AncestryProject

Leon Joshua Gensel

2022-09-09

Disclaimer:

If you want to run this report as .Rmd file, you will need your data organized the same way it is in the repository. Also you will need the tidyverse, ggplot2, reshape2 and betareg packages for R.

Introduction

Genetic Data is commonly used to reconstruct human history and paint a picture of ancient populations. Usually the genomic sequences are used to construct phylogenetic trees. Due to their uniparental inheritance and lack of recombination the mitochondrial DNA and the Y-chromosome are particullarly well suited for inferring population history based on their phylogenies (here maybe example studies).

Yet, there is no clear guidance on what the limitations of such approaches are, and how to avoid over-interpretation of such signal. The main aim of this study is to estimate the accuracy of using phylogenetic trees constructed from mtDNA (mitochondrial DNA) and MSY (Male-specific region of the Y-Chromosome) genomic data to infer information about a populations history. Furthermore I will try to estimate the source or part of the methodology which introduces the error in the final estimations.

To achieve this I ran forward simulations of an admixture event of two source populations with randomized admixture proportions and varying demographic parameters. These simulations output a recorded tree sequence of randomly sampled individuals alive in the final generation and the genetic data of their mtDNA and Y-chromosome. In addition the current proportions of the source populations in the admixed population were output in every generation. The genetic data then was used to construct a phylogenetic tree. Based on this tree the admixture proportions can be inferred. This allowed me to compare the inferred admixture proportions to the simulated ones in the current and initial state of the population. This gave insight into how demographic parameters (like the degree of divergence and the time since admixture) can limit the accuracy of this method. Furthermore I tried to quantitatively estimate how much error each factor (e.g. sampling of individuals, using the trees structure, constructing) contributed to the inferred proportions.

