UPPSALA
UNIVERSITET

UPTEC X 21026
Examensarbete 30 hp
Augusti 2021

Developing new methods for estimating population divergence times from sequence data

Karl Svärd

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Developing new methods for estimating population divergence times from sequence data

Karl Svärd

Methods for estimating past demographic events of populations are powerful tools in order to get insights of otherwise hidden pasts. The genetic data of people is a valuable resource for these purposes as patterns of variation can inform of the past evolutionary forces and historical events that generated them. There is, however, a lack of methods within the field that uses this information to its full extent. That is why this project has looked at developing a set of new alternatives for estimating demographic events.

The work done has been based on modifying the purely sequence based method TTo (Two-Two-outgroup) for estimating divergence times of two populations. The modifications consisted of using beta distributions to model the polymorphic diversity of the ancestral population in order to increase the max sample size possible. The finished project resulted in two implemented methods: TT-beta and a partial variant of MM. TT-beta was able to produce estimations in the same region as TTo and showed that the usage of beta distributions had real potential. For MM there only was a partial implementation able to be done, but this one also showed promise and the ability to use varying sample sizes to estimate demographic values.

Handledare: Per Sjödin
Ämnesgranskare: Matthew Webster
Examinator: Pascal Milesi
ISSN: 1401-2138, UPTEC X 21026
Tryckt av: Uppsala

Nya metoder för att få studera vår historia

Vår historia är något som intresserar folk av alla åldrar och ursprung. Var kommer våra förfäder ifrån, vilka händelser har de upplevt, hur såg det ut på deras tid? Alla frågor som många idag söker svar på. Till deras hjälp används arkeologiska fynd, gamla berättelser och bevarade skrifter, klassiska metoder som gett oss större delen av våran nutida förståelse. Vi kan följa våran utveckling som en art, från överlevnad i små grupper av jägare till medborgare i komplexa civilisationer. Men, som med det mesta, så har även dessa metoder sina nackdelar. Luckor befinner sig fortfarande i våran historia och mycket bygger på antaganden gjorda på enstaka källor.

Men med modern teknik behöver detta inte nödvändigtvis vara fallet. De senaste årtiondena har visat en explosiv ökning av potential inom studerandet av våran interna databank, den genetiska arvsmassan. Nya metoder har gjort det mycket lättare att få tillgång och extrahera den djupa mängd information som gömmer sig inombords. Detta har visat att en persons genetiska kod säger mer än enbart vilken ögonfärg man har och hur snabbt man kan springa. Mönster av variation sedda över hela dataset i större folkgrupper har visats sig kunna berätta om historiska och evolutionära händelser som skett förr i tiden.

Tecken på att detta är ett relativt nytt område märks dock tydligt. Det nuvarande utbudet av metoder som har potentialen att utvinna historiska händelser från genetisk data är för tillfället begränsat. Detta är i kombination med att många av metoderna bygger på antaganden gjorda av en populations historia innan dess att resultat ens är färdiga. Det här projektet har haft som mål att försöka uppdatera fältet med några nya metoder baserade på att man som användare slipper ge alltför begränsande antaganden, och istället ge resultat direkt från given genetisk data.

Metoderna som har utvecklats är en fortsättning på en befintligt metod inom fältet för uppskattning av hur länge sedan två populationer delade sig från varandra. Arbetet har bestått av att modifiera denna metod för att kunna arbeta med större mängder data och då teoretiskt ge uppskattningar av bättre kvalité. För att uppnå detta så är de nya metoderna baserade på användningen av numeriska metoder för att lösa en serie med komplexa och informativa ekvationer om två populationers demografiska förflutna. Detta har resulterat i två metoder som tillsammans visar stark potential för fortsatt arbete och utveckling. Varianterna visar uppskattningarna av splittider som är åtminstone lika bra som metoden den är baserad på, men är fortfarande långt från att presenteras som en helt färdig metod på grund av dess instabilitet. Idéer på framtida arbetssteg är som tur många och förhoppningsvist så kan grunden definierat i det här arbetet leda till en full lösning i framtiden.

Table of contents

1 Background **1**

1.1 Coalescent Theory . . . 2

1.2 Simulating Genetic Ancestry . . . 2

1.3 Estimating Demographic Events . . . 3

1.3.1 Model Based Methods . . . 3

1.3.2 TT Method . . . 3

1.4 Modelling allele frequencies with Beta Distributions . . . 4

1.5 Aims . . . 4

2 Material and Methods **5**

2.1 Implementations . . . 5

2.1.1 Setting up TT-beta . . . 5

2.1.2 Setting up MM . . . 7

2.2 Simulating data . . . 8

2.3 Testing on real datasets . . . 8

3 Results **9**

3.1 Implementation of TT-beta . . . 9

3.1.1 TT-beta performance . . . 10

3.2 Implementation of MM . . . 12

3.2.1 MM performance . . . 13

3.3 Tests on real data . . . 15

4 Discussion **15**

4.1 TT-beta as a new method . . . 15

4.1.1 Run time and accuracy . . . 16

4.1.2 Compared to TT and TTo . . . 16

4.1.3 Runs on real data . . . 16

4.2 MM and its potential . . . 17

4.3 Future work . . . 17

4.4 Conclusion . . . 18

5 Acknowledgements **18**

Abbreviations

ABC	Approximate Bayesian computation
DNA	Deoxyribonucleic acid
MM	Many-Many method
TT	Two-Two method
TTo	Two-Two-outgroup method

1 Background

The strive to understand our history and events which occurred long past has always been present in people throughout the years. Initially in the form of stories orally passed down through generations, then followed by written records and archaeological findings in more modern times. These studies have made it possible for us to follow the advancement of us as a species, from simple hunter-gatherers to members of complex civilizations. But not everything about our history can clearly be interpreted from these sources. Significant gaps or otherwise unrefined assumptions exist, especially from our early years as a species where mostly sparse skeletal remains are our only clues. While a lot of knowledge still can be gathered from these with the classical methods, the real potential lies within in the form of information dense genetic material.

The vast genomic data existing within each individual contain more information than solely the genetic code deciding personal looks and abilities. Patterns of genetic variation among individuals in populations can also be used for inferring the evolutionary forces and historical events that generated them. This discovery, in addition to the increasing availability of sequenced human DNA, has opened many possibilities to uncover or refine our knowledge of the demographic history of the human race. Genetic data gathered from sources such as skeletal remains can be used to estimate events such as ancient migrations and size changes of populations that otherwise would be hard to interpret (Schraiber & Akey 2015).

However, areas of improvement still exists within the field as the number of available analytical methods is limited, and in many cases, are too reliant on prerequisite assumptions of a population's past. This master's thesis project aims at improving the selection of current demographic inference methods by continuing the previous work of developing a novel alternative called 'MM' (Many-Many). It is a method for inferring demographic events of two populations, that in theory will circumvent these prior assumptions and thus be able to give more accurate estimations. The project will be based on modifying an already existing method named 'TT' (Two-Two) which can estimate population divergence times from given sample data (Sjödin *et al.* 2021). The modifications consist of using the family of probability density functions called beta distributions to model the diversity in the ancestral population. In theory, this will increase the maximum sample size from each population, which in turn should lead to an increase of statistical power and make it possible to estimate more parameters such as historic changes in population size.

1.1 Coalescent Theory

One of the most commonly used tools within population genetics is the stochastic process known as 'the coalescent', or alternatively 'Kingman's coalescent' named after its first discoverer John Kingman (Kingman 1982). While being purely based on probabilistic mathematics, it has showed real prowess of modeling the genealogical history of a set of individuals. The process is based on the approach of modeling a set of genetic samples backwards in time from a given starting state. Followed by retroactively adding any possible mutations that could have occurred on the constructed lineages. This has proven on many occasions to both be effective and give solid predictions (Wakeley 2009).

The principle of the coalescent has the added bonus that it can be applied on a wide range of different models of evolution, from neutral models where no selection or population size changes occur, to more complex ones when combining with other methods for simulation. The process consists of iteratively sampling (with replacement) the parents to all samples in the current generation. This makes it possible for two or more lineages to receive the same parent, or in other words coalesce. These steps are then continued until all lineages coalesce into one (Wakeley 2009).

1.2 Simulating Genetic Ancestry

One of the major benefits resulting from the coalescent's effectiveness is the set of computationally efficient and fast computer algorithms made possible for simulating genetic samples and their ancestry (Wakeley 2009). This is especially useful for generating in silico data to be compared with observed values gathered from real life scenarios in order to get increased insights of a population's past. However, implementations of this have faced issues when scaling the tasks to the sizes required for modern day genome wide analyses on hundreds of thousands of samples. A result of the constant improvements happening within the field of genetic sequencing. Newer alternatives has thus been developed with better scalability and better processing of extensive simulations. One such software is the population genetics simulator msprime, based on the open source tree sequence software tskit (Kelleher *et al.* 2016). A tree sequence, or succint tree sequence as they are formally called, represents the relationships between genetic sequences in a way in which informs on the full genetic ancestry and gives a lossless compression of DNA datasets (Kelleher *et al.* 2019). The use of these give msprime an efficient simulation process even for large datasets, and the ability to save finished simulations more space-efficiently in the form of TreeSequence files.

1.3 Estimating Demographic Events

In the absence of sufficient records or other types of concrete proof, estimation of past demographic events is our best bet to get insights on our past. Many different approaches for this exists, but they all rely on the gold mine of information that is genetic sequences.

1.3.1 Model Based Methods

Many already existing methods for estimating demographic events like divergence times rely on performing simulations on sets of models of evolution. The results from these tests are then compared to what can be observed empirically in order to assess the plausibility of the used model. Approximate Bayesian Computation (ABC) is one such method which uses a rejection algorithm in order to decide how well a model fits to given observed data. However, the task of choosing which models to test and what parameters to use can be quite the challenging task for methods such as these as the set of different models to choose from is very large. This also increases the risk of missing out on some of the underlying demographic information existing within the data (Beaumont *et al.* 2002).

1.3.2 TT Method

Recent developments have tried to circumvent the issues inherent to the earlier mentioned methods by instead estimating demographic parameters without specifying the exact underlying model. One example is the so-called 'Two-Two (TT)' method for estimating divergence time by relying on sequence data from two sampled haploid genomes from each of two populations. The method assumes a general split model without migration between the two sampled populations and a constant ancestral population size, but is robust to low levels of admixture and requires no assumptions about past population sizes in the daughter populations. Model parameters are then estimated by solving probabilistic equations based on the observed genetic variation and coalescent theory. The genetic variation in this case comes from the observed instances of eight possible sample configurations regarding if a position is conserved from the ancestral population or derived (0, 1 or 2 per population). The total number of sample configurations is eight in this case as the two monomorphic cases (0,0 and 2,2) is combined into one. Population time estimates are ultimately gathered from each of the two branches in the model, resulting in two final estimates (Schlebusch *et al.* 2017) (Sjödin *et al.* 2021).

A closely related method named 'Two-Two-outgroup (TTo)' is another alternative which is a continuation of the prior method. By additionally sampling from an outgroup popu-

lation, that split sometime before the main investigated divergence event, you can ascertain if observed derived variants are older than the split or not. This makes it possible to ignore the possibility of novel mutations occurring down the line and the method does not require the assumption of a constant ancestral population size. New equations can thus be derived that are more robust to past changes in the ancestral population size (Sjödin *et al.* 2021).

The MM method is the next logical step of the TT and TTo method. By modeling the polymorphic diversity with beta distributions, you free up a lot of previous restrictions and make it possible for the method to work with flexible number of samples per population. The extra information gathered from the potential increase of sample data also results in the method in theory being able to estimate historic changes in the sizes of the daughter populations.

1.4 Modelling allele frequencies with Beta Distributions

Beta distributions are a family of continuous probability distributions within statistics. They are characterized by their two positive shape parameters, a and b, that give them a wide range of possible shapes. The density of a beta distribution, defined on the interval $[0, 1]$, follows the formula:

$$\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$$

The high level of flexibility of the formula has led to uses within several fields of study where you want to model a distribution within limited ranges. One such is within population genetics, where the use of beta distributions has seen much use. Different forms of the beta distributions has been used in several cases to model the allele frequencies in populations. One such example is the Balding–Nichols model, presented in 1995, which models the allele frequencies in specifically sub-divided populations (Balding & Nichols 1995).

