks such as Uncorrelated Relaxed Clock, Random local clock, and Fixed local clock that can be used for relaxing the strict clock constant rate assumption. These models allow for variation in evolutionary rates among branches of the genealogical tree.

The number of demographic parameters for a nucleotide substitution model increases as we move from simple models like JC69 to the more complex ones such as the GTR model. Prior distributions for these parameters are chosen for sampling from the posterior distribution of these parameters given the observed dataset. The goal is to provide proper or reasonable priors to the parameters in order to draw maximum independent samples from the true posterior distribution. The parameters involved in the four substitution models include the substitution rate (clock rate), base frequencies (π_A, π_T, π_C, and π_G), Bayesian skyline population sizes, κ (transition-transversion ratio) etc. The site heterogeneity models such as Γ and Γ+I involve the shape parameter (α) for the Gamma distributed site variation rates. The variation in rates for different sites is dependent on the values of α; choosing a prior distribution for α is important in order to capture the variation in rates effectively. BEAST offers a wide range of prior distributions that can be used for these parameters and hyper parameters used in these hierarchical models. For instance, the TN93 + Γ model involves parameters such as κ_1 (transition-transversion ratio 1), κ_2 (transition-transversion ratio 2), base frequencies (π_i, where i = A, T, C and G), Bayesian skyline population sizes and shape parameter (α) for gamma-distributed rates across sites.

Prior Distributions

The transition-transversion ratios (κ) must be positive real number. It follows lognormal prior distribution.

$$\kappa \sim logNormal(\mu = 1, \sigma = 1.25)$$

The initial value for κ is chosen as 2.

The shape parameter (α) for Gamma distributed rates across various sites of the sequence follows Exponential prior distribution

$$\alpha \sim Exp(\lambda = 2)$$

A large proportion of sites in the sequence are invariable results in low values of α. Exponential distribution accounts for low values of α. The initial value for α is taken as 0.5.

The base frequencies (π_A, π_T, π_C, π_G) must lie between 0 and 1, and sum to 1. Hence, Dirichlet distribution is used as prior distribution for the base frequencies (McGuire et al. 2001).

$$\boldsymbol{\pi} \sim Dirichlet(1,1,1,1)$$

The base frequencies are equal for JC69 model. It is useful for the other models which allow for unequal base compositions.

The prior distribution for skyline population size is the Uniform distribution.

$$\boldsymbol{\Phi} \sim Uniform(0,1)$$

The analysis will be carried out by taking a fixed nucleotide substitution rate (μ) of 1.26×10^{-8} substitutions per site per year (Mishmar D et al. 2003).

The final step is to define the MCMC chain where the number of iterations and the sampling frequencies are provided. The MCMC samples generated from the posterior distribution using BEAST can be visualised using Tracer (v1.7.1) (Rambaut A et al. 2018). Tracer is a software package for analysing the MCMC output files generated through Bayesian demographic inference. The Bayesian Skyline Plot helps in investigating the demographic history of the gene sequences visually.

4 Analysis & Results

The dataset used for my analysis is the mtDNA sequence dataset for 53 individuals reported in the literature (Ingman et al.2000). The mtDNA sequences are re-aligned using *Bio Edit* (Hall 1999; Alzohairy 2011) before running the model for calculating the maximum likelihood function or estimating the parameters using Bayesian statistics. There are mutations in the mtDNA sequence involving insertions and deletions of the nucleotides to/from the sequence data. Insertions are mutations in which extra base pairs are inserted into a new place in the DNA. Deletions are mutations in which a section of DNA is lost, or deleted. For these reasons, Insertions/Deletions change the length of the mtDNA sequence. In order to make the length of all the sequences equal, Insertions are introduced at some sites of the sequence data using the Bio-Edit. The coding region of the mtDNA sequence comprises of sites from 577 to 16023 (Andrew et al 1999). The coding region was extracted for all the individuals from the complete mtDNA sequence data using the Bio-Edit before running the data.

4.1 Optimization Model

The log-likelihood value and the maximum likelihood estimates of the demographic model parameters are evaluated using PAML. The baseml PAML program is used for the maximum likelihood calculation. The analysis is carried out using the four nucleotide substitution models namely JC69, HKY85, TN93 and GTR. For each of the models, the strict molecular clock is chosen which assumes that every branch in the genealogical tree evolves with the same evolutionary rate. The discrete Gamma Site heterogeneity model with number of Gamma categories K =10, is used for the calculation of the Maximum likelihood estimates. The shape parameter (α) for the model is estimated with initial value = 0.5 assigned in the PAML program. Similarly, for models like HKY85 and TN93, the transition-transversion ratio (κ) is also estimated in the analysis with initial value for $\kappa = 2$ assigned in the model. The PAML output consists of branch lengths, shape parameter (α), transition-transversion ratio (κ, for HKY85 and TN93), number of parameters and the log-likelihood value. AIC (Akaike Information Criterion) values are calculated using the log-likelihood value (lnL) and the number of parameters (n_p) for each model. The AIC values for the four substitution models useful my analysis is calculated and given by:-

Substitution Model	Log-Likelihood (lnL)	No of Parameters (n_p)	AIC = $-2\times lnL + 2\times n_p$
JC69	-26939.20	53	53984.39
HKY85	-25687.45	54	51482.91
TN93	-25656.14	55	51422.28
GTR	-25652.93	58	51421.86

Table 4.1 Log-Likelihood and AIC values for the four mutation models

The AIC value is a useful test statistic to evaluate the relative goodness of fit among the different substitution models. The model with the lowest AIC value is chosen as the best substitution model. The TN93 and GTR models have approximately same AIC values. However, the GTR model is the most complex model among all the four models having the maximum number of parameters to be estimated. The GTR model is a computationally inefficient model. A complex model will always lead to a lower AIC value but it will be penalized for the complexity. Therefore, the TN93 model is chosen for further demographic analysis of the dataset.