Program structure

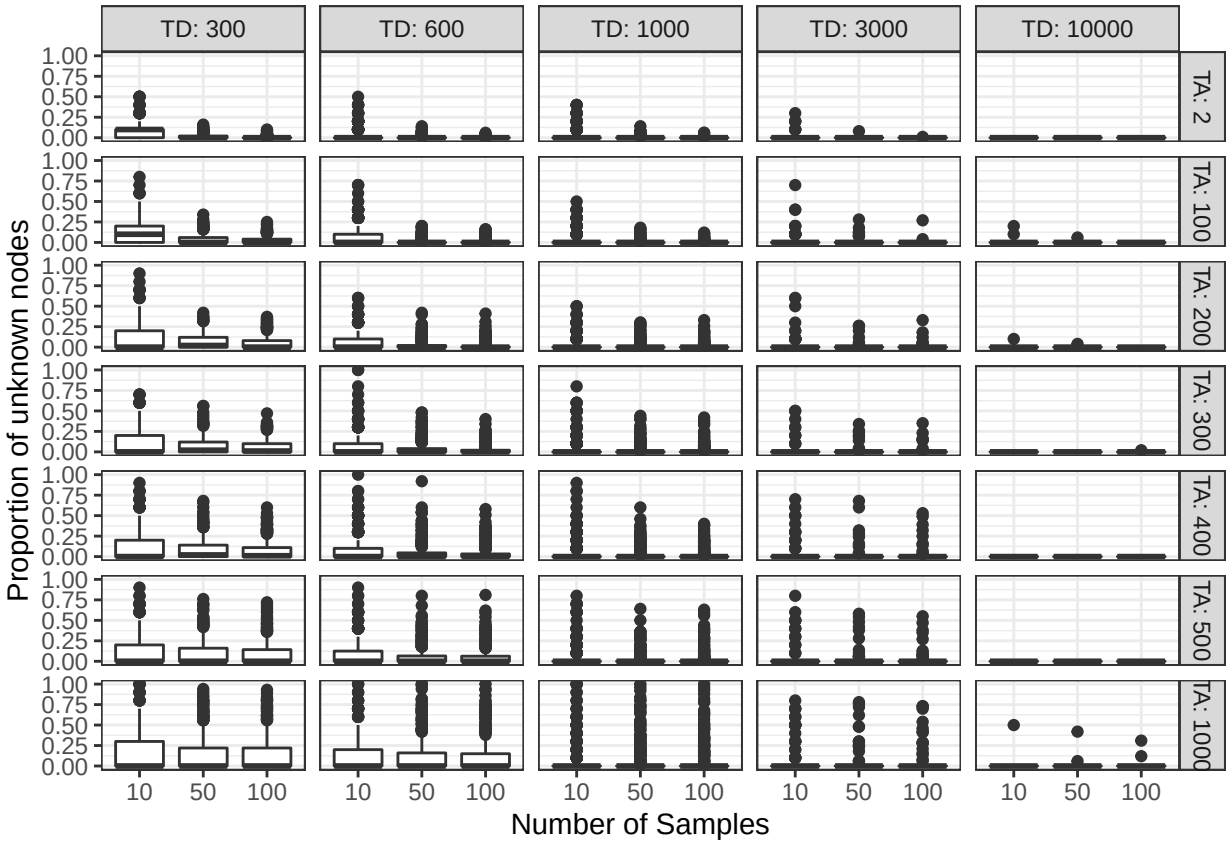
SLiM Simulations

Processing Pipeline

Exploratory analysis

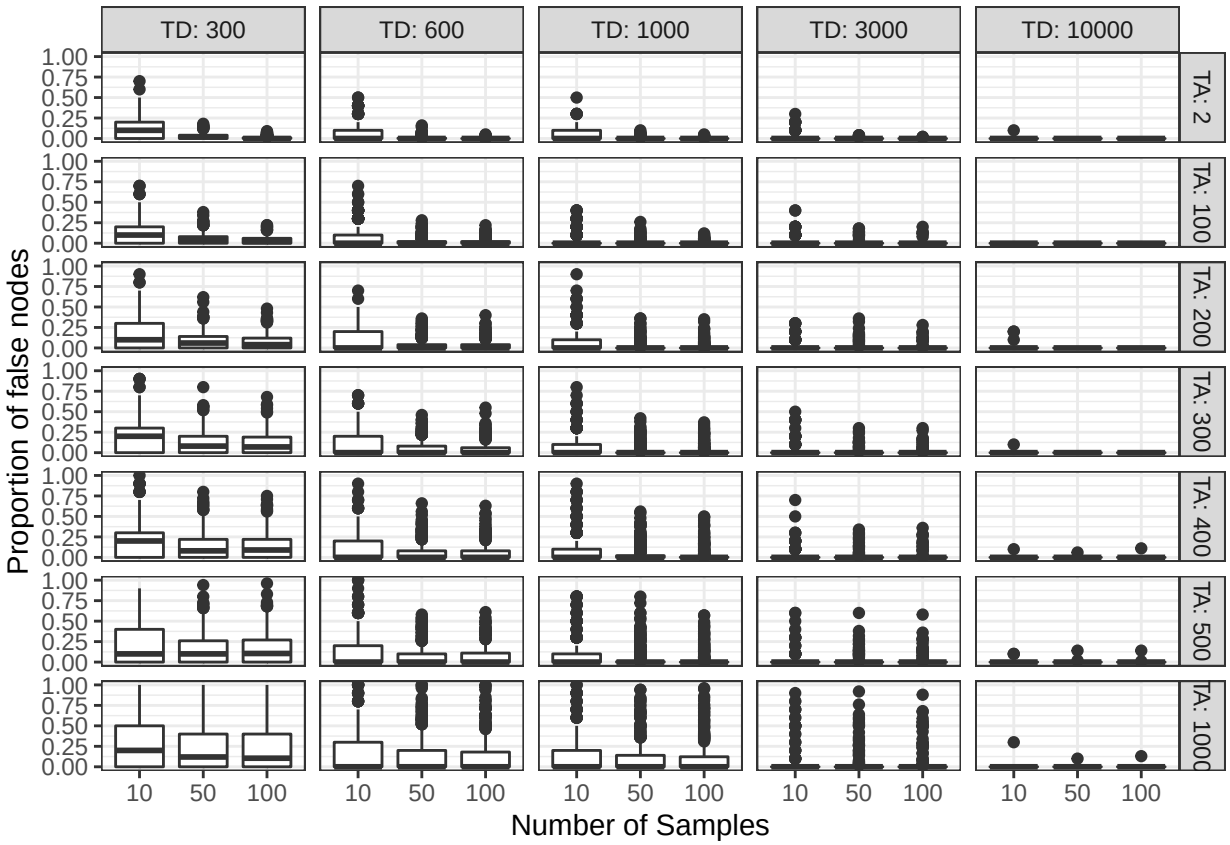
First I'll explore the genealogy data. For this all metrics were calculated from the recorded tree sequence of every simulation run. This means the trees reflect the true history of the sampled individuals. First we will take a look at unknown nodes. These are P3 nodes which cluster with any number of both P1 and P2 leafs and therefore cannot be assigned to one of the source populations with certainty.