1.5 Aims

The goal of this project is first to develop a method called TT-beta, a modified version of TTo that has been modified to use beta distributions to model the polymorphic frequency spectra of the ancestral population. After this, focus will be shifted towards trying to implement MM based on the same principles as TT-beta. The performance of

both methods will be assessed on simulated data, with the addition of testing it on real data in the case of TT-beta.

2 Material and Methods

The work done during the project can be divided up into several steps: implementing method, testing results on simulated genetic data and, if the prior tests showed enough promise, running the method on available actual sequence data from different human populations.

2.1 Implementations

The implementations has been done primarily in Python, with some additional C scripts in the case of the MM method. Bash scripts was also written for running large tasks on the computing cluster resources provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX.

2.1.1 Setting up TT-beta

Before trying to implement the MM method, another more slimmed down alternative currently named 'TT-beta' was worked on first. This was a version of the TTo method that had been modified to rely on beta distributions in a similar manner as the theory behind MM. Following the calculations in Appendix A, the equation set

$$r_1 \equiv \frac{1}{4}\frac{2p_{2,1} + p_{1,1}}{\alpha_2}$$
$$r_2 \equiv \frac{1}{4}\frac{2p_{1,2} + p_{1,1}}{\alpha_1}$$
$$s \equiv \frac{1}{4}\frac{p_{1,1}}{\alpha_1\alpha_2}$$

can be derived from sample configuration data. These r and s values don't represent any real life values by themselves, but gives rise to further valuable conclusions (such as the estimated divergence time) when basing the ancestral population's polymorphic frequency spectra on beta distributions. $p_{i,j}$ is the probability of the sample configuration

with i derived and $2-i$ ancestral in population 1, and j derived and $2-j$ ancestral in population 2. This is in this case estimated by the counts of a sample configuration ($m_{i,j}$) divided by the total count of all configurations (m_{tot}). The α_i parameters are conditional estimates taken from the theory behind the TTo method and informs of the probability of two lineages in the population coalescing before the estimated time of the split (Sjödin *et al.* 2021). The counts used in this case is conditional on the outgroup, meaning that only sites where the derived variant has been observed in the outgroup sample considered. These counts are marked with a * to make this difference clear.

$$\alpha_1 = 2\frac{m^*_{1,0} + m^*_{1,2} + m^*_{1,1}}{2(m^*_{1,0} + 2m^*_{2,0} + m^*_{2,1}) + m^*_{1,1}}$$
$$\alpha_2 = 2\frac{m^*_{0,1} + m^*_{2,1} + m^*_{1,1}}{2(m^*_{0,1} + 2m^*_{0,2} + m^*_{1,2}) + m^*_{1,1}}$$

As further shown in Appendix A, basing the ancestral population's polymorphic frequency distribution on the beta distribution results in the third degree equations

$$r = \frac{a(a+1)b}{(a+b)(a+b+1)(a+b+2)}$$
$$s = \frac{a(a+1)b(b+1)}{(a+b)(a+b+1)(a+b+2)(a+b+3)}$$

where r denotes either $r1$ or $r2$ as they result in the same equation. Using the previously known $r1$, $r2$ and s values you then can estimate which a and b parameters gives the best fit to the given sample data. This was done with numerical optimization methods from the Python package SciPy, where the full selection of suitable algorithms were tested in order to find a solution to a and b based on the r and s equations. This process was done twice for each of the two possible variants of the r function as they result in two different equation systems to be solved, which each one giving their own divergence time estimates T per population. The 4 in total estimates were finally calculated with:

$$T_{1,1} = \frac{2m_{1,0} + m_{1,1} + 4m_{2,0} + 2m_{2,1}}{4m_{tot}} - \frac{a_1 b_1}{(a_1+b_1)(a_1+b_1+1)}$$
$$T_{1,2} = \frac{2m_{1,0} + m_{1,1} + 4m_{2,0} + 2m_{2,1}}{4m_{tot}} - \frac{a_2 b_2}{(a_2+b_2)(a_2+b_2+1)}$$
$$T_{2,1} = \frac{2m_{0,1} + m_{1,1} + 4m_{0,2} + 2m_{1,2}}{4m_{tot}} - \frac{a_1 b_1}{(a_1+b_1)(a_1+b_1+1)}$$
$$T_{2,2} = \frac{2m_{0,1} + m_{1,1} + 4m_{0,2} + 2m_{1,2}}{4m_{tot}} - \frac{a_2 b_2}{(a_2+b_2)(a_2+b_2+1)}$$

Where $T_{1,1}$ and $T_{1,2}$ are the two estimates of the divergence time of population 1 using, respectively, the $r1$ and $r2$ based estimates of a and b. $T_{2,1}$ and $T_{2,2}$ are the two estimates of the divergence time of population 2.

2.1.2 Setting up MM

The implementation of the MM method was limited to be able to fit within the range of this thesis project. Instead of aiming for a complete method for estimating divergence times, only a subset of parameters were estimated: a, b, τ_A and τ_B. Following Appendix B, the two first parameters are from the beta distribution (like TT-beta) and the two τ values are the drift parameters in the daughter populations A and B. They are defined as

$$\tau = \int_0^t \frac{dz}{2N(z)}$$

where t is the divergence time in generations away from the population and $N(z)$ is the diploid population size z generations away in the same direction.

The main part of the MM method heavily relies on the two probabilities g (Tavaré 1984) and h (Slatkin 1996). The first denoting the probability of there being k ancestors at the scaled time τ in a sample of n gene-copies:

$$g_{n,k}(\tau) = \frac{1}{\binom{k}{2}} \sum_{i=k}^{n} e^{-\binom{i}{2}\tau} \binom{i}{2} \prod_{l=k, l\neq i}^{n} \frac{\binom{l}{2}}{\binom{l}{2} - \binom{i}{2}}$$

when $n < k \leq 2$, with special cases $g_{n,n}(\tau) = e^{-\binom{n}{2}\tau}$ and $g_{n,1}(\tau) = 1 - \sum_{k=2}^{n} g_{n,k}(\tau)$. The other probability is defined by

$$h(m, l; i, j) = \frac{\binom{m-1}{i-1}\binom{l-1}{j-1}}{\binom{m+l-1}{i+j-1}}$$

and describes the probability of having m derived lineages and l ancestral lineages when there are $m + l$ lineages in total, given that there were i derived and j ancestral lineages when there were $i + j$ lineages in total.

The process of estimating demographic parameters then comes from determining which scenario results in the highest probability given the observed counts of the possible sample configurations (m_A, l_A, m_B, l_B). The limitations put in place for the MM method in this case led to that only a partial solution was looked into where you ignored the possibility that the derived mutations are younger than the ancestral population. In practise

ignoring sample configurations where the derived allele only occurs in one population, with the exception of monomorphic cases. You get

$$P\left(S_B = \{m_B, l_B\} \wedge S_A = \{m_A, l_A\}\right)$$
$$= \sum_{i_A=0}^{m_A} \sum_{j_A=0}^{l_A} \sum_{i_B=0}^{m_B} \sum_{j_B=0}^{l_B} \frac{B(a + i_A + i_B, b + j_A + j_B)}{B(a, b)} h^*(\tau_A; m_A, l_A; i_A, j_A) h^*(\tau_B; m_B, l_B; i_B, j_B)$$

with

$$h^*(\tau; m, l; i, j) = \binom{i+j}{i} h(m, l, i, j) g_{m+l, i+j}(\tau)$$

and $B(x, y)$ is the beta function. The SciPy package was also used in this case in order to find the most probable values for a, b, τ_A and τ_B. A Python function was set up that took the 4 parameters as input and returned the total probability from the formula above. The output was in the end inverted (multiplied by -1) in order to turn it into a minimization problem better fit for SciPy's wide range of minimization functions.

2.2 Simulating data

The simulated data used to test the effectiveness of the created variants of TT beta and MM was created with the software msprime and ran on the UPPMAX cluster to facilitate runs with up to 10000 replicates. Performances where calculated from a set of demographic models simulating different historical events that could have occurred around the time of a population split. These include changes to the ancestral population size: constant, bottle-neck, expansion or shrinking. A model with a bottleneck event occurring in one of the daughter populations was also tested. The simulations ran with 2 generated samples per daughter population, resulting in a total sample size of 4. Exceptions from this came when testing the MM implementation as runs with 2, 4, 6, 8 and 10 samples per population was simulated. Saved performance values plotted in the form of histograms from Python's Matplotlib package for presentation purposes.

2.3 Testing on real datasets

The TT-beta implementation was as a final test also ran on real sample data available from human populations. The dataset used was the same as in Sjödin et al, with 11 individuals from the HGDP (Human Genome Diversity Project) and the Denisovan genome

and Altai Neanderthal genome from (Meyer *et al.* 2012) and (Prüfer *et al.* 2014). The HGDP samples contained individuals of, amongst others, French, Sardinian and african populations (Sjödin *et al.* 2021). All were compared against each other using three different outgroups: baa, archaic and Mbunti populations (see Sjödin et al for details). Estimates of divergence times, the beta distribution and its shape defining parameters a and b were collected and then plotted for visualisation purposes with a range of R scripts.

3 Results

The TT-beta method, in addition to the partial MM implementation, was successfully implemented in Python with the `optimize` package from SciPy. Both alternatives were also able to be tested on large simulated datasets produced from a range of different demographic modules that were able to be executed on UPPMAX. As this process was relatively problem free, enough time was also available to perform tests on real human population data.

3.1 Implementation of TT-beta

The r and s equation system were written in Python as a function taking the vector `[a,b]` as input and returning the reformed equations: `[a*b*(a+1)/((a+b)*(a+b+1)*(a+b+2)) - r, a*b*(a+1)*(b+1)/((a+b)*(a+b+1)*(a+b+2)*(a+b+3))-s]` that approaches zero when a and b is estimated correctly. This is in accordance to the format required for the root solving algorithms of `scipy.optimize`. Mainly the root finding functions `fsolve` and `root` were tested for this purpose because of their wide range of different algorithms available, but also other alternatives like `least_squares` and `minimize` were tested candidates. In the end, `root` using its `hybr` algorithm was the best performing alternative with both fast running time and accurate results.

However, one important observation during testing was the big impact that the starting guess had on the method's final guess. Results could in some cases be extremely different just because it was handed a starting value in a hard to solve region. This was accounted for by creating a specialized root finding function, using `root`, that for a minimum of 500 iterations randomizes the starting value within a given search space of $a = [0, 1]$ and $b = [1, 10000]$, as the assumptions for a panmitic ancestral population

requires $0 < a < 1$ and $1 < b$. The best root found from this search is finally given as output.

Two additional variants of this TT-beta implementation was also experimented with when it was observed that guesses not always followed the assumption of a panmitic ancestral population with b values in some cases being outside its given range. The alternatives TT-beta 'bound' and 'converted' were set up, where the former used `least_squares` to only look for solutions within the allowed interval, and the later that simply converted the results from the normal TT-beta method by changing any values outside the allowed interval to the closest one within.

3.1.1 TT-beta performance

The methods from each of the three different versions of TT-beta were tested on the set of demographic models created with msprime. All simulations ran with 10000 replicates and were also tested on classical TT and TTo. Results for TT-beta were promising and showed the majority of values around the true divergence time of 50,000 years. However, in some cases extreme miscalculations were found. As shown in figure 1 and 2.