4.2 Site Heterogeneity Models

The analysis of exploring the population dynamics for the sequence dataset requires drawing samples from the posterior distribution using the BEAST software (Drummond AJ et al. 2012). The best substitution model is chosen as TN93 for the dataset (Ingman et al. 2000) using the AIC test statistic. The TN93 model is used as the substitution model for analysing the demographic history of the sampled dataset. As the constant rate assumption for all the different sites of the sequence is an underestimation of the true distance between the individuals (Tateno *et al.* 1994; Huelsenbeck 1995a; Yang 1996c; Sullivan and Swofford 2001), the constant rates model is not used for the analysis. The Invariable rates model (I) takes into account the majority of invariable sites but restricts other sites of the sequence with same rate and there is a very less information about the mutation rates. The Γ and Γ + I models are used for the phylogenetic analysis. The number of Gamma categories K = 10 and the strict clock assumption is used for setting up the BEAST model. The Bayesian Skyline Plot (Drummond et al. 2005) with number of population groups m = 10 is used for reconstructing the genealogy of the sampled individuals. A piecewise-constant skyline model is assumed for estimating the population size and the divergence times for the dataset. The analysis is carried out by taking a fixed nucleotide substitution rate (μ) of 1.26×10^{-8} substitutions per site per year (Mishmar D et al. 2003). The MCMC algorithm for both the models TN93 + Γ and TN93 + Γ + I is run for 10Mn iterations, taking samples every 1000 iterations with a burn-in period of 10% (i.e. 1Mn) iterations.

The Γ + I model is an extension to the Γ model having an additional proportion of invariant sites parameter (p_{inv}) added to the Γ model. The Γ + I model encounters issue of non-identifiability due to presence of two strongly correlated parameters p_{inv} and α (Yang 1993; Sullivan *et al.* 1999). The gamma distribution with $\alpha \leq 1$ already accounts for different rates among sites of the sequence. It also takes into account that most of the sites are invariable since it peaks at zero, but has a long tail so can account for few high mutation rates. Adding a proportion of invariable sites (p_{inv}) parameter creates a strong correlation between p_{inv} and α and the two parameters cannot be optimised independently of each other (Mayrose et al. 2005).

Trace Plots are useful for accessing the mixing of a chain. Fig. 4.2 shows the trace plots of α and p_{inv} for the TN93 + Γ + I. The MCMC chain iterates through a long range of sample space for α and p_{inv}. There is not great randomness throughout the chain which shows that there is a serial correlation between the draws of the chain and a small number of mutually independent samples are generated from the model.

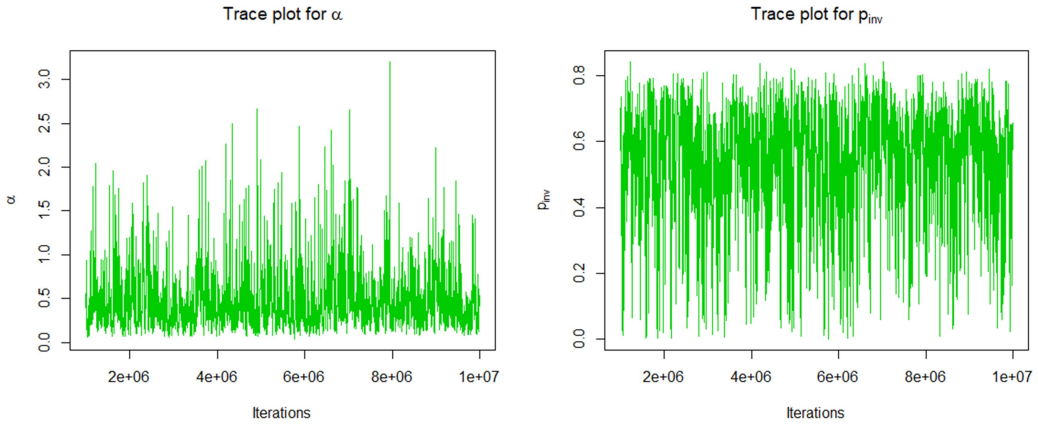


Figure 4.2 Trace plots of α and p_{inv} for the Γ + I model illustrating serial correlation between the terms of the chain.

Joint Distribution

The joint distribution for α and p_{inv} is illustrated with the help of a scatterplot (Fig. 4.3). For most of the samples for α and p_{inv} drawn from the MCMC algorithm, the relationship between α and p_{inv} suggests that for large values of p_{inv}, there are smaller values of α. High values of p_{inv} indicate most of the sites in the sequence are invariable to any mutational changes. Corresponding to high values of p_{inv}, there are low values for α which suggests that α itself is sufficient enough to account for high number of invariable sites of the sequence. Apart from the invariable sites, there are few sites in the sequence which evolve at very high or low mutation rates and α takes into account the rate variation for these sites as well (Yang 1994a). The application of the proportion of invariant sites (p_{inv}) is unidentifiable as the rate variation among the sites of the sequence could be already accounted for by the scale parameter α of the Gamma distribution. The additional parameter (p_{inv}) only increases the complexity of the MCMC algorithm and the chain requires longer time to run, making it computationally inefficient.