In the following plot we can see the proportion of unknown nodes in P3 for different values of TD, TA and the number of sampled individuals.



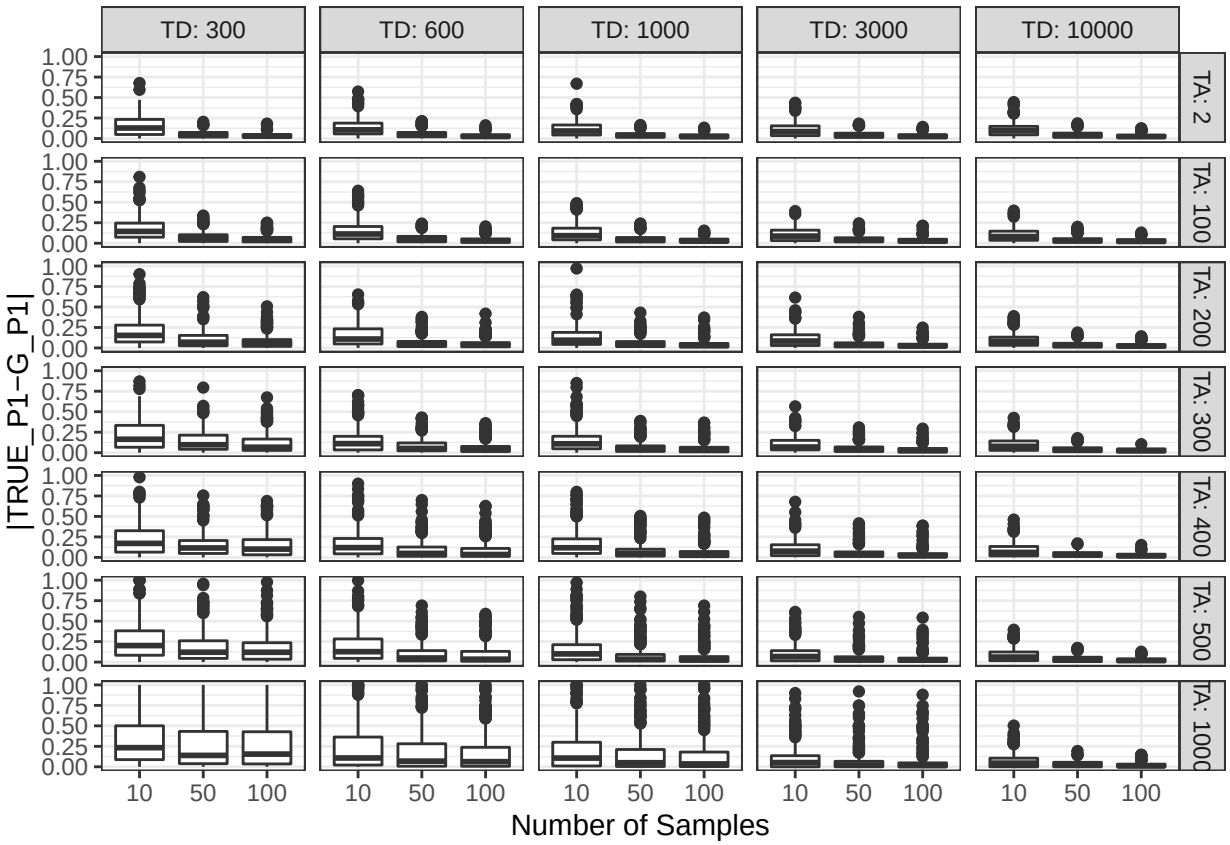
We can already see that to avoid unknown nodes you ideally would want high TD and samples, but low TA. Also overall the number of unknown nodes is low unless TD is very low and TA very high.