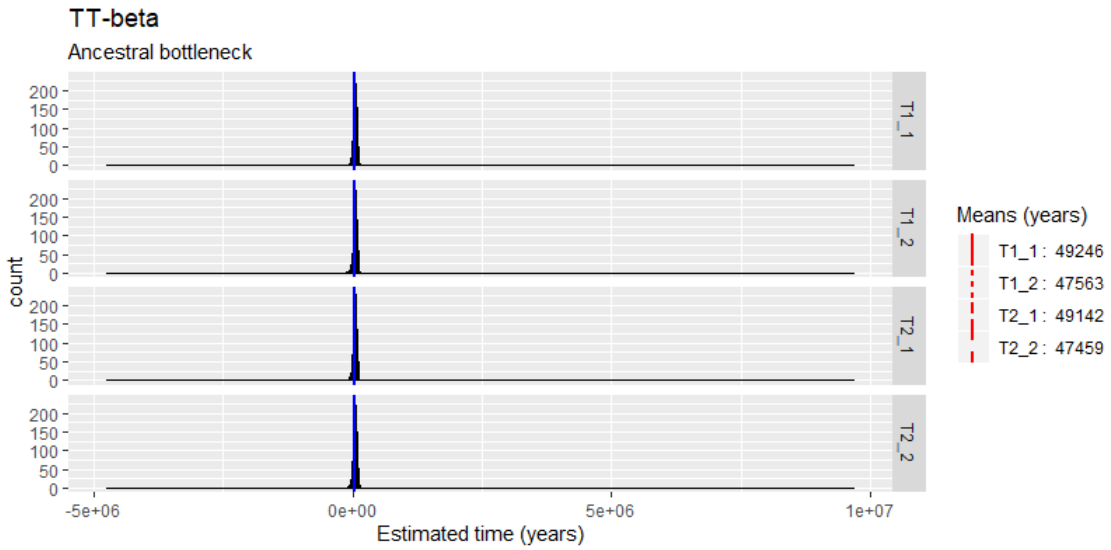


Figure 1: Histogram showing the estimates of normal TT-beta from simulated data of an ancestral bottleneck. The blue line represents the true divergence time.

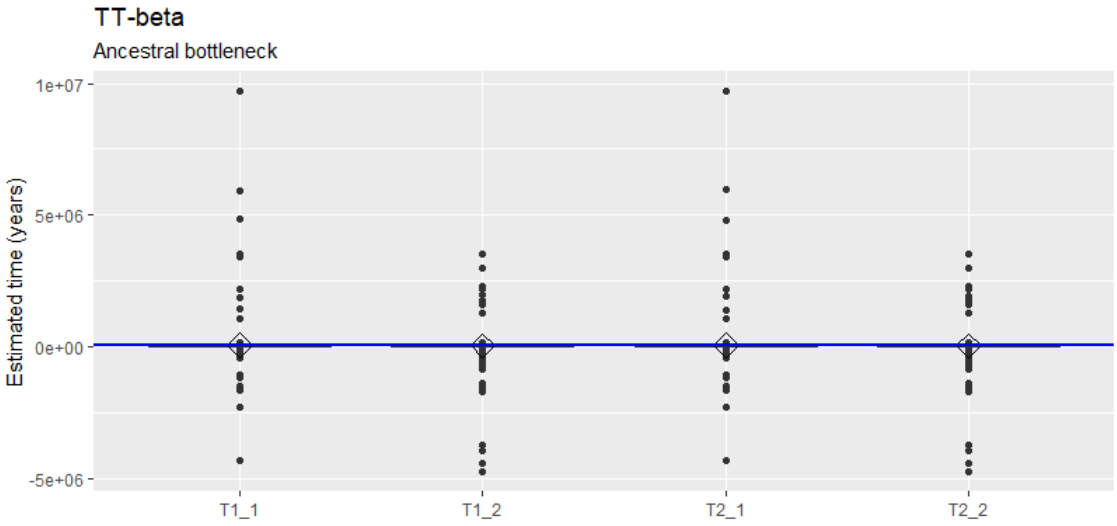


Figure 2: Boxplot showing the estimates of normal TT-beta from simulated data of an ancestral bottleneck. The blue line represents the true divergence time and black dots the instances of outliers.

Using R, outliers were extracted from the boxplots (like figure 2) and then hidden from the dataset they came from. This resulted in trimmed plots that were much easier to interpret (figure 3).

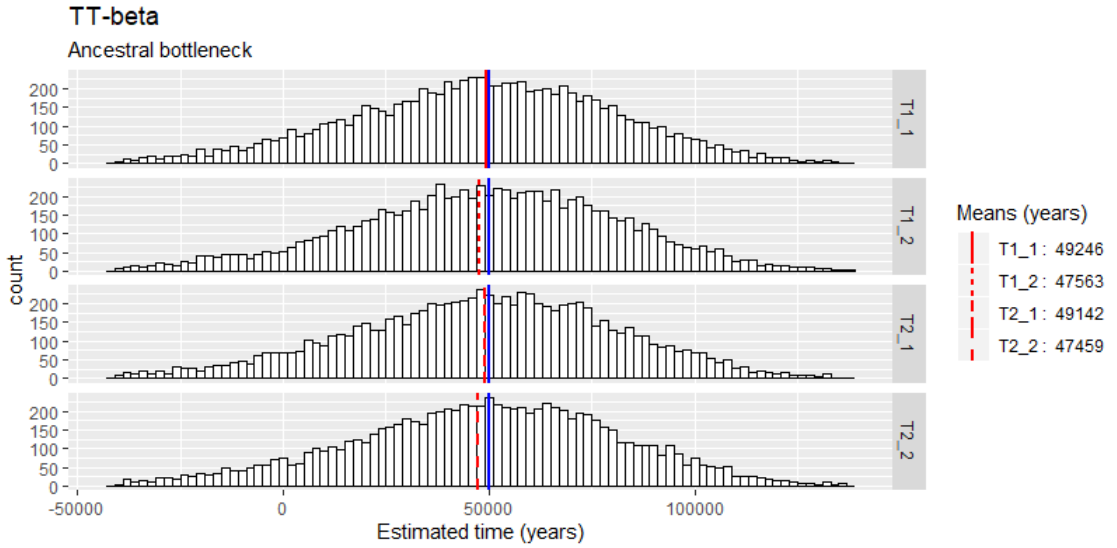


Figure 3: Trimmed histogram showing the estimates of normal TT-beta from simulated data of an ancestral bottleneck, after outliers have been removed. The blue line represents the true divergence time.

This shows quite a wide distribution of different estimates, but with a clear weight on the true divergence time. An interesting observation from this spread is the presence of also negative time estimates. The trimmed performances across all tested demographic models on the normal TT-beta is summarized in figure 4 and shows that this is a common trend for most of the observed demographic models. A wide distribution of different estimates, ranging from highs of 150,000 plus years to lows of almost negative 50,000 years, but with the calculated mean still notably close to the true divergence time value.

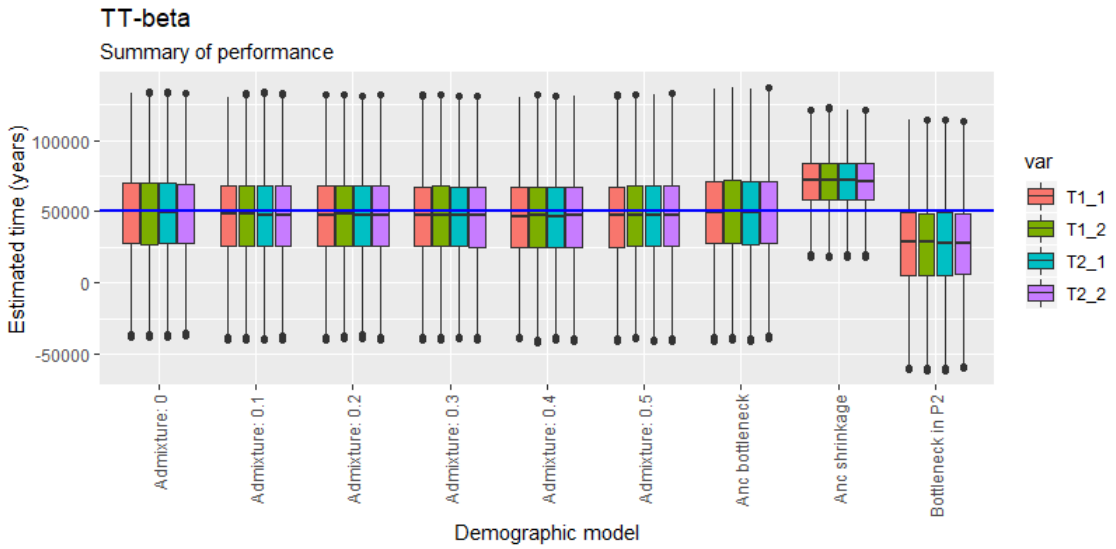


Figure 4: Summary of all trimmed boxplots from the runs of the normal TT-beta method.

Summaries of the bound and converted TT-beta alternatives showed a similar spread of guesses but with mean values farther away from the truth (see Appendix C). In the appendix you can also see how the base TT and TTo performed on the same data.

3.2 Implementation of MM

The probabilistic functions h and $h*$, central to the MM method, were implemented as a series of Python functions containing all the parts required according to the theory. The g function was written in C++ and could be connected to the otherwise fully Python based system. A function to be optimized, called f, was created from these parts in such a way that took a, b, τ_A and τ_B as input and returned the probability of them existing according to the already given information: samples sizes nA, nB, the observed set of configurations and their counts m. The output was also inverted (multiplied with -1) in

order to fit the requirements of SciPy's `minimize` function. Other alternatives were also tested but `minimize` with its default solving algorithm performed the best in accuracy and with relatively good speed compared to the other solving algorithms from function. The optimization was performed within the boundaries possible of the variables: $a > 0$, $b > 0$, $\tau_A > 0$, $\tau_B > 0$. One difference from the TT-beta implementation was that only one starting guess was used ($\tau_A = 0.2$, $\tau_B = 0.8$, $a = 0.0005$, $b = 1$), as the running time took several times as long and it would not be feasible to use a similar method of finding the optimal one for each case.

3.2.1 MM performance

Only the estimated τ parameters could be compared to their true values as a result of the partial implementation of MM performed in this report . This is because a and b for the moment don't have any formulas in this scenario which could be used to derive them (but a should probably be close to 0 and b larger than 1). It is for this reason that only the τ values will be observed from the performance tests.

Simulations were run on four different demographic modules with 4 samples per population: basic split model, ancestral bottleneck, ancestral shrinking population and 0.2 levels of admixture after split. An extra four runs were also performed that tested the effect of sample size, with: 2, 6, 8 or 10 samples per population on a basic split model. All observed models should in theory have the same τ values as the daughter populations are constant of size 10000 and the divergence event is constant at 50,000 years ago. This should result in $\tau_A = 0.1$ and $\tau_B = 0.1$.

Similarly to the results of TT-beta, instances of extreme outliers also persisted in some cases for the MM method and required trimming to be able to see the true distribution (see Appendix D). This showed an interesting trend that fewer samples seems to drastically worsen the quality of the estimations. As shown in figure 5 where results from tests where 4 and 8 samples per population is compared.

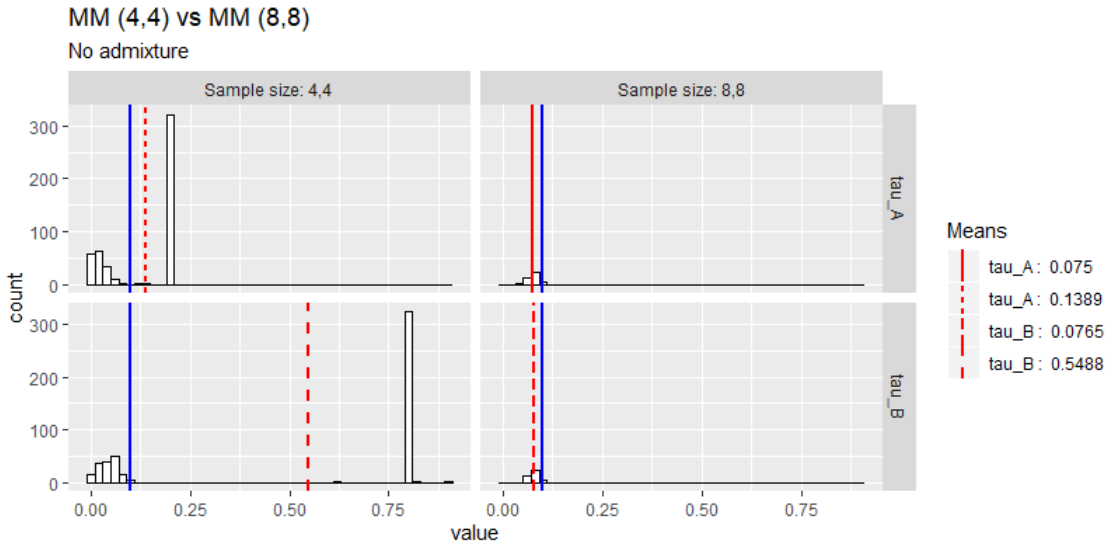


Figure 5: Comparison of τ estimations from MM when using 4 samples per population or 8. Unfortunately, drastic differences in number of replicates had to be used, with 1000 for (4,4) and 50 for (8,8).