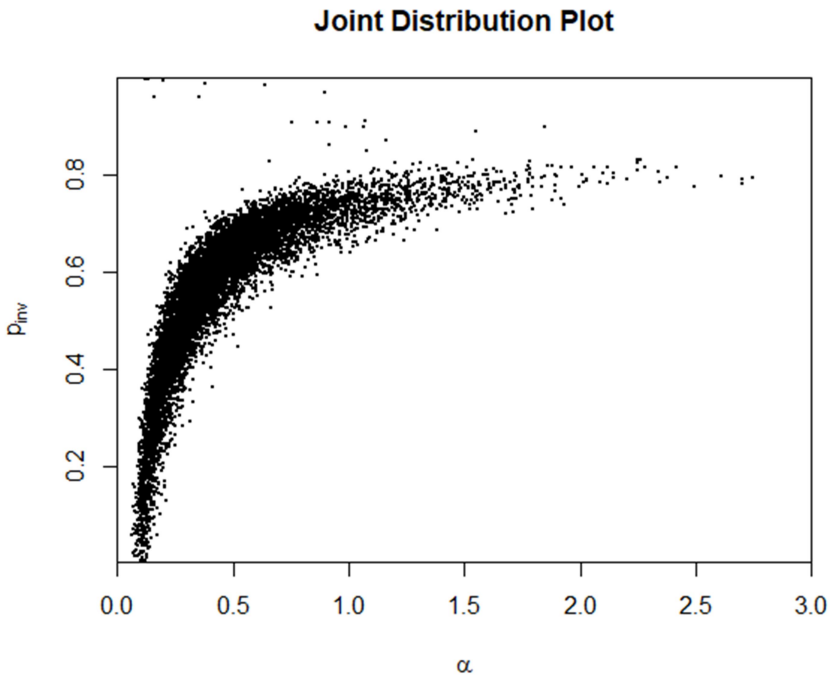


Figure 4.3 Plot illustrating strong correlation between shape parameter(α) and proportion of invariant sites(p_{inv}).

Summary Statistics

Table 4.2 gives the different summary statistics for α and p_{inv} for the Γ + I and Γ models. The shape parameter (α) for Γ + I model has a very high standard deviation compared to the Γ model. The MCMC algorithm for both the models Γ + I and Γ is run for 10Mn iterations, taking samples every 1000 iterations with a burn-in period of 10% (i.e. 1Mn) iterations.

	Γ + I		**Γ**
Statistic	**α**	**p_{inv}**	**A**
Mean	0.472	0.554	0.0948
Std. Dev.	0.3156	0.1562	0.0203
ESS	643	549	5983

Table 4.2 Summary statistics for parameters α and p_{inv} showing small ESS for Γ + I model compared to Γ model

The Effective sample size (ESS) of a parameter sampled from an MCMC is the number of effectively independent draws from the target posterior distribution. The ESS for α in Γ model is 5983, whereas the ESS for α in Γ + I model is 643 which is very low compared to the Γ model. It implies that the samples drawn from the MCMC chain are not independent draws from the target posterior for Γ+I model and the chain takes longer time for convergence (Geyer 2011). The ESS can be improved by running the chain for longer duration with increased number of iterations, but this would not be computationally efficient and will exploit more resources for the purpose of chain convergence. The Γ+I model is criticised due to the involvement of two highly correlated parameters α and p_{inv} (Mayrose et al. 2005). The Γ + I model is an inefficient method compared to Γ model for the dataset. The further analysis for exploring the demographic history for the sampled dataset is carried out using the Γ model.

4.3 Chain Mixing and Convergence

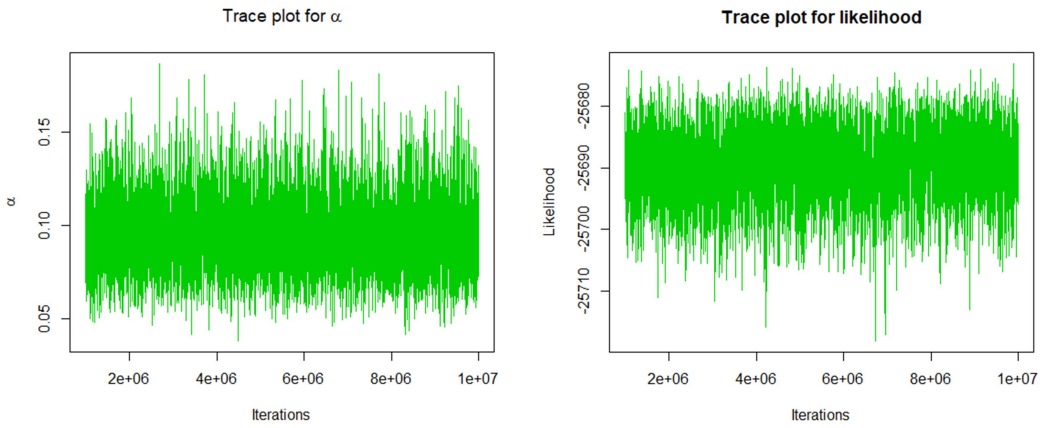
The basic idea of an MCMC algorithm is to create samples that have a stationary distribution similar to the target posterior distribution. The MCMC chain is assumed to be converged when the samples generated from the algorithm are similar to the target distributions.
MCMC diagnostics are tools that can used to check the quality of the samples generated from the MCMC algorithm. There are two major problems that could be observed in MCMC output:-

1. a large proportion of the samples are drawn from distributions that are significantly different from the target distributions.
2. the independent sample size is too small.