Next we'll look at false nodes. Those are P3 individuals that cluster together with one population in the tree but actually are ancestors of the other one. This would mean that based on their sister clade you would misclassify their origin. In similar fashion as with the unknown nodes the plot shows the proportion of false nodes for the different simulation parameters.



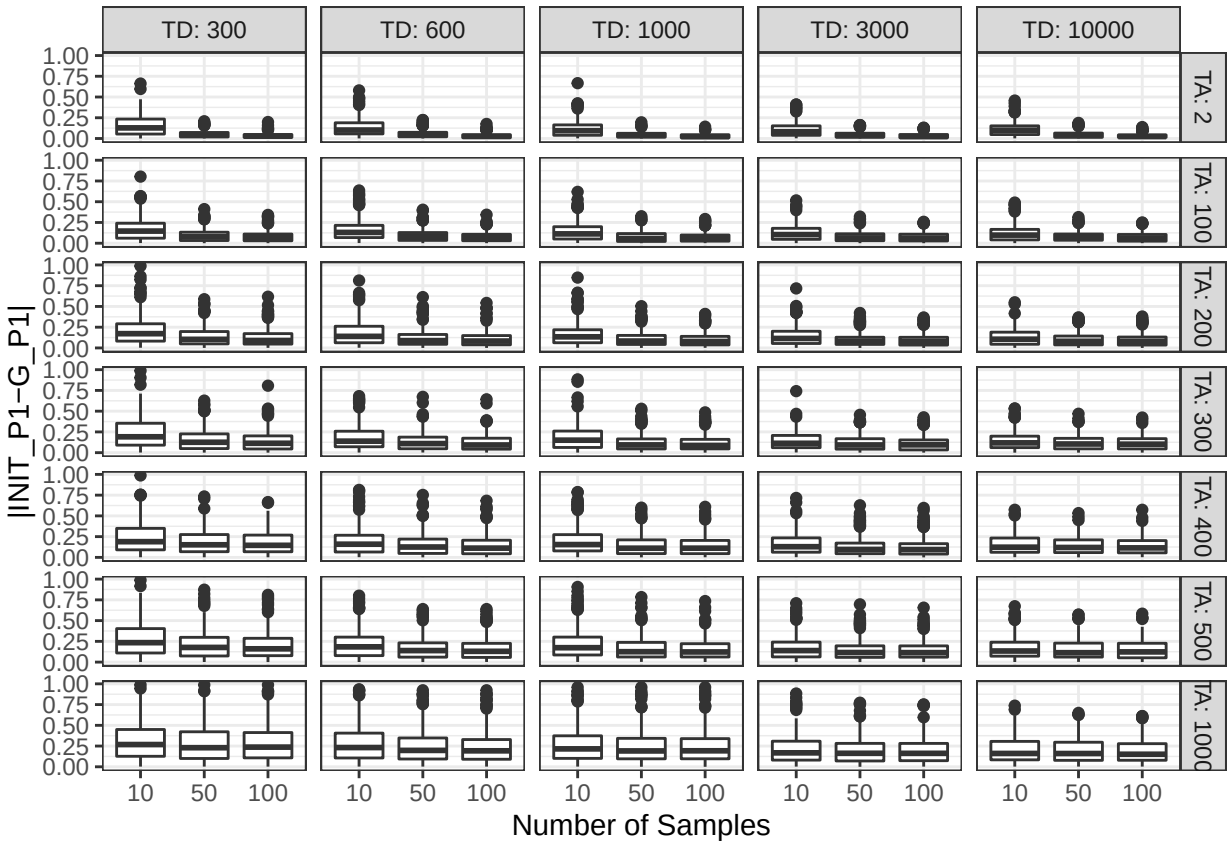
We see the same pattern as with unknown nodes. This already suggest that for a tree to represent the history of individuals accurately you need enough samples, deeply diverged source populations and the admixture should be recent.