This trend was made even more obvious when comparing to the rest of the different sample setups used (figure 6). The limited testing on different demographic models, however, didn't seem to have any notable effect on the estimations.

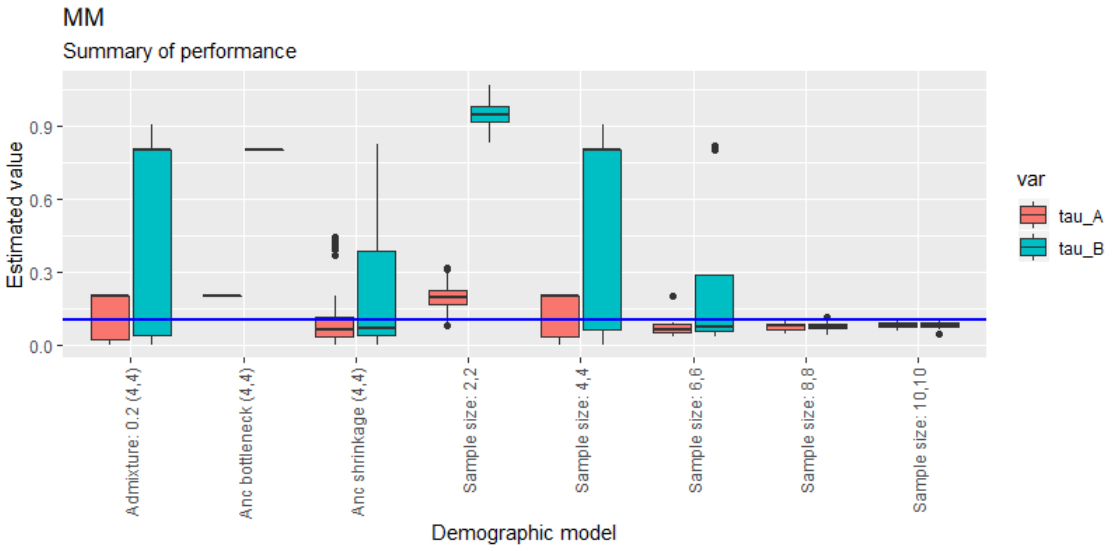


Figure 6: Summary of all trimmed boxplots from the runs of MM on different samples sizes and demographic models.

3.3 Tests on real data

The runs of the normal TT-beta method on the previously mentioned population data gave rise to divergence times estimates as shown in Appendix D, and interestingly displays some of the tendencies already seen on the runs of simulated data. Both time estimate alternatives (T1 and T2) result in more or less the same values and some combinations of populations have managed to result in negative divergence times.

Observing the estimated a and b parameters from these tests (see Appendix D) also showed that almost all of them landed within the acceptable and expected intervals derived earlier, with a values just above 0 and b's around 1. The only exceptions of this came from the Neanderthal - Denisovan pairing, which gave negative values far from the others, and the Dai - Han one that failed to return any at all. Plotting all the resulting beta distributions from these estimates also gave interestingly non-diverse ancestral frequency spectra (see Appendix D).

4 Discussion

The aim of this project was to explore the possibility of developing a novel method for estimating past divergence times of populations. The results show a real possibility behind the theory of the two investigated methods, even though the circumstances of the project resulted in only TT-beta being able to be properly tested and assessed.

4.1 TT-beta as a new method

The successful implementation of the TT-beta method was not something certain beforehand as it had not been explored practically before. So just the fact that it is able to give values in the same region as unmodified TT and TTo is immediately a good sign for the possibilities of using beta distribution to model ancestral frequency spectra. This is especially the case as the numerical methods tested for the purpose of finding these values were limited to only those available to the SciPy package. Extending the search of numerical methods, or possibly even developing one tailor made for the task at hand, could have beneficial effects on the running time and accuracy.

4.1.1 Run time and accuracy

The performance of the TT-beta method also highlighted some issues that could stem from the reliance on numerical solvers: relatively slow running times and a certain volatility that could result in extreme instances of incorrect estimations. Like mentioned before, this could be improved with the choice of implementation, but it can never be faster than the purely formula based TT and TTo alternatives. The largest effect of this extra running time will be seen in cases that require large number of runs, but single use cases won't notice this effect as much as the run time at most only is a few tenths of a second longer. The volatility of the TT-beta results are probably easier to fix with either the fine tuning of the numerical method used or by increasing the number of different start guesses leading to more robust estimates.

4.1.2 Compared to TT and TTo

When excluding the outliers, the performance of the TT-beta method really resembled those of TT and TTo (see Appendix C). All methods performed well on the demographic models with varying admixture between the populations directly after the split, with only some loss of accuracy as the amount of admixture increases. The effect of changes in the daughter population sizes was however made clear, as a bottleneck event occurring in population 2 resulted in averages far below the true value. But with the clear distinction that both TTo and TT-beta were able to handle it better than TT. A bottleneck in the ancestral population, on the other hand, proved to be less of a problem for the methods. Effects on performance was then later seen again as the model with a declining ancestral population seemed to greatly affect TT, but not TTo or TT-beta as much. In summary, these results points towards TT-beta at least replicating the performance of TTo, and improves on the results of the normal TT method.

One major difference of the TT-beta version is, however, the seemingly increased chance of failing to give any estimates at all. A decent chunk (approximately 15%) was constantly returned as NaN-values, for the most part completely unrelated to which demographic model used. It should however be noted that this didn't seem to impact the overall distribution of values, but should be kept in mind when running the method.

4.1.3 Runs on real data

The tests on real population data cemented a lot of the different observations seen earlier. Some populations produced negative time estimates (which of course is an error), one pairing failed to give any estimates at all and one gave values far from what you would expect. But one of the most interesting findings came from plotting the resulting

beta distributions from the estimated a and b values (see Appendix D). It seems that all populations gave estimations very similar to each other, with surprisingly little variation. The reason for this is unclear, and could either be a sign of worry that the model is too constrictive and biased, or that it is a common pattern within the observed populations. This is however impossible to know with the limited information at hand as further testing would be required to determine such things.

4.2 MM and its potential

The implementation and testing of the MM method was quite limited as a result of the time available at that stage of the project. Because of this, it is hard to give an honest and concrete assessment on it as a method for estimating divergence times. What can be said is that potential absolutely exists, and that the central theory behind it have proven to work practically in the form of TT-beta.

A few trends could also be seen from the small set of runs conducted on the partial MM implementation estimating the τ values for each of the two daughter populations. Sample size seemed to have the largest effect on the estimates with all tests with fewer than 8 samples per population experiencing a spread out set of τ values, with τ_B seemingly being more inaccurate. This probably comes from the fact that τ_B worked with a starting guess of 0.8 (compared to 0.2 of τ_A) and seemed to get stuck there in instances with fewer samples. But when using 8 or more they instead seem more accurate and precise. The small amount of admixture didn't seem to have any notable effect on the estimates with values almost identical to those without admixture, but larger demographic events like a declining ancestral population had a large impact on both mean value and spread. These conclusions should however be taken with grain of salt as the larger runs required lowering the number of replicates from 1000 to 50, partly because the MM method became significantly slower when scaling the problem size.

4.3 Future work

The work within this area is far from done, with more needed to produce a method in the way as described in the beginning if this report. A first thing would be to perform more extensive analyses of both TT-beta and the current version of MM in order to get a better grasp of how the methods perform in more complex demographic model. It would also be interesting to test different sample configurations of samples per population in the case of MM. These tests would also be a good opportunity to try other sets of numeri-

cal methods, or alternatively to optimize ones currently in use as that could be further explored.

The next step would then be to try to implement the rest of the theory behind MM in order to get the full method up and running. This would hopefully reach the original goal of the report to produce a novel method with the potential of more statistically significant divergence time estimates and the ability infer past changes in the daughter population sizes.

4.4 Conclusion

The limited MM implementation, and especially the associated TT-beta method, has shown potential for the use of beta distributions to create novel methods for estimating divergence times. TT-beta was able to be fully implemented in Python and gave promising results, performing on at least the level of classical TTo. But coming with the cost of being slightly slower and having a decreased reliability. Tests on real human population data cemented these findings and shed light on possible issues with negative estimates being returned. Time limitations resulted in only a partial MM implementation, but it still achieved functionality and could give quite accurate estimations when the sample size used was large enough.

5 Acknowledgements

The computations were enabled by resources in project [SNIC 2021/22-197] provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX, partially funded by the Swedish Research Council through grant agreement no. 2018-05973

I would personally like to thank my supervisor, Per Sjödin, for the continuous support and advice during the project. Thanks also go out to James McKenna for helping me and providing with starting resources for setting up the simulations on msprime. Would also like to thank my subject reader, Matthew Webster, for the critical feedback and guidance to keep the project focused. Lastly, I would also like to thank my examiner Pascal Milesi, and the course administrator Lena Henriksson for always being available and answering any potential questions that arose.

References

Balding DJ, Nichols RA. 1995. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96: 3–12.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. Genetics 162: 2025–2035.

Kelleher J, Etheridge AM, McVean G. 2016. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLOS Computational Biology 12: e1004842. Publisher: Public Library of Science.

Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G. 2019. Inferring whole-genome histories in large population datasets. Nature Genetics 51: 1330–1338. Number: 9 Publisher: Nature Publishing Group.

Kingman JFC. 1982. The coalescent. Stochastic Processes and their Applications 13: 235–248.

Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, Filippo Cd, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV, Derevianko AP, Patterson N, Andrés AM, Eichler EE, Slatkin M, Reich D, Kelso J, Pääbo S. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science 338: 222–226. Publisher: American Association for the Advancement of Science Section: Research Article.

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, Li H, Mallick S, Dannemann M, Fu Q, Kircher M, Kuhlwilm M, Lachmann M, Meyer M, Ongyerth M, Siebauer M, Theunert C, Tandon A, Moorjani P, Pickrell J, Mullikin JC, Vohr SH, Green RE, Hellmann I, Johnson PLF, Blanche H, Cann H, Kitzman JO, Shendure J, Eichler EE, Lein ES, Bakken TE, Golovanova LV, Doronichev VB, Shunkov MV, Derevianko AP, Viola B, Slatkin M, Reich D, Kelso J, Pääbo S. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505: 43–49.

Schlebusch CM, Malmström H, Günther T, Sjödin P, Coutinho A, Edlund H, Munters AR, Vicente M, Steyn M, Soodyall H, Lombard M, Jakobsson M. 2017. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. Science 358: 652–655. Publisher: American Association for the Advancement of Science Section: Report.

Schraiber JG, Akey JM. 2015. Methods and models for unravelling human evolutionary history. Nature Reviews Genetics 16: 727–740. Number: 12 Publisher: Nature Publishing Group.

Sjödin P, McKenna J, Jakobsson M. 2021. Estimating divergence times from DNA sequences. Genetics .

Slatkin M. 1996. Gene genealogies within mutant allelic classes. Genetics 143: 579–587.

Tavaré S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. Theoretical Population Biology 26: 119–164.

Wakeley J. 2009. Coalescent Theory: An Introduction. Roberts & Company Publishers, Greenswood Village, Colorado, 1st edition.