Convergence diagnostics such as Trace Plots, Autocorrelation Plots, Effective Sample Size and Gelman Rubin Statistic are used for evaluating the Chain Convergence for the MCMC model. The TN93 + Γ model is used for the demographic analysis for this thesis. MCMC samples are drawn from the posterior distribution using the BEAST software.

4.3.1 Trace Plots

A Trace plot shows the sampled values of a parameter over the length of the MCMC chain. It is an important tool used for accessing the chain convergence. The MCMC algorithm for TN93 + Γ is run for 10Mn iterations, taking samples for every 1000 iterations with a burn-in period of 10% (i.e. 1Mn) iterations. The trace plots for α, κ_1 and κ_2 (Fig 4.4) suggests that the MCMC chain is converged and there is a low serial correlation between the parameter samples i.e. a substantial amount of samples are mutually independent.



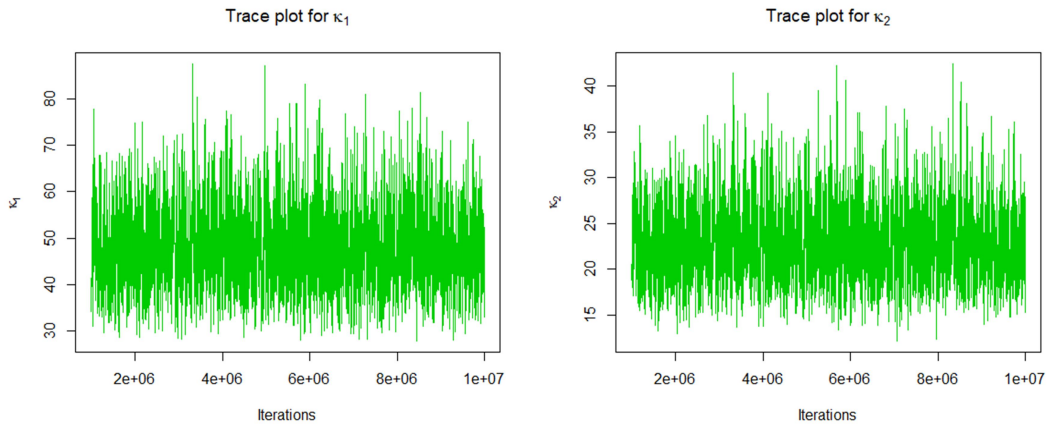


Figure 4.4 Trace plots for the model parameters showing the chain is exploring the sample space very fast.

4.3.2 Auto Correlation Plots:-

Autocorrelation is the coefficient of linear correlation between two terms of a sequence of random variables. Autocorrelation plots are used to measure the correlation between the draws of the chain precisely. The Autocorrelation plots of α, κ_1 and κ_2 (Fig. 4.5) for the TN93 + Γ model shows that the sample autocorrelation between the terms of the chain decreases as the lag increases. The autocorrelation is large for short lags but then goes to zero for larger values of lags.