Next let's take a look at the difference between the "true"/tracked admixture proportion and the admixture proportion inferred from the genealogy. TRUE_P1 here means the tracked proportion of P1 ancestors in P3 at the final generation of a simulation. G_P1 is the P1 proportion inferred based on the genealogy. The Plot is structured as the previous ones.



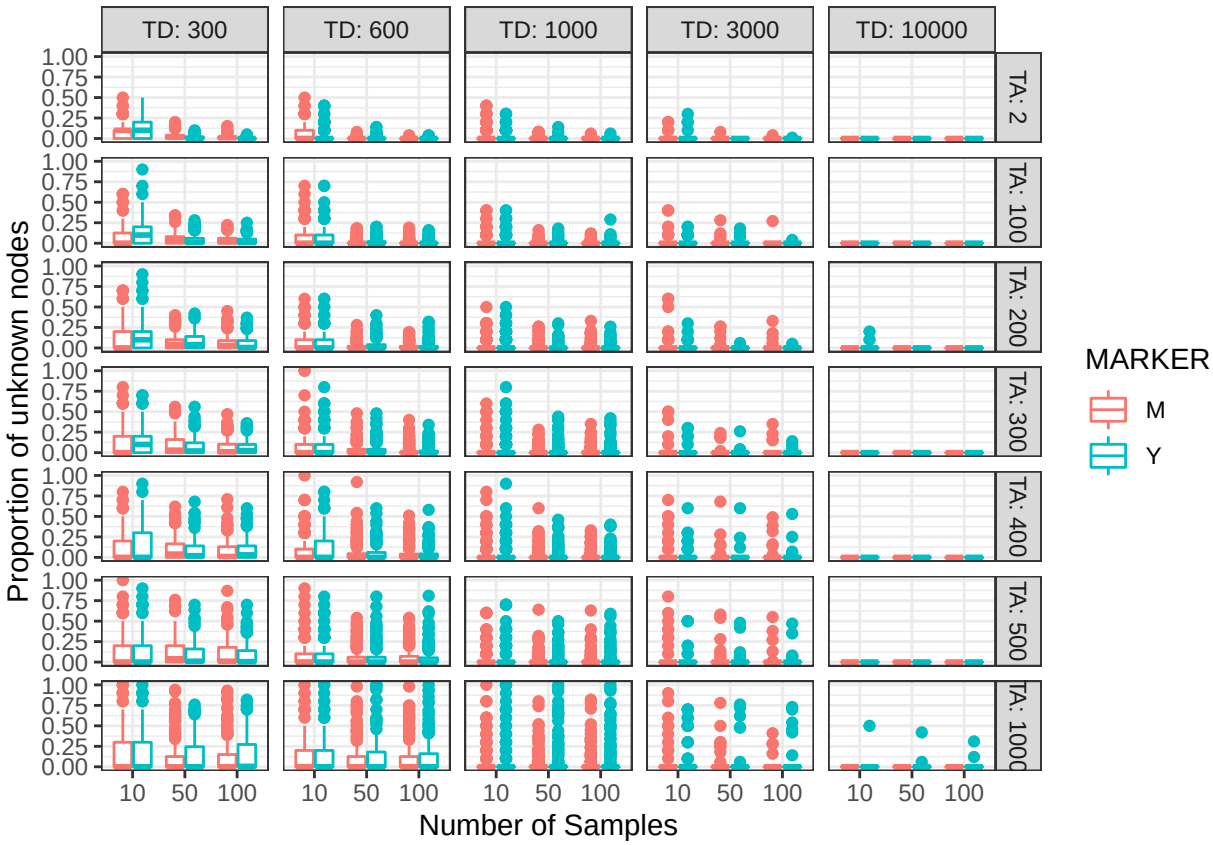
Again there is the same pattern showing that for accurate inference high TD and number of samples as well as low TA is desirable. Different here is that overall the difference between "true"/simulated proportions and inferred proportions is larger then the proportions of unknown and false nodes. This suggests that the error in inference based on trees is not only due to the methodology of relying on the clustering of nodes in a tree. I will investigate this further in later sections of this study.