Appendix A - Theory behind TT-beta

June 2, 2021

1 Intro

The density of a Beta distribution with parameters a and b is

$$\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$$

This is a density for a stochastic variable in the interval $(0,1)$ so that the integral from 0 to 1 is 1. It follows that

$$\begin{aligned}&\int_0^1 x^{m+a-1}(1-x)^{l+b-1}dx\\&= B(m+a,l+b)\int_0^1 \frac{x^{m+a-1}(1-x)^{l+b-1}}{B(m+a,l+b)}dx = B(m+a,l+b)\end{aligned}$$

In a population with *polymorphic* derived frequency distribution modelled by a $Beta(a,b)$ distribution, the probability to obtain m derived and l ancestral alleles in a sample of size $m+l$ at a polymorphic site is

$$\begin{aligned}&\binom{m+l}{m}\int_0^1 x^m(1-x)^l\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}dx\\&= \binom{m+l}{m}\frac{1}{B(a,b)}\int_0^1 x^{m+a-1}(1-x)^{l+b-1}dx\\&= \binom{m+l}{m}\frac{B(m+a,l+b)}{B(a,b)}\end{aligned}$$

Since

$$\Gamma(n+x) = (n+x-1)\Gamma(n+x-1)$$

so that

$$\begin{aligned} B(m+a,l+b) &= \frac{\Gamma(m+a)\Gamma(l+b)}{\Gamma(m+l+a+b)} \\ &= \frac{(m+a-1)(l+b-1)}{(m+l+a+b)}\frac{\Gamma(m+a-1)\Gamma(l+b-1)}{\Gamma(m+l+a+b-1)} \\ &= \frac{\Pi_{i=0}^{m-1}(a+i)\Pi_{j=0}^{l-1}(b+j)}{\Pi_{i=0}^{m+l-1}(a+b+i)}\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \\ &= \frac{\Pi_{i=0}^{m-1}(a+i)\Pi_{j=0}^{l-1}(b+j)}{\Pi_{i=0}^{m+l-1}(a+b+i)}B(a,b) \end{aligned}$$

we get

$$\binom{m+l}{m}\int_0^1 x^m(1-x)^l\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}dx = \binom{m+l}{m}\frac{\Pi_{i=0}^{m-1}(a+i)\Pi_{j=0}^{l-1}(b+j)}{\Pi_{i=0}^{m+l-1}(a+b+i)}$$

Reformulating TT

Relating to a_{nm} (with $n = m + l$) in the TT-paper we get

$$
\begin{aligned}
a_{21} &= \binom{2}{1}\frac{\Pi_{i=0}^{0}(a+i)\Pi_{j=0}^{0}(b+j)}{\Pi_{i=0}^{1}(a+b+i)} = 2\frac{ab}{(a+b)(a+b+1)} \\
a_{31} &= \binom{3}{1}\frac{\Pi_{i=0}^{0}(a+i)\Pi_{j=0}^{1}(b+j)}{\Pi_{i=0}^{2}(a+b+i)} = 3\frac{ab(b+1)}{(a+b)(a+b+1)(a+b+2)} \\
&= \frac{3}{2}\frac{b+1}{a+b+2}a_{21} \\
a_{32} &= \binom{3}{2}\frac{\Pi_{i=0}^{1}(a+i)\Pi_{j=0}^{0}(b+j)}{\Pi_{i=0}^{2}(a+b+i)} = 3\frac{a(a+1)b}{(a+b)(a+b+1)(a+b+2)} \\
&= \frac{3}{2}\frac{a+1}{a+b+2}a_{21} \\
a_{41} &= \binom{4}{1}\frac{\Pi_{i=0}^{0}(a+i)\Pi_{j=0}^{2}(b+j)}{\Pi_{i=0}^{3}(a+b+i)} = 4\frac{ab(b+1)(b+2)}{(a+b)(a+b+1)(a+b+2)(a+b+3)} \\
&= \frac{4}{2}\frac{(b+1)(b+2)}{(a+b+2)(a+b+3)}a_{21} = \frac{4}{3}\frac{b+2}{a+b+3}a_{31} \\
a_{42} &= \binom{4}{2}\frac{\Pi_{i=0}^{1}(a+i)\Pi_{j=0}^{1}(b+j)}{\Pi_{i=0}^{3}(a+b+i)} = 6\frac{a(a+1)b(b+1)}{(a+b)(a+b+1)(a+b+2)(a+b+3)} \\
&= \frac{6}{2}\frac{(a+1)(b+1)}{(a+b+2)(a+b+3)}a_{21} = \frac{6}{3}\frac{a+1}{a+b+3}a_{31} = \frac{6}{3}\frac{b+1}{a+b+3}a_{32} \\
a_{43} &= \binom{4}{3}\frac{\Pi_{i=0}^{2}(a+i)\Pi_{j=0}^{0}(b+j)}{\Pi_{i=0}^{3}(a+b+i)} = 4\frac{a(a+1)(a+2)b}{(a+b)(a+b+1)(a+b+2)(a+b+3)} \\
&= \frac{4}{2}\frac{(a+1)(a+2)}{(a+b+2)(a+b+3)}a_{21} = \frac{4}{3}\frac{a+2}{a+b+3}a_{32}
\end{aligned}
$$

This is furthermore consistent with (from the TT-paper):

$$
\begin{aligned}
a_{21} &= \frac{1}{2}a_{41} + \frac{2}{3}a_{42} + \frac{1}{2}a_{43} \\
a_{31} &= \frac{3}{4}a_{41} + \frac{1}{2}a_{42} \\
a_{32} &= \frac{1}{2}a_{42} + \frac{3}{4}a_{43}
\end{aligned}
$$

Plugging in the expressions for a_{41}, a_{42} and a_{43} in the expressions for the probability of the different sample configurations making the fewest assumptions (see

TT paper) we get:

$$
\begin{aligned}
p_{1,0} =& 2(1-\alpha_1)\mu\nu_1 + 2\alpha_1\mu t_1 \\
&+ 2\alpha_1 \frac{ab(b+1)}{(a+b)(a+b+1)(a+b+2)} \left(1 - \alpha_2 \frac{a+1}{a+b+3}\right) \\
p_{0,1} =& 2(1-\alpha_2)\mu\nu_2 + 2\alpha_2\mu t_2 \\
&+ 2\alpha_2 \frac{ab(b+1)}{(a+b)(a+b+1)(a+b+2)} \left(1 - \alpha_1 \frac{a+1}{a+b+3}\right) \\
p_{2,0} =& (1-\alpha_1)\mu t_1 - (1-\alpha_1)\mu\nu_1 \\
&+ \frac{ab}{(a+b)(a+b+1)} \left(1 - \alpha_1 \frac{b+1}{a+b+2} - \alpha_2 \frac{a+1}{a+b+2} + \alpha_1\alpha_2 \frac{(a+1)(b+1)}{(a+b+2)(a+b+3)}\right) \\
p_{0,2} =& (1-\alpha_2)\mu t_2 - (1-\alpha_2)\mu\nu_2 \\
&+ \frac{ab}{(a+b)(a+b+1)} \left(1 - \alpha_2 \frac{b+1}{a+b+2} - \alpha_1 \frac{a+1}{a+b+2} + \alpha_1\alpha_2 \frac{(a+1)(b+1)}{(a+b+2)(a+b+3)}\right) \\
p_{1,1} =& 4\alpha_1\alpha_2 \frac{a(a+1)b(b+1)}{(a+b)(a+b+1)(a+b+2)(a+b+3)} \\
p_{2,1} =& 2\alpha_2 \frac{a(a+1)b}{(a+b)(a+b+1)(a+b+2)} \\
&- 2\alpha_1\alpha_2 \frac{a(a+1)b(b+1)}{(a+b)(a+b+1)(a+b+2)(a+b+3)} \\
p_{1,2} =& 2\alpha_1 \frac{a(a+1)b}{(a+b)(a+b+1)(a+b+2)} \\
&- 2\alpha_1\alpha_2 \frac{a(a+1)b(b+1)}{(a+b)(a+b+1)(a+b+2)(a+b+3)}
\end{aligned}
$$

Implying that

$$
\begin{aligned}
p_{1,0} + \frac{1}{2}p_{1,1} =& 2(1-\alpha_1)\mu\nu_1 + 2\alpha_1\mu t_1 + 2\alpha_1\frac{ab(b+1)}{(a+b)(a+b+1)(a+b+2)} \\
p_{0,1} + \frac{1}{2}p_{1,1} =& 2(1-\alpha_2)\mu\nu_2 + 2\alpha_2\mu t_2 + 2\alpha_2\frac{ab(b+1)}{(a+b)(a+b+1)(a+b+2)} \\
p_{2,0} + \frac{1}{2}p_{2,1} =& (1-\alpha_1)\mu t_1 - (1-\alpha_1)\mu\nu_1 - \alpha_1\frac{ab(b+1)}{(a+b)(a+b+1)(a+b+2)} + \frac{ab}{(a+b)(a+b+1)} \\
p_{0,2} + \frac{1}{2}p_{1,2} =& (1-\alpha_2)\mu t_2 - (1-\alpha_2)\mu\nu_2 - \alpha_2\frac{ab(b+1)}{(a+b)(a+b+1)(a+b+2)} + \frac{ab}{(a+b)(a+b+1)} \\
p_{2,1} + \frac{1}{2}p_{1,1} =& 2\alpha_2\frac{a(a+1)b}{(a+b)(a+b+1)(a+b+2)} \\
p_{1,2} + \frac{1}{2}p_{1,1} =& 2\alpha_1\frac{a(a+1)b}{(a+b)(a+b+1)(a+b+2)} \\
p_{1,1} =& 4\alpha_1\alpha_2\frac{a(a+1)b(b+1)}{(a+b)(a+b+1)(a+b+2)(a+b+3)}
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{1}{2}p_{1,0} + \frac{1}{4}p_{1,1} + p_{2,0} + \frac{1}{2}p_{2,1} =& \mu t_1 + \frac{ab}{(a+b)(a+b+1)} \\
\frac{1}{2}p_{0,1} + \frac{1}{4}p_{1,1} + p_{0,2} + \frac{1}{2}p_{1,2} =& \mu t_2 + \frac{ab}{(a+b)(a+b+1)}
\end{aligned}
$$

Reformulating TTo

If α_1 and α_2 is known (as under the TTo set up), we get the equations

$$
\begin{aligned}
r_1 \equiv \frac{1}{4}\frac{2p_{2,1} + p_{1,1}}{\alpha_2} =& \frac{a(a+1)b}{(a+b)(a+b+1)(a+b+2)} \\
r_2 \equiv \frac{1}{4}\frac{2p_{1,2} + p_{1,1}}{\alpha_1} =& \frac{a(a+1)b}{(a+b)(a+b+1)(a+b+2)} \\
s \equiv \frac{1}{4}\frac{p_{1,1}}{\alpha_1\alpha_2} =& \frac{a(a+1)b(b+1)}{(a+b)(a+b+1)(a+b+2)(a+b+3)}
\end{aligned}
$$

writing r for either r_1 or r_2

$$r = \frac{a(a+1)b}{(a+b)(a+b+1)(a+b+2)}$$
$$s = \frac{a(a+1)b(b+1)}{(a+b)(a+b+1)(a+b+2)(a+b+3)}$$

Appendix B - Theory behind MM

June 2, 2021

1 Intro

The density of a Beta distribution with parameters a and b is

$$\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$$

This is a density for a stochastic variable in the interval $(0,1)$ so that the integral from 0 to 1 is 1. It follows that

$$\int_0^1 x^{m+a-1}(1-x)^{l+b-1}dx$$
$$= B(m+a,l+b)\int_0^1 \frac{x^{m+a-1}(1-x)^{l+b-1}}{B(m+a,l+b)}dx = B(m+a,l+b)$$

Also

$$\tau = \int_0^t \frac{dz}{2N(z)}$$

where t is the time in generations away from population A and $N(z)$ is the diploid population size z generations away from population A (in the same direction as t) We also rely on the being the probability of there being k ancestors at time τ to a sample of n gene-copies, $g_{n,k}(\tau)$. Specifically

$$g_{n,k}(\tau) = \frac{1}{\binom{k}{2}}\sum_{i=k}^{n} e^{-\binom{i}{2}\tau}\binom{i}{2}\prod_{l=k,l\neq i}^{n}\frac{\binom{l}{2}}{\binom{l}{2}-\binom{i}{2}}$$

for $n < k \leq 2$ with special cases $g_{n,n}(\tau) = e^{-\binom{n}{2}\tau}$ and $g_{n,1}(\tau) = 1-\sum_{k=2}^{n} g_{n,k}(\tau)$ ([**?**, **?**]).

2 Sample distribution of m_A derived and l_A ancestral in one and m_B derived and l_B ancestral in another population

Now we center the derivation on the derived frequency in the population where population A and population B eventually merge (population C). Write a and

b for the shape parameters of the frequency distribution in population C. We also define

$$h(m,l;i,j) = \frac{\binom{m-1}{i-1}\binom{l-1}{j-1}}{\binom{m+l-1}{i+j-1}}$$

which is the probability of m derived lineages when there are $m+l$ lineages given that there were i derived lineages when there were $i+j$ lineages ([**?**]).