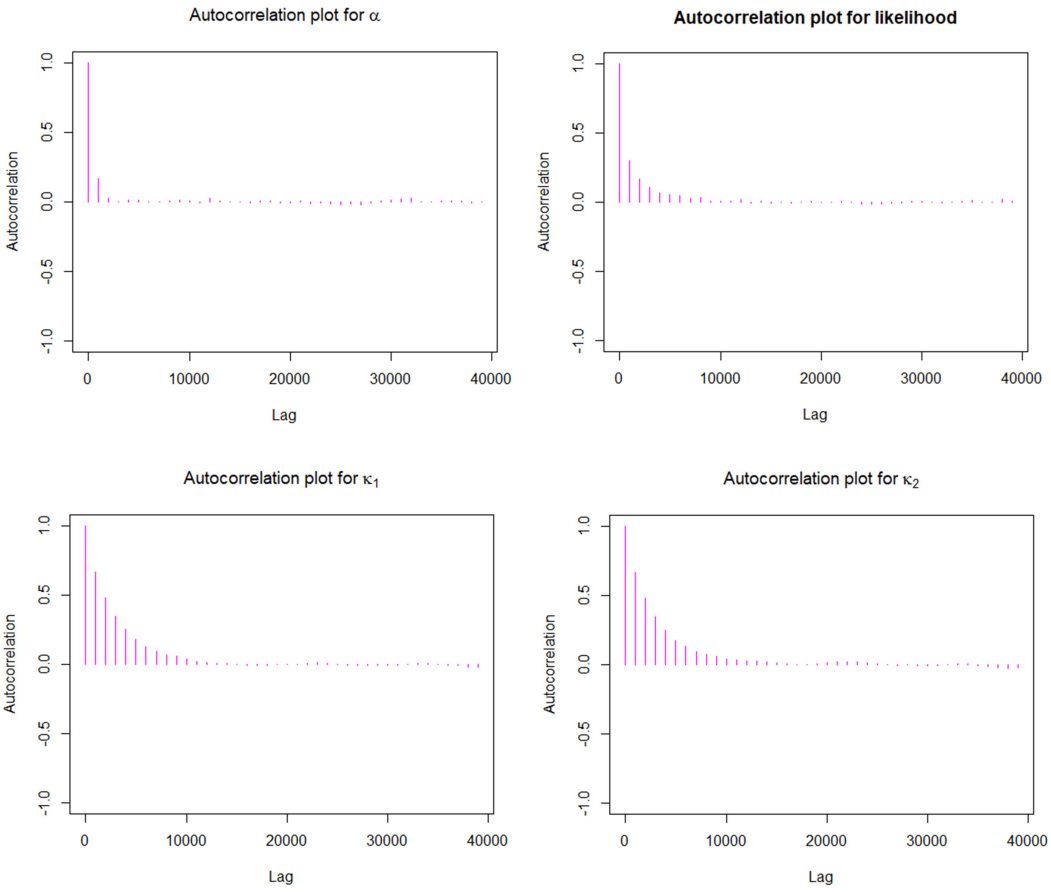


Figure 4.5 Autocorrelation plots illustrating chain convergence for the parameters of the TN93 + Γ model

4.3.3 Effective Sample Size

The Effective sample size (ESS) of a parameter sampled from an MCMC is the number of effectively independent draws from the target posterior distribution. The ESS values in Table 4.3 shows that for the chain with 10Mn iterations, taking samples every 1000 iterations, with a burn in period of 10% (i.e. 1Mn) iterations, the ESS is acceptable for the estimated TN93 + Γ model parameters. The samples are independent and the chain is assumed to be converged.

TN93 + Γ	ESS
α	5982.9
κ_1	1573.4
κ_2	1527.7
π_A	3314.5
π_T	3213.6
π_C	3474.5
π_G	3365.9

Table 4.3 Effective Sample Size for the TN93 + Γ model parameters

4.3.4 Gelman Rubin Statistic

The Gelman-Rubin (GR) Statistic or potential scale reduction factor diagnostic (Gelman & Rubin, 1992) is calculated for evaluating the MCMC chain convergence. The convergence is assessed by comparing the estimated between-chains and within-chain variances for each model parameter. Large differences between these variances indicate non-convergence. The GR test statistic or potential scale reduction factor (denoted by R) is calculated for each variable. Approximate convergence is diagnosed when the potential scale reduction factor (or shrink factor, R) is close to 1. Brooks and Gelman (1997) have suggested that, if R < 1.1 for the model parameters, it can be inferred that the convergence has reached.

Gelman-Rubin Diagnostic

TN93 + Γ	Point Estimate	Upper C.I.
α	1.000656	1.002486
κ_1	1.000391	1.001189
κ_2	1.000590	1.001492
π_A	1.000343	1.001067
π_T	1.000277	1.001133
π_C	1.000028	1.000166
π_G	1.000420	1.001691
Tree height	1.000109	1.000119

Table 4.5 The point estimates and upper limits for the GR statistics of the TN93 + Γ model parameters

The point estimate and the upper limit for the Gelman-Rubin (GR) diagnostic for the model parameters are shown in Table 4.4. The parameters include shape parameter (α) for gamma distributed rates across sites, transition-transversion ratio (κ), base frequencies (π_i, where i = A, T, C and G) for the four nucleotides and the Tree Height.

The MCMC algorithm for TN93 + Γ model is run for three chains with over dispersed initial values for each parameter. Each chain is run for 10Mn iterations taking samples for every 1000 iterations with a burn-in period of 10% (i.e. 1Mn iterations). The shrink factors for α, κ_1 and κ_2 are very close to 1. The upper limits for shrink factors are also close to 1.

4.4 Reconstructed Population Size

The chain convergence diagnostics indicate that the chain is converged and independent samples are generated from the target posterior distribution. The Effective Sample Size for the parameters of the distribution is acceptable in respect of the MCMC chain of length 10Mn iterations with the first 10% (i.e. 1Mn) iterations discarded as burn-in and model parameters were sampled every 1000 iterations thereafter. The analysis is carried out using the TN93 + Γ nucleotide substitution model where number of population groups (m) was set to 10 and number of Gamma categories (K) was set to 10 for running the BEAST model.

The reconstruction of the population sizes through time under the TN93 + Γ model is illustrated using the Bayesian Skyline Plot (Fig 4.6), constructed using the posterior samples drawn from the BEAST model. The posterior means, medians and the 95% highest posterior density (HPD) intervals of population sizes and branch lengths given by the BEAST output are used to find a relationship between a population's demographic history and its genealogy. The shape parameter (α) for gamma distributed rates across various sites, the transition-transversion ratios (κ_1 and κ_2) and the base frequencies (π_A, π_T, π_C, π_G) are co-estimated along the parameters of the Bayesian Skyline plot model and the ancestral genealogy.