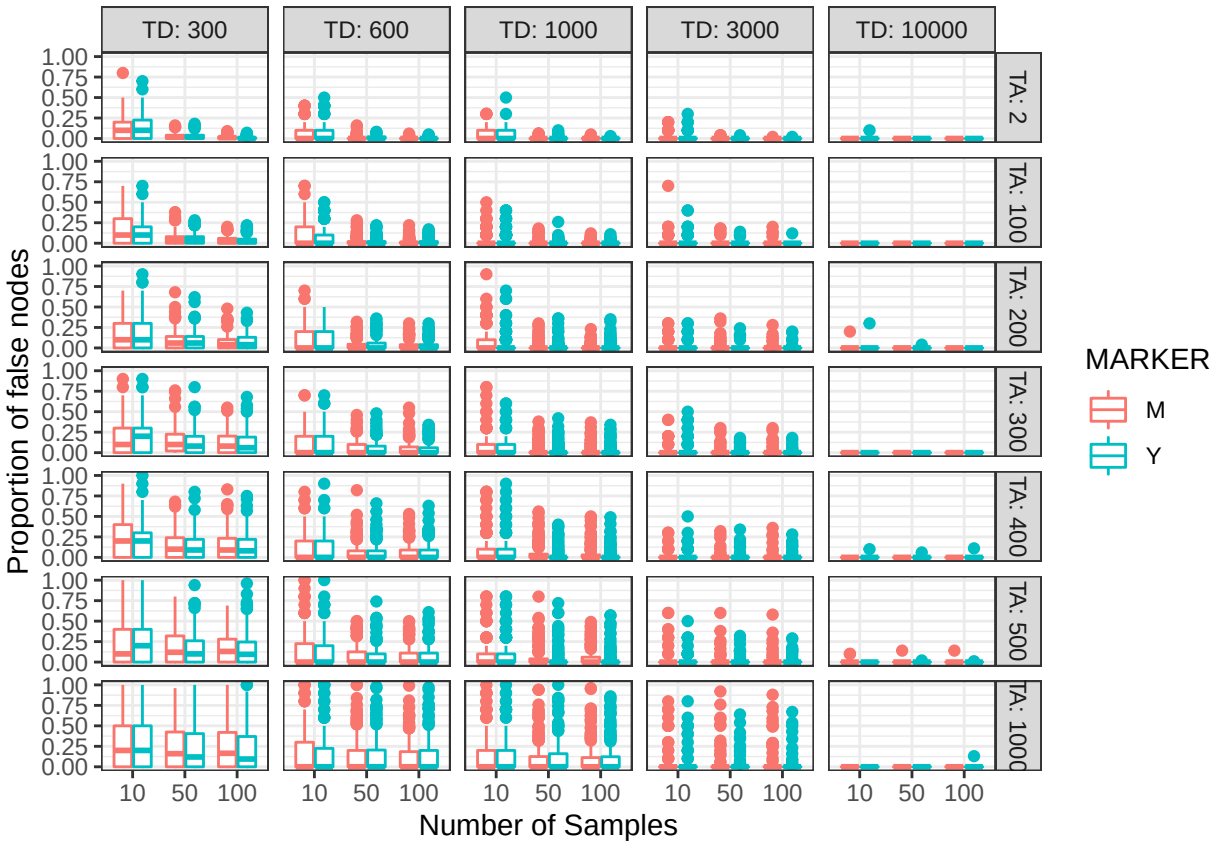
Lastly we can show the same plot as above, just now for the difference between inferred proportion and the initial proportion set up in the simulation. INIT_P1 here denotes the initial P1 proportion in P3, right at admixture.