2.1 Ignoring branch specific mutations

To derive $P\left(S_B = \{m_B, l_B\} \wedge S_A = \{m_A, l_A\}\right)$, first ignore the possibility that the derived mutation is younger than population C (note that this is not even possible if both $m_B > 0$ and $m_A > 0$) and write

$$\begin{aligned}
&P\left(S_B = \{m_B, l_B\} \wedge S_A = \{m_A, l_A\}\right) \\
&= \int_0^1 \left(\sum_{i_A=0}^{m_A} \sum_{j_A=0}^{l_A} g_{n_A, i_A+j_A}(\tau_A) h(m_A, l_A; i_A, j_A) \binom{i_A+j_A}{i_A} x^{i_A} (1-x)^{j_A} \right) \\
&\times \left(\sum_{i_B=0}^{m_B} \sum_{j_B=0}^{l_B} g_{n_B, i_B+j_B}(\tau_B) h(m_B, l_B; i_B, j_B) \binom{i_B+j_B}{i_B} x^{i_B} (1-x)^{j_B} \right) \\
&\times \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} dx \\
&= \frac{1}{B(a,b)} \sum_{i_A=0}^{m_A} \sum_{j_A=0}^{l_A} \sum_{i_B=0}^{m_B} \sum_{j_B=0}^{l_B} g_{n_A, i_A+j_A}(\tau_A) g_{n_B, i_B+j_B}(\tau_B) \\
&\times h(m_A, l_A; i_A, j_A) h(m_B, l_B; i_B, j_B) \binom{i_A+j_A}{i_A} \binom{i_B+j_B}{i_B} \\
&\times \int_0^1 x^{a+i_A+i_B-1} (1-x)^{b+j_A+j_B-1} dx \\
&= \sum_{i_A=0}^{m_A} \sum_{j_A=0}^{l_A} \sum_{i_B=0}^{m_B} \sum_{j_B=0}^{l_B} \frac{B(a+i_A+i_B, b+j_A+j_B)}{B(a,b)} h^*(\tau_A; m_A, l_A; i_A, j_A) h^*(\tau_B; m_B, l_B; i_B, j_B)
\end{aligned}$$

with

$$h^*(\tau; m, l; i, j) = \binom{i+j}{i} h(m, l, i, j) g_{m+l, i+j}(\tau)$$

2.2 Only considering branch specific mutations

If we ignore derived variants that are older than the split, the only sample configurations that are possible are those with at least one of m_A m_B being 0. Thus we only need to derive the probability for $P(S = \{m, l\})$ given that the mutation occurs on one of the lineages in the genealogy more recently than τ.

For this purpose define μ to be the mutation rate per generation and locus, to be the expected time *in generations* with k lineages in theμ to be the mutation rate per generation and locus, $\nu_{k,i}$ to be the time *in generations* spent with k lineages given that there are i lineages left after τ scaled time units, and ν_k to be the expected time *in generations* with k lineages in the genealogy within τ units of time. Then

$$\nu_k = \sum_{i=1}^{k} \nu_{k,i} g_{n,i}(\tau)$$

and for $l > 0$

$$\begin{aligned} P(S = \{m, l\}) &= \sum_{k=2}^{l+1} h(m, l; 1, k-1) k\mu \sum_{i=1}^{k} \nu_{k,i} g_{n,i}(\tau) \\ &= \mu \sum_{k=2}^{l+1} h(m, l; 1, k-1) k\nu_k = \mu \sum_{k=1}^{l} h(m, l; 1, k)(k+1)\nu_{k+1} \end{aligned}$$

and for $l = 0$

$$\begin{aligned} P(S = \{n, 0\}) &= h(n, 0; 1, 0)\mu\nu_{1,1} g_{n,1}(\tau) \\ &= \mu h(n, 0; 1, 0)\nu_1 = \mu\nu_1 \end{aligned}$$

so that we can write

$$P(S = \{m, l\}) = \mu \sum_{k=0}^{l} h(m, l; 1, k)(k+1)\nu_{k+1}$$

for $l \geq 0$.
We have for $0 < m_B \leq n_B$, $m_A = 0$

$$\begin{aligned} P(S_B = \{m_B, l_B\} \wedge S_A = \{0, n_A\}) &= \mu \sum_{k=0}^{l_B} h(m_B, l_B; 1, k)(k+1)\nu_{k+1}^B \\ &= \sum_{k=0}^{l_B} h(m_B, l_B; 1, k) V_{k+1}^B \end{aligned}$$

with $V_k^B = k\mu\nu_k^B$.

For $0 < m_A \leq n_A$, $m_B = 0$

$$P(S_B = \{0, n_B\} \wedge S_A = \{m_A, l_A\}) = \sum_{k=0}^{l_A} h(m_A, l_A; 1, k) V_{k+1}^A$$

with $V_k^A = k\mu\nu_k^A$.

2.3 Full solution

Here we make the key assumption that the probability of both picking a derived variant in the ancestral population C and having a branch specific mutation in the genealogy is vanishingly small.

For $m_A = 0$ and $0 < m_B \leq n_B$

$$
\begin{aligned}
&P(S_B = \{m_B, l_B\} \wedge S_A = \{0, n_A\}) \\
&\quad = \sum_{k=0}^{l_B} h(m_B, l_B; 1, k) V_{k+1}^B \\
&\quad + \sum_{i_A=0}^{0} \sum_{j_A=0}^{n_A} \sum_{i_B=0}^{m_B} \sum_{j_B=0}^{l_B} \frac{B(a + i_A + i_B, b + j_A + j_B)}{B(a,b)} h^*(\tau_A; m_A, l_A; i_A, j_A) h^*(\tau_B; m_B, l_B; i_B, j_B) \\
&= \sum_{k=0}^{l_B} h(m_B, l_B; 1, k) V_{k+1}^B + \sum_{j_A=0}^{n_A} \sum_{i_B=0}^{m_B} \sum_{j_B=0}^{l_B} \frac{B(a + i_B, b + j_A + j_B)}{B(a,b)} h^*(\tau_A; n_A, 0; i_A, 0) h^*(\tau_B; m_B, l_B; i_B, j_B) \\
&= \sum_{k=0}^{l_B} h(m_B, l_B; 1, k) V_{k+1}^B + \sum_{j_A=0}^{n_A} \sum_{i_B=0}^{m_B} \sum_{j_B=0}^{l_B} \frac{B(a + i_B, b + j_A + j_B)}{B(a,b)} g_{n_A, j_A}(\tau_A) h^*(\tau_B; m_B, l_B; i_B, j_B) \\
&= \sum_{k=0}^{l_B} h(m_B, l_B; 1, k) V_{k+1}^B + \sum_{j_A=1}^{n_A} \sum_{i_B=1}^{m_B} \sum_{j_B=0}^{l_B} \frac{B(a + i_B, b + j_A + j_B)}{B(a,b)} g_{n_A, j_A}(\tau_A) h^*(\tau_B; m_B, l_B; i_B, j_B)
\end{aligned}
$$

and for $m_B = 0$ and $0 < m_A \leq n_A$

$$
\begin{aligned}
&P(S_B = \{0, n_B\} \wedge S_A = \{m_A, l_A\}) \\
&= \sum_{k=0}^{l_A} h(m_A, l_A; 1, k) V_{k+1}^A + \sum_{i_A=1}^{m_A} \sum_{j_A=0}^{l_A} \sum_{j_B=1}^{n_B} \frac{B(a + i_A, b + j_A + j_B)}{B(a,b)} g_{n_B, j_B}(\tau_B) h^*(\tau_A; m_A, l_A; i_A, j_A)
\end{aligned}
$$

If none of the above is true (no sample is monomorphic for the ancestral allele) and that there is at least 1 ancestral gene copy in the sample, we have instead

$$
\begin{aligned}
&P(S_B = \{m_B, l_B\} \wedge S_A = \{m_A, l_A\}) \\
&\quad = \sum_{i_A=0}^{m_A} \sum_{j_A=0}^{l_A} \sum_{i_B=0}^{m_B} \sum_{j_B=0}^{l_B} \frac{B(a + i_A + i_B, b + j_A + j_B)}{B(a,b)} h^*(\tau_A; m_A, l_A; i_A, j_A) h^*(\tau_B; m_B, l_B; i_B, j_B)
\end{aligned}
$$

2.4 Estimation

In total, there will be $4 + n_A + n_B$ parameters to estimate ($a, b\tau_A, \tau_B$ as well as $V_1^A, \cdots V_{n_A}^A$ and $V_1^B, \cdots V_{n_B}^B$) and there are $(n_A + 1)(n_B + 1)$ possible sample

configurations. Of these,

1 monomorphic class: $\{0, n_B\}, \{0, n_A\} + \{n_B, 0\}, \{n_A, 0\}$

n_B configurations with information about V_i^B: $\{1, n_B - 1\}, \{0, n_A\} \cdots \{n_B, 0\}, \{0, n_A\}$

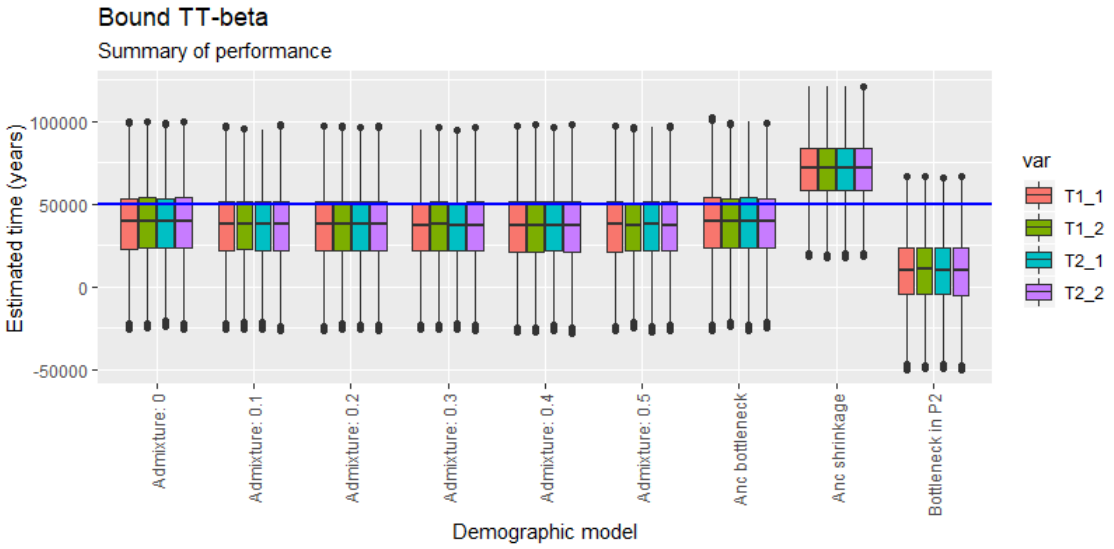
n_A configurations with information about V_i^A: $\{0, n_B\}, \{1, n_A - 1\} \cdots \{0, n_B\}, \{n_A, 0\}$

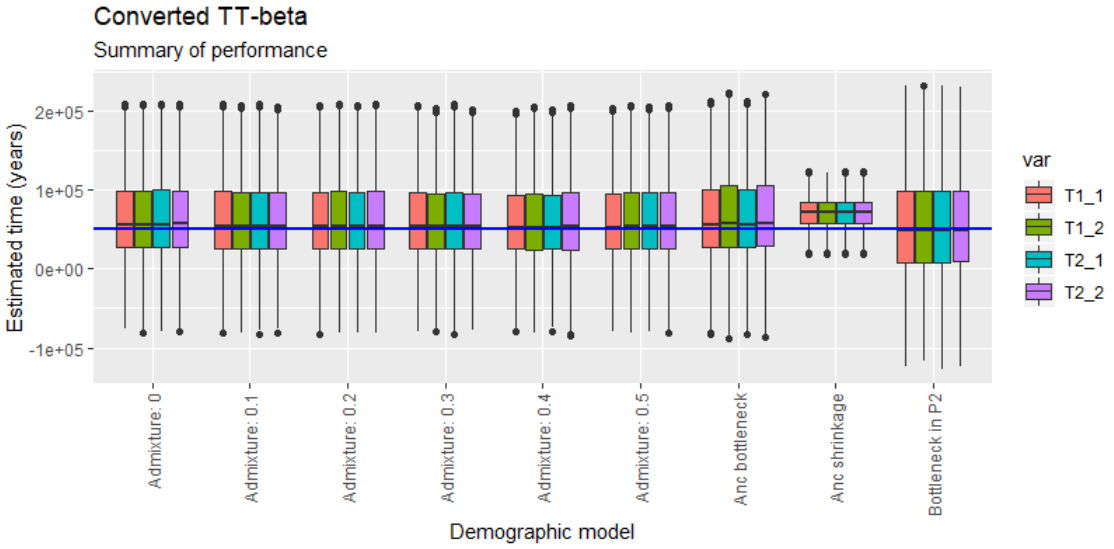
In other words, there are $(n_A + 1)(n_B + 1) - 1$ configurations to base our estimates on and of these, $(n_A + 1)(n_B + 1) - 2$ are polymorphic for both variants. Furthermore, $(n_A+1)(n_B+1)-2-n_B-n_A = n_An_B-1$ configurations are polymorphic in both samples and thus only informative of a, b, τ_B and τ_A. Thus, as long as $n_An_B - 1 \geq 4 \Rightarrow n_An_B \geq 5$ we have sufficient power to estimate a, b, τ_B and τ_A. As an example, if only diploid samples are considered, then one individual from each population (the TT setup) does not suffice but any sample with at least two individuals from one of the populations will. Also, if a haploid sample is retrieved from one population, more than two diploid individuals from the other population is required.
Assuming that $n_An_B \geq 5$, we should be able to estimate a, b, τ_A and τ_B so that the right side of