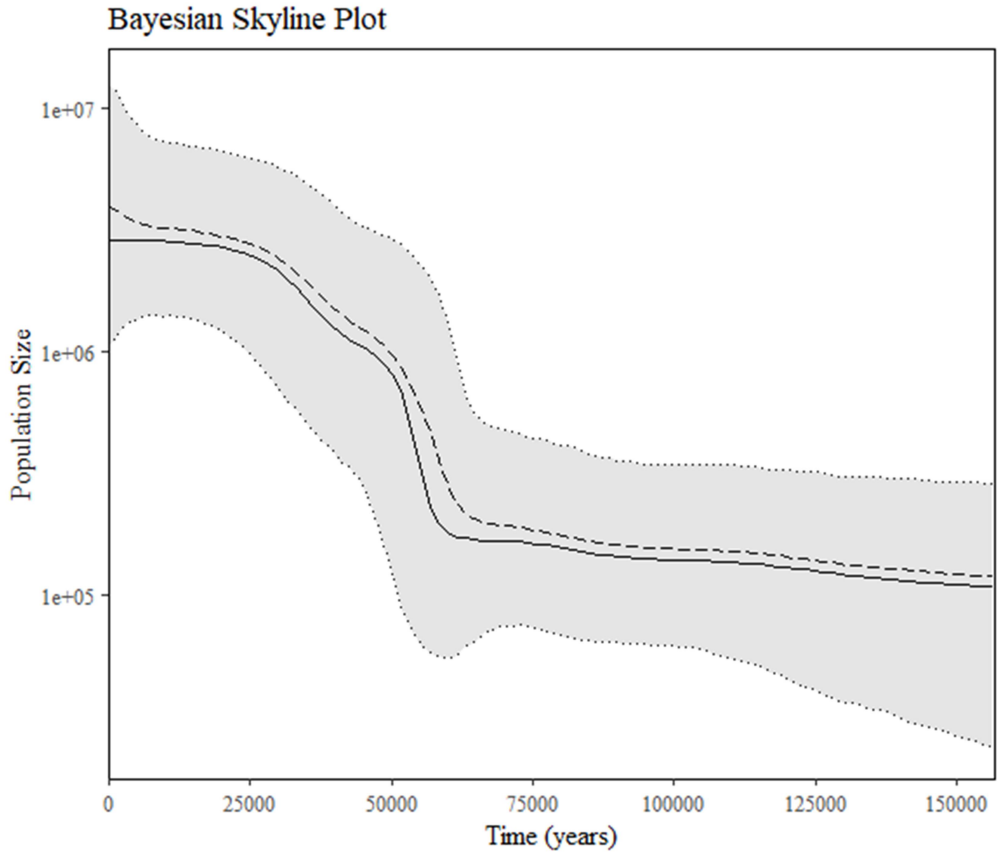


Figure 4.6 A Bayesian Skyline Plot (m=10) derived from a sample of mtDNA coding region sequences for 53 individuals of diverse geographical locations. The x axis is in units of years, and the y axis is the estimated population size through time in base – 10 log scale. The thick solid line is the median estimate and the dotted lines show the 95% HPD limits. The thick dashed line shows the mean estimate for the population size. The plot shows a rapid increase in the population size between 60000-50000 years ago, probably due to the migration of modern humans from Africa to different parts of the world.

The Bayesian Skyline Plot shows the summarized results for the estimated mean and median population sizes through each time with an associated measure of uncertainty bounded by the 95% HPD intervals (Fig 4.6). The plot describes a slow constant increase in the population size from the period between 160,000-60,000 years ago. Previously published work on early migrations from Africa to other geographical locations of the world suggests that the Homo sapiens were most likely developed in the Horn of Africa between 300,000-200,000 years ago (Mellars P et al. 2019). This marks as evidence that the slow increase in population size during the period of 160,000-60,000 years ago is due to presence of human population in a small confined region of the world.

The plot describes an explosive growth in the estimated population size ~ 60,000 years ago. The population size estimates increased by 5 times during the period 60000-50000 years ago. The rapid increase in the population size could be due to the migrations out of Africa to different parts of the world which caused a major outbreak of population growth in the world. It is consistent with the early southern Asian growth phase with high mtDNA diversity in populations from the Indian Ocean Coast (Kivisild et al. 2003) as well as recent archaeological evidences (Mellars 2006a), that strongly supports a rapid coastal migration along the ‘‘Southern Route’’ from Africa into Southern Asia. The early migration of Africans to Europe and Australia is also supported by Macaulay et al. (2005) and Rasmussen et al. (2011). The estimated population size is increasing after the early migration movements from Africa to major parts of the world. This could be due to the medical advancements which resulted in improved mortality and fertility rates.