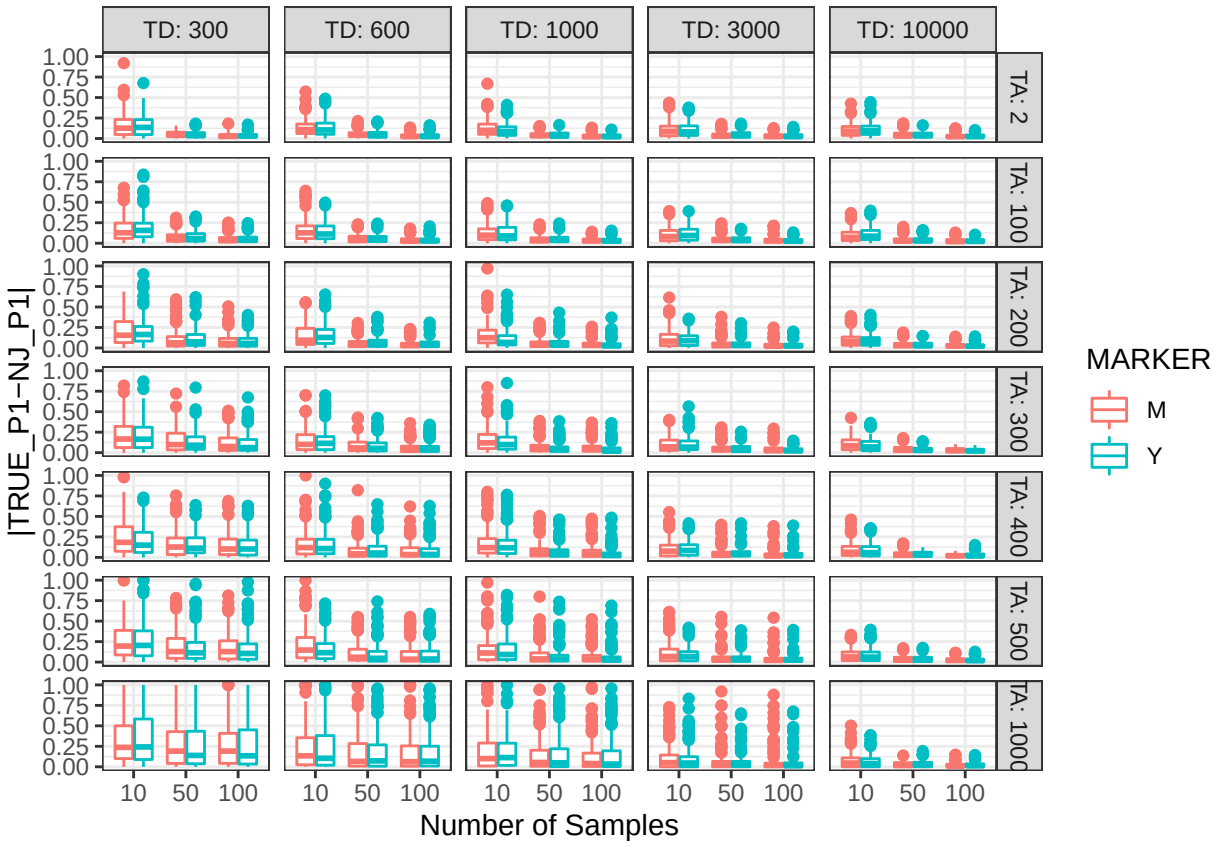


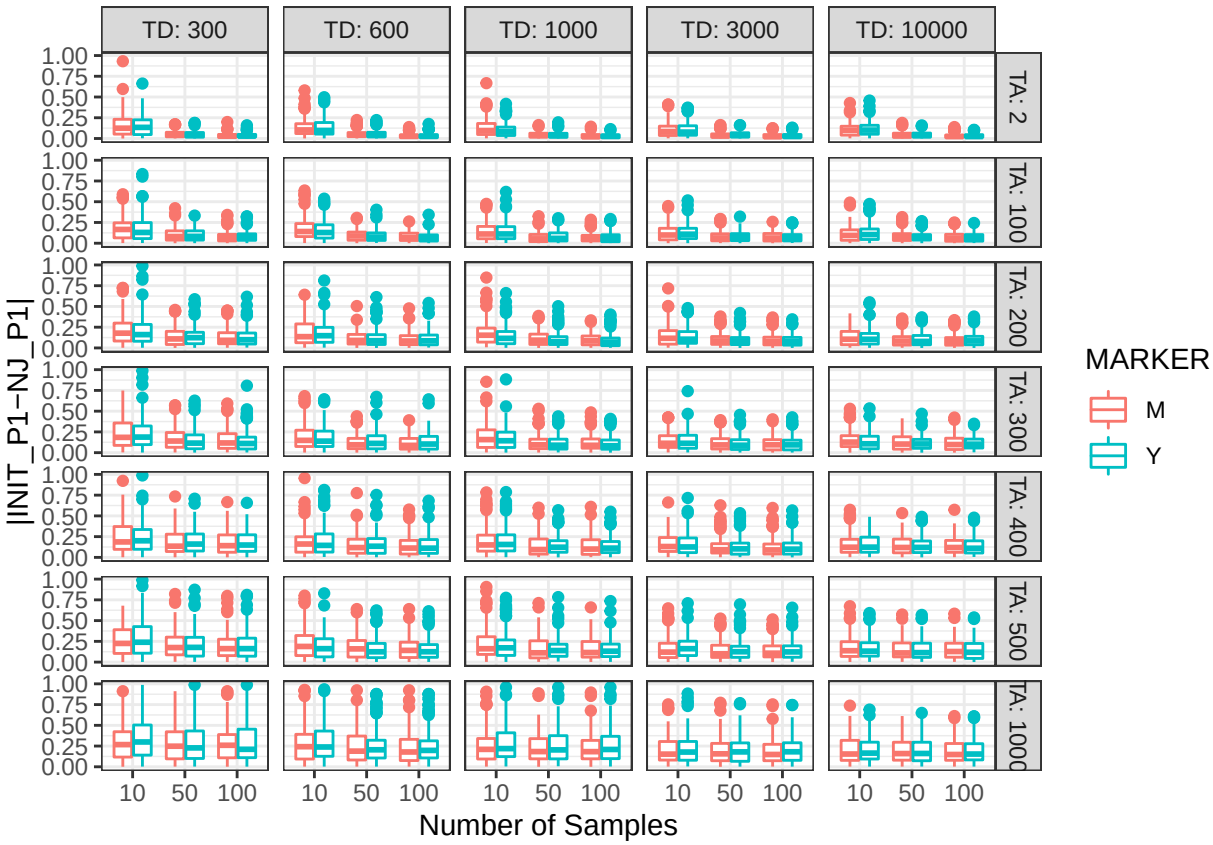
We can see that now TA has an even stronger effect because high TA means more time for the proportions to drift away from the initial proportions in the population.

Now I will provide the same plots but for the metrics of the constructed genetic tree. The genetic tree was constructed from the mtDNA and Y-chromosome sequences of the sampled individuals with the neighbor-joining method. Therefore the metrics from the genetic trees are denoted with NJ, e.g. NJ_P1. Also now since we used genetic data we need to distinguish between phylogenetic trees constructed for female individuals from mtDNA data (denoted as M and colored in red) and trees constructed for male individuals from Y-chromosome data (denoted as Y and colored in blue).


TD: 300
TD: 600
TD: 1000
TD: 3000
TD: 10000
TA: 2
TA: 100
TA: 200
TA: 300
TA: 400
TA: 500
TA: 1000
Proportion of unknown nodes
1.00
0.75
0.50
0.25
0.00
10
50
100
Number of Samples
MARKER
M
Y


TD: 300
TD: 600
TD: 1000
TD: 3000
TD: 10000
TA: 2
TA: 100
TA: 200
TA: 300
TA: 400
TA: 500
TA: 1000
Proportion of false nodes
1.00
0.75
0.50
0.25
0.00
10
50
100
Number of Samples
MARKER
M
Y


TD: 300
TD: 600
TD: 1000
TD: 3000
TD: 10000
TA: 2
TA: 100
TA: 200
TA: 300
TA: 400
TA: 500
TA: 1000
|TRUE_P1−NJ_P1|
1.00
0.75
0.50
0.25
0.00
10
50
100
Number of Samples
MARKER
M
Y



In summary the data of the phylogenetic trees does not look very different from the genealogy data. This indicates that the methodology of constructing phylogenetic trees from genetic data of uniparental markers is fairly robust and delivers similar results to the "true" trees (genealogies). I am surprised to not see a noticeable difference between using mtDNA and Y-chromosome data, which seemed logical to me due to the difference in genome size and mutation rate.

Estimating accuracy of Inference

Estimating accuracy against the current population state

Estimating accuracy against the initial population state

Estimating errors

Error of the inference method

Error of using genetic data to construct trees

Error of sampling individuals

Error of genetic drift after admixture

Conclusion