AncestryProject

Leon Joshua Gensel

2022-09-09

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 overinterpretation 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

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