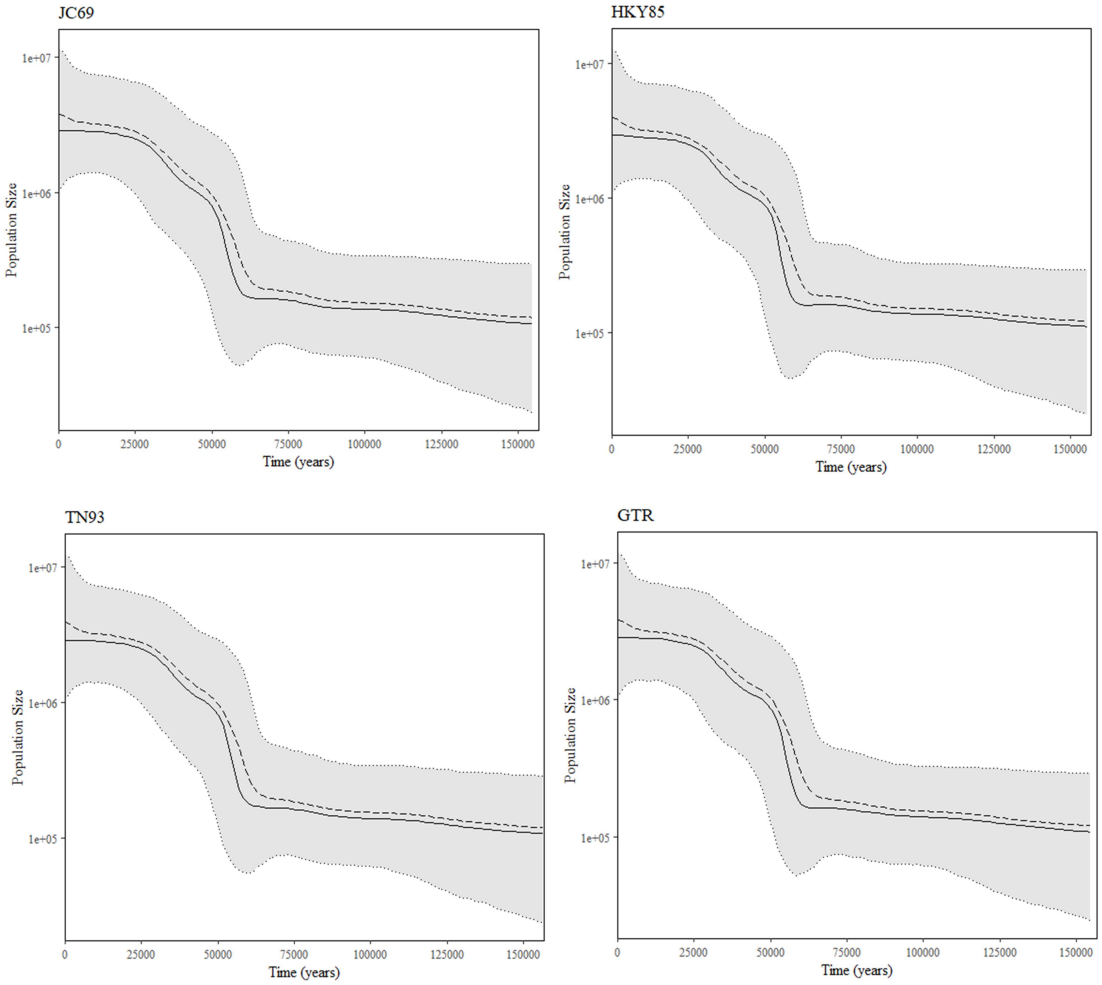


Figure 4.7 Bayesian Skyline plots for different models giving similar estimated population sizes through time.

The demographic history reconstructed using the Bayesian Skyline plot for the four substitution models is illustrated in Fig 4.7. The sensitivity analysis using the maximum likelihood methods chooses the TN93 model as the best nucleotide substitution model for the dataset. However, the other three models namely JC69, HKY85 and GTR give similar population dynamics for the dataset.

5 Further Analysis

The Bayesian Skyline Plot is used to describe the demographic history of 53 sampled sequences using the summarized results of estimated posterior means, medians and the 95% HPD limits for the population size through time. The population size for each coalescent interval is estimated using a reconstructed genealogical tree. The reconstruction of the genealogical tree is carried out using structure and estimated branch lengths provided by the BEAST output. The genealogical tree for the sampled sequences can be graphically analysed using the FigTree (v1.4.4) software (available from http://tree.bio.ed.ac.uk/software/figtree/). The mean Tree height for the sample is evaluated by tracing back the branch lengths along the lineages of the genealogical tree to the Most Recent Common Ancestor for the sample (Ingman et al. dataset 2000). It is also known as the Time to the Most Recent Common Ancestor (TMRCA). The TMRCA of human mtDNA for our analysis based on the mtDNA sequence dataset for 53 individuals is estimated to be 184,500 years before present (YBP). In order to date the MRCA of human mtDNA, the TN93 + Γ substitution model was used and the substitution rate (μ) was taken to be 1.26×10^{-8} substitutions per site per year (Mishmar et. al 2003). Ingman et al. (2000) estimated the TMRCA of human mtDNA to be 171,500 YBP using these 53 individuals. They carried their analysis using the substitution rate of 1.70×10^{-8} substitutions per site per year estimated from the mean genetic distance between all humans and one chimpanzee sequence (Kumar and Hedges 1998). Similar attempts for studying the evolution of human mtDNA have been implemented by other researchers like Horai et al. (1995) and Macaulay et al. (2005). The TMRCA of human mtDNA estimated using age estimates from Horai et al. (1995) is 190,000 YBP (H.-J. Bandelt et al. 2005). It is calculated based on the assumption that human-chimp mtDNA split dated back to 6.5 million years ago (Mishmar et al. 2003). For the dataset of Macaulay et al. (2005), the TMRCA of human mtDNA is estimated to be 202,000 YBP (H.-J. Bandelt et al. 2005). The structure and the estimated TMRCA of the reconstructed genealogical tree of this thesis shows favourable results in accordance with the previously published research work on evolution of human DNA sequences.