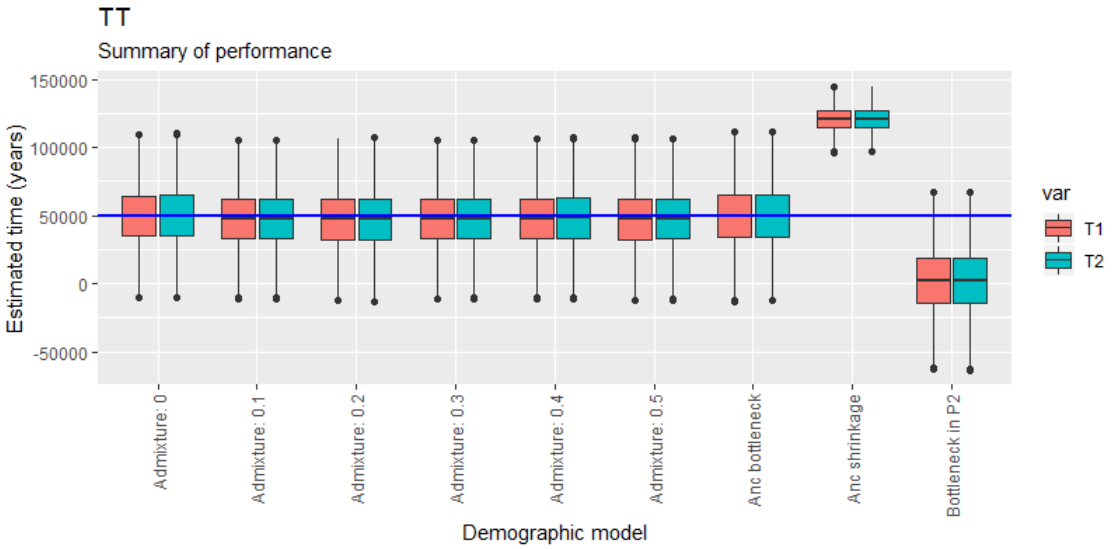
$$\begin{aligned}
&\sum_{k=0}^{l_B} h(m_B, l_B; 1, k) V_{k+1}^B \\
&\quad = P(S_B = \{m_B, l_B\} \wedge S_A = \{0, n_A\}) \\
&\qquad - \sum_{j_A=1}^{n_A} \sum_{i_B=1}^{m_B} \sum_{j_B=0}^{l_B} \frac{B(a + i_B, b + j_A + j_B)}{B(a, b)} g_{n_A, j_A}(\tau_A) h^*(\tau_B; m_B, l_B; i_B, j_B)
\end{aligned}$$

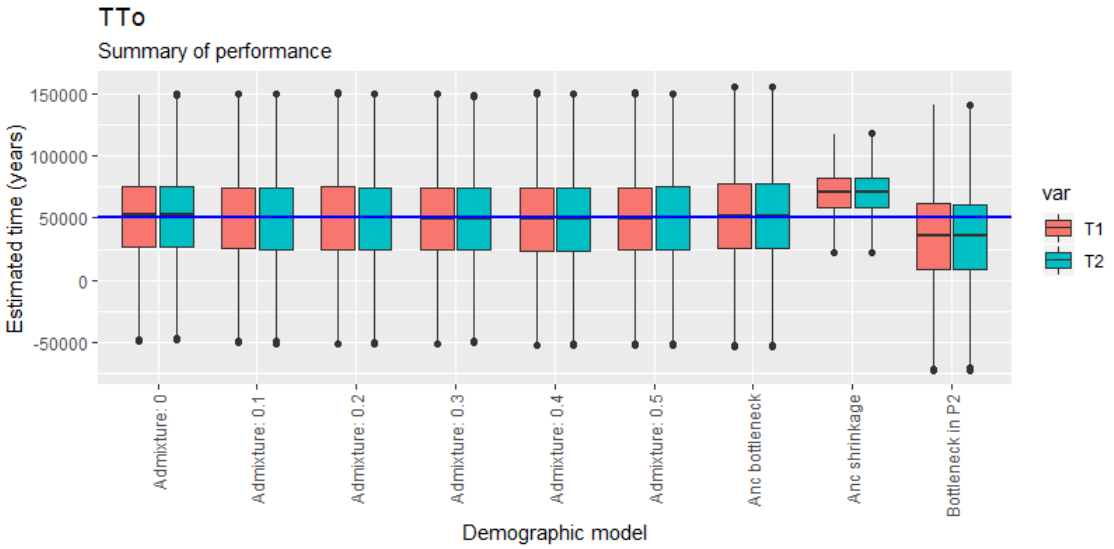
can be considered known when estimating V_i^B.