6 Conclusion

The phylogenetic analysis for inferring the past population dynamics of a sampled dataset is implemented using the Bayesian Skyline Plot model. The analysis is carried out using a diverse global mtDNA sample. The Bayesian Skyline plot model is an efficient model as it uses the MCMC sampling procedures to estimate a posterior distribution of population size through directly from a sample of mtDNA sequences. Unlike the Classic Skyline Plot and Generalized Skyline Plot, the Bayesian skyline plot accounts for phylogenetic error for the sampled dataset. The phylogenetic error is pertinent for less variable dataset and the Bayesian Skyline plot accounts for these slowly evolving sequences. The MCMC method used for implementing the Bayesian Skyline plot is a very efficient algorithm as it co-estimates the ancestral genealogy and the substitution model parameters along with the demographic parameters. It gives smoother estimates for the parameters of the model. Unlike previous models, the Bayesian Skyline plot gives high posterior density (HPD) intervals for the estimated population size to represent both phylogenetic and coalescent uncertainties.

The maximum likelihood analysis reveals the best chosen mutation model as the TN93 model. The Γ + I model is criticised due to strong correlation between the two variables α and p_{inv}. The Γ model is chosen to account for variable substitution rates across different sites of the sequence. The TN93 + Γ model is chosen for exploring the demographic history of the sampled sequences. However, the Bayesian Skyline plot of the dataset (Ingman et al. 2000) for the other three models (JC69, HKY85 and GTR) show similar results about the population sizes through time.

The Bayesian skyline plot constructed shows a slow constant growth before the period of 60,000 years ago and a rapid increase in population size following that from the period between 60000-50000 years ago. Previously published work suggests that the period from 60000-50000 years ago was the period of early migrations from Africa to different parts of the world (Macaulay et al. 2005). The analysis carried out in the thesis excellently shows the past population dynamics of a sampled dataset. The population sizes estimated from the analysis accounts for the early migrations from Africa. The dataset used for the analysis was a small dataset of 53 diverse individuals from the world and still the results of the analysis were commendable. The estimated TMRCA of the sampled dataset is close to the time of the evolution of human mtDNA estimated by many researchers. The Bayesian Skyline Plot model gave appreciating results for estimating the demographic history of the sampled dataset. The analysis was carried out using the coding region of the mtDNA sequences. However, if I had another six months for carrying forward the research, I would have explored the population dynamics using the control region of the mtDNA sequence. The genealogical tree reconstructed from the analysis could be studied for different sub groups of the sampled dataset. The dataset used was a small dataset and choosing a dataset containing more number of sequences could have given even better results.

References

Max Roser, Hannah Ritchie and Esteban Ortiz-Ospina (2013) - "World Population Growth". *Published online at OurWorldInData.org.* Retrieved from: 'https://ourworldindata.org/world-population-growth' [Online Resource]

Sigurğardóttir S, Helgason A, Gulcher JR, Stefansson K, Donnelly P. The mutation rate in the human mtDNA control region. *Am J Hum Genet*. 2000;66(5):1599-1609. doi:10.1086/302902

Vivek Jayaswal, John Robinson, Lars Jermiin, Estimation of Phylogeny and Invariant Sites under the General Markov Model of Nucleotide Sequence Evolution, *Systematic Biology*, Volume 56, Issue 2, April 2007, Pages 155–162, https://doi.org/10.1080/10635150701247921

Smith, G., Vijaykrishna, D., Bahl, J. *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459,** 1122–1125 (2009). https://doi.org/10.1038/nature08182).

Henn Ojaveer, Ain Lankov, Marilyn Teder, Mart Simm, Riina Klais, Hydrobiologia, 2017. Feeding patterns of dominating small pelagic fish in the Gulf of Riga, Baltic Sea. 792; Dochtermann, N. A., & Peacock, M. M. (2013). Inter- and intra-specific patterns of density dependence and population size variability in Salmoniformes. *Oecologia*

Human Evolutionary Genetics (2nd Edition) – Mark Jobling, Edward Hollox, Matthew Hurles, Toomas Kivisild, and Chris Tyler Smith 2013

Computational Molecular Evolution – Ziheng Yang 2006

Tateno *et al.* 1994; Huelsenbeck 1995a; Yang 1996c; Sullivan and Swofford 2001)

Kuhner, Yamato, and Felsenstein 1998; Pybus, Rambaut, and Harvey 2000; Drummond et al. 2002)

Griffiths and Marjoram 1996; Kuhner, Yamato, and Felsenstein 2000; Fearnhead and Donnelly 2001)

Bahlo and Griffiths 2000; Beerli and Felsenstein 2001), and variable population size (Kuhner, Yamato, and Felsenstein 1998; Beaumont 1999; Drummond et al. 2002

Drummond, A.J., Suchard, M.A., Xie, D. and Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution*, *29*(8), pp.1969-1973.

Strimmer, K. and Pybus, O.G., 2001. Exploring the demographic history of DNA sequences using the generalized skyline plot. *Molecular Biology and Evolution*, *18*(12), pp.2298-2305.

Felsenstein, J., 2005. Theoretical evolutionary genetics joseph felsenstein. *University of Washington, Seattle*.

Ingman, M., Kaessmann, H., Pääbo, S. and Gyllensten, U., 2000. Mitochondrial genome variation and the origin of modern humans. *Nature*, *408*(6813), pp.708-713.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005 - Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol. 22:1185–1192.

Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory - Jotun Hein, Mikkel H. Schierup, and Carsten Wiuf 2004