Appendix C - Performances of different TT versions

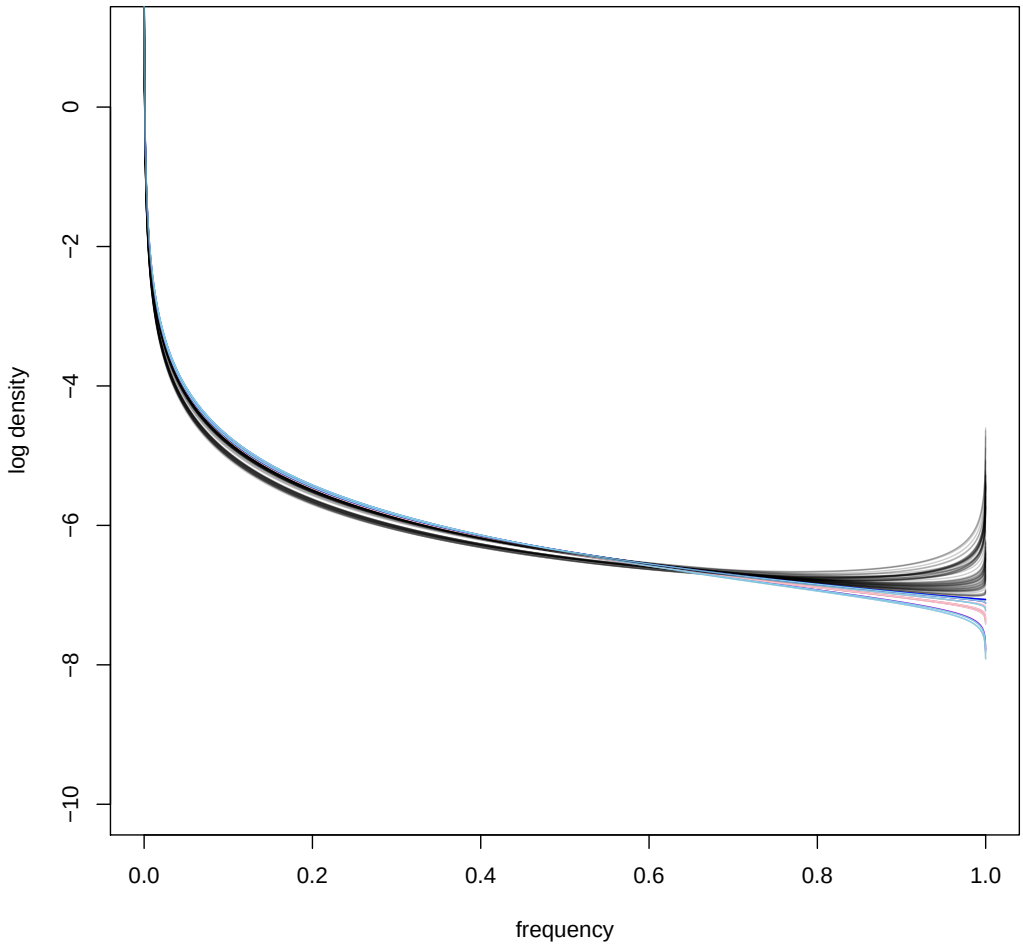


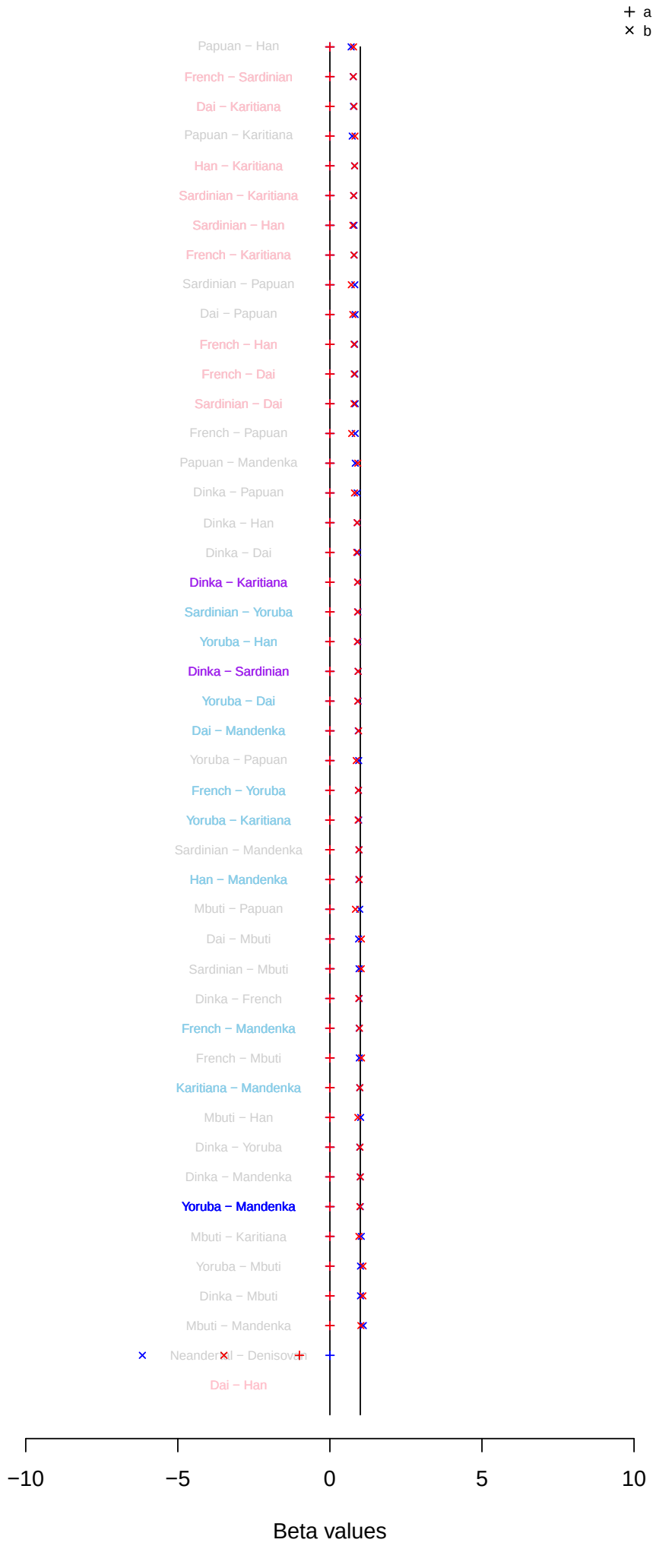


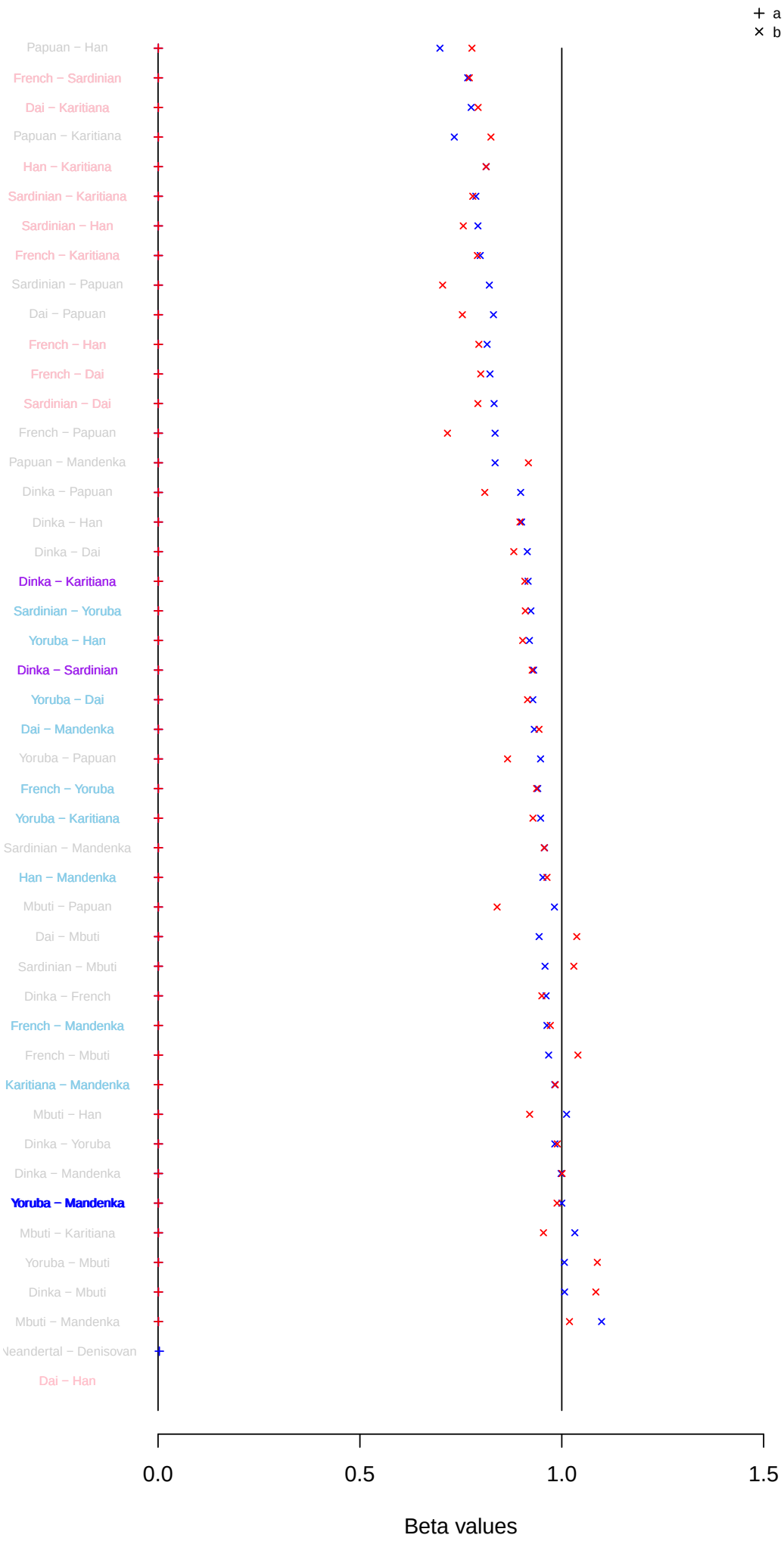

TT
Summary of performance
Estimated time (years)
150000
100000
50000
0
-50000
Admixture: 0
Admixture: 0.1
Admixture: 0.2
Admixture: 0.3
Admixture: 0.4
Admixture: 0.5
Anc bottleneck
Anc shrinkage
Bottleneck in P2
Demographic model
var
T1
T2

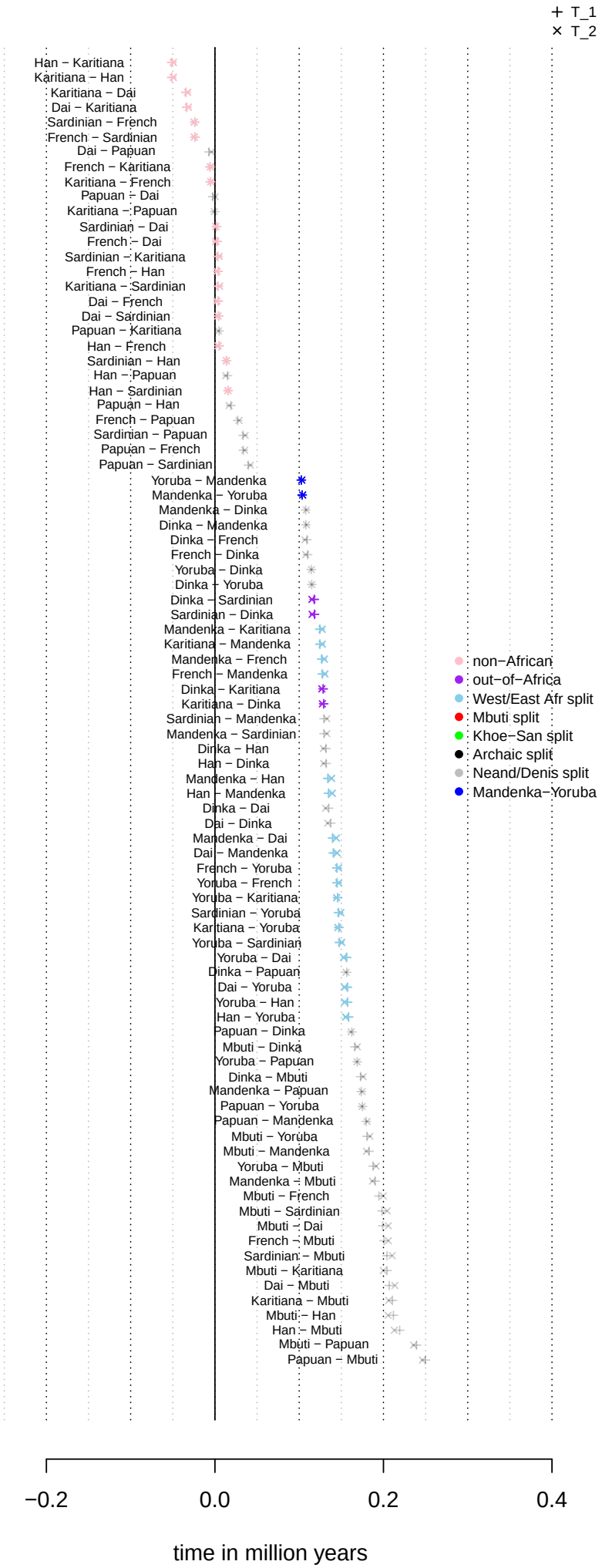

TTo
Summary of performance
Estimated time (years)
150000
100000
50000
0
-50000
Admixture: 0
Admixture: 0.1
Admixture: 0.2
Admixture: 0.3
Admixture: 0.4
Admixture: 0.5
Anc bottleneck
Anc shrinkage
Bottleneck in P2
Demographic model
var
T1
T2

Appendix D - Results from real data




a
b
Papuan − Han
French − Sardinian
Dai − Karitiana
Papuan − Karitiana
Han − Karitiana
Sardinian − Karitiana
Sardinian − Han
French − Karitiana
Sardinian − Papuan
Dai − Papuan
French − Han
French − Dai
Sardinian − Dai
French − Papuan
Papuan − Mandenka
Dinka − Papuan
Dinka − Han
Dinka − Dai
Dinka − Karitiana
Sardinian − Yoruba
Yoruba − Han
Dinka − Sardinian
Yoruba − Dai
Dai − Mandenka
Yoruba − Papuan
French − Yoruba
Yoruba − Karitiana
Sardinian − Mandenka
Han − Mandenka
Mbuti − Papuan
Dai − Mbuti
Sardinian − Mbuti
Dinka − French
French − Mandenka
French − Mbuti
Karitiana − Mandenka
Mbuti − Han
Dinka − Yoruba
Dinka − Mandenka
Yoruba − Mandenka
Mbuti − Karitiana
Yoruba − Mbuti
Dinka − Mbuti
Mbuti − Mandenka
Neanderthal − Denisovan
Dai − Han
−10
−5
0
5
10
Beta values


a
b
Papuan − Han
French − Sardinian
Dai − Karitiana
Papuan − Karitiana
Han − Karitiana
Sardinian − Karitiana
Sardinian − Han
French − Karitiana
Sardinian − Papuan
Dai − Papuan
French − Han
French − Dai
Sardinian − Dai
French − Papuan
Papuan − Mandenka
Dinka − Papuan
Dinka − Han
Dinka − Dai
Dinka − Karitiana
Sardinian − Yoruba
Yoruba − Han
Dinka − Sardinian
Yoruba − Dai
Dai − Mandenka
Yoruba − Papuan
French − Yoruba
Yoruba − Karitiana
Sardinian − Mandenka
Han − Mandenka
Mbuti − Papuan
Dai − Mbuti
Sardinian − Mbuti
Dinka − French
French − Mandenka
French − Mbuti
Karitiana − Mandenka
Mbuti − Han
Dinka − Yoruba
Dinka − Mandenka
Yoruba − Mandenka
Mbuti − Karitiana
Yoruba − Mbuti
Dinka − Mbuti
Mbuti − Mandenka
Neandertal − Denisovan
Dai − Han
0.0
0.5
1.0
1.5
Beta values


+ T_1
× T_2
Han − Karitiana
Karitiana − Han
Karitiana − Dai
Dai − Karitiana
Sardinian − French
French − Sardinian
Dai − Papuan
French − Karitiana
Karitiana − French
Papuan − Dai
Karitiana − Papuan
Sardinian − Dai
French − Dai
Sardinian − Karitiana
French − Han
Karitiana − Sardinian
Dai − French
Dai − Sardinian
Papuan − Karitiana
Han − French
Sardinian − Han
Han − Papuan
Han − Sardinian
Papuan − Han
French − Papuan
Sardinian − Papuan
Papuan − French
Papuan − Sardinian
Yoruba − Mandenka
Mandenka − Yoruba
Mandenka − Dinka
Dinka − Mandenka
Dinka − French
French − Dinka
Yoruba − Dinka
Dinka − Yoruba
Dinka − Sardinian
Sardinian − Dinka
Mandenka − Karitiana
Karitiana − Mandenka
Mandenka − French
French − Mandenka
Dinka − Karitiana
Karitiana − Dinka
Sardinian − Mandenka
Mandenka − Sardinian
Dinka − Han
Han − Dinka
Mandenka − Han
Han − Mandenka
Dinka − Dai
Dai − Dinka
Mandenka − Dai
Dai − Mandenka
French − Yoruba
Yoruba − French
Yoruba − Karitiana
Sardinian − Yoruba
Karitiana − Yoruba
Yoruba − Sardinian
Yoruba − Dai
Dinka − Papuan
Dai − Yoruba
Yoruba − Han
Han − Yoruba
Papuan − Dinka
Mbuti − Dinka
Yoruba − Papuan
Dinka − Mbuti
Mandenka − Papuan
Papuan − Yoruba
Papuan − Mandenka
Mbuti − Yoruba
Mbuti − Mandenka
Yoruba − Mbuti
Mandenka − Mbuti
Mbuti − French
Mbuti − Sardinian
Mbuti − Dai
French − Mbuti
Sardinian − Mbuti
Mbuti − Karitiana
Dai − Mbuti
Karitiana − Mbuti
Mbuti − Han
Han − Mbuti
Mbuti − Papuan
Papuan − Mbuti
non−African
out−of−Africa
West/East Afr split
Mbuti split
Khoe−San split
Archaic split
Neand/Denis split
Mandenka−Yoruba
−0.2
0.0
0.2
0.4
time in million years