INVESTIGATING HUMAN ANCESTRY BY SIMULATIONS

Exploring the limitations of using uniparental markers for the detection of sex-biased admixture

Gutachter

Prof. Dr. Katja Nowick

Prof. Dr. Britta Tietjen

Author: Leon Joshua Gensel

Investigating Human Ancestry by simulations:

Exploring the limitations of using uniparental genetic markers for the detection of sex-biased admixture

Contents

1 Introduction

1.1 Human population genetics and genetic markers

A modern approach to investigating Human Ancestry and Evolution is based on the idea that evolutionary processes leave their trace in the genomes of every individual (1, p1). Demographic changes or environmental challenges can alter the genetic diversity of a population over time (1, p2). By investigating different genetic markers with population genetic approaches, one can get insights into those past events that shaped our genomes.

Autosomal chromosomes (22 pairs) make up a majority of our genome (which is approximately 3 billion bp long) (1, p20). Each parent contributes one of their DNA molecules to the chromosome pair of their offspring (1, p29; Figure 1). During meiosis crossing over events lead to the exchange of segments between the two chromosomal homologs (1, p34). This genetic shuffling process, called recombination, and the rate at which it occurs is an important property to consider when analyzing autosomal data. Also, autosomal data, depending on the location in the genome, can be from coding or non-coding regions, being important in the regards of selective pressures (1, p27).

In addition to the 44 autosomes the human genome contains a pair of sex chromosomes which differ between the sexes (~155 Mb long X-chromosome, and ~60 Mb long Y-chromosome; Figure 1)(1 p30). Females have two copies of the X-chromosome (XX) and males one X- and one Y-chromosome (XY) (1 p30). The male passing on either their X- or Y-chromosome determines the sex of the offspring. The recombination of the X chromosome occurs only within the female individuals and it is absent in males, except for the very small parts of the pseudo-autosomal region that recombines between the X and the Y chromosomes (1 p36).

The Y-chromosome, passed on from father to sons, consists of a pseudo-autosomal region (PAR), making up approximately 10%, and the male-specific region (MSY) making up about 90% (Figure1; 1 p37). The PAR is homolog to and recombines with the X-chromosome, but the MSY does not recombine (1 p37). This makes the MSY a uniparental marker. The lack of recombination and the paternal inheritance make the MSY a useful marker to study the paternal history of populations (1 p36, 2).

Lastly, the mitochondrial DNA (mtDNA) is another uniparental marker of different molecular origin (Figure 1). The human mtDNA is a 16569 base-pair-long circular DNA molecule found in mitochondria (3). A human cell contains around 100 mitochondria, and each mitochondrion contains up to 15 mtDNA molecules making it abundant for sequencing (4). Additionally, the high mutation rate of the mtDNA results in a high number of polymorphisms, which makes diversity easy to detect (1 p603, 5). The mtDNA is inherited uniparentally from mothers to her offspring and does not recombine (1 p37,8). Even though mtDNA is present in both sexes the lack of recombination and strictly maternal inheritance make it analogous to the MSY.

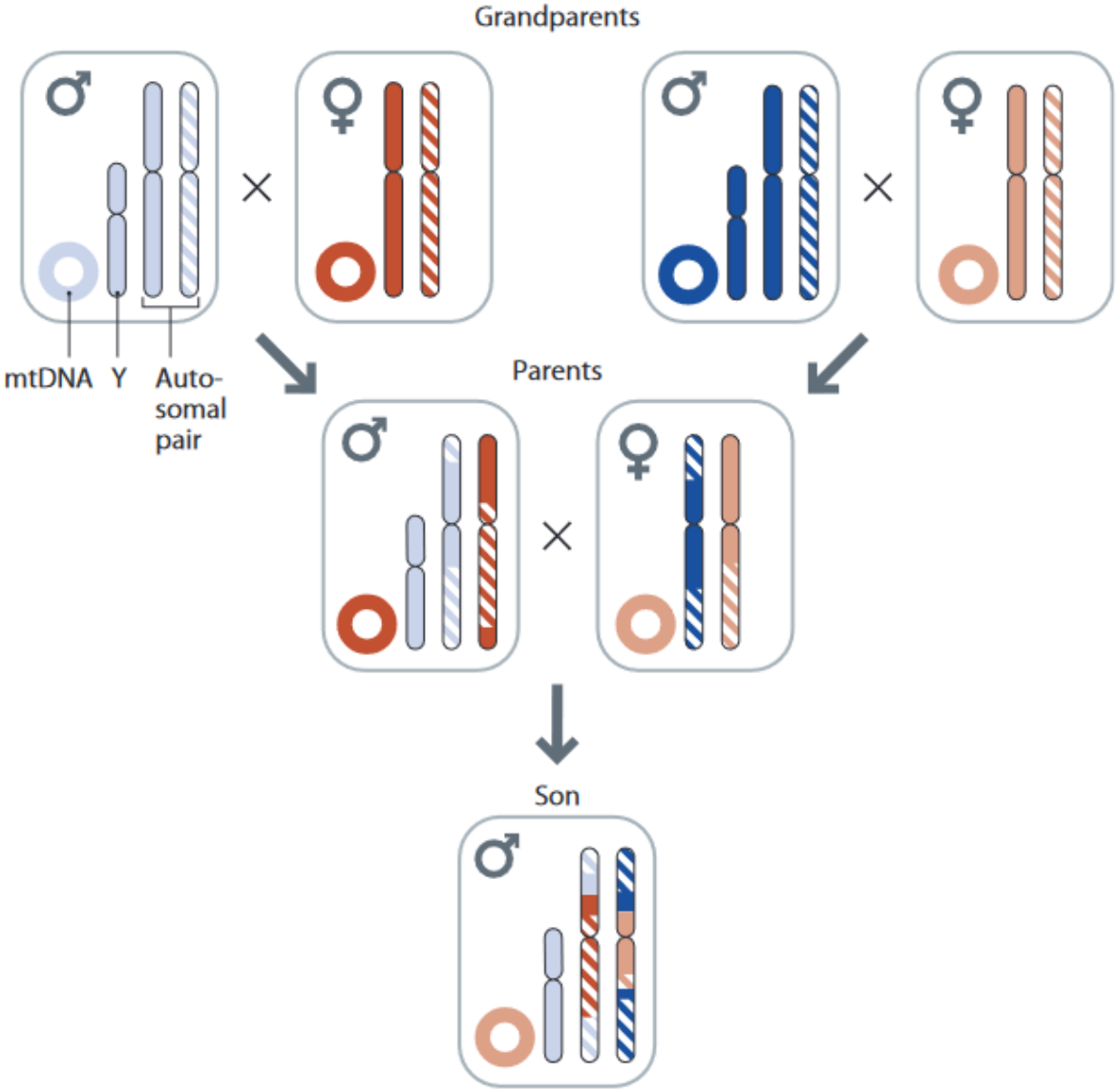

Figure 1. *Example of the inheritance of different genetic markers. Shown is the strictly maternal inheritance of the mtDNA, the strictly paternal inheritance of the MSY and the biparental inheritance of a pair of autosomes with recombination events in every generation (Source: 1 p36).*

1.2 Phylogeny and Haplogroups of the mtDNA and the MSY

A method to visualize the genetic relationships between different entities (for example individuals or populations) is to construct a phylogenetic tree (1 p182). Such a tree describes the genetic history and all the ancestral relationships for every entity (1 p182). Terminology for these gene trees partially varies from the tree-terminology in computer science and is depicted in Figure 2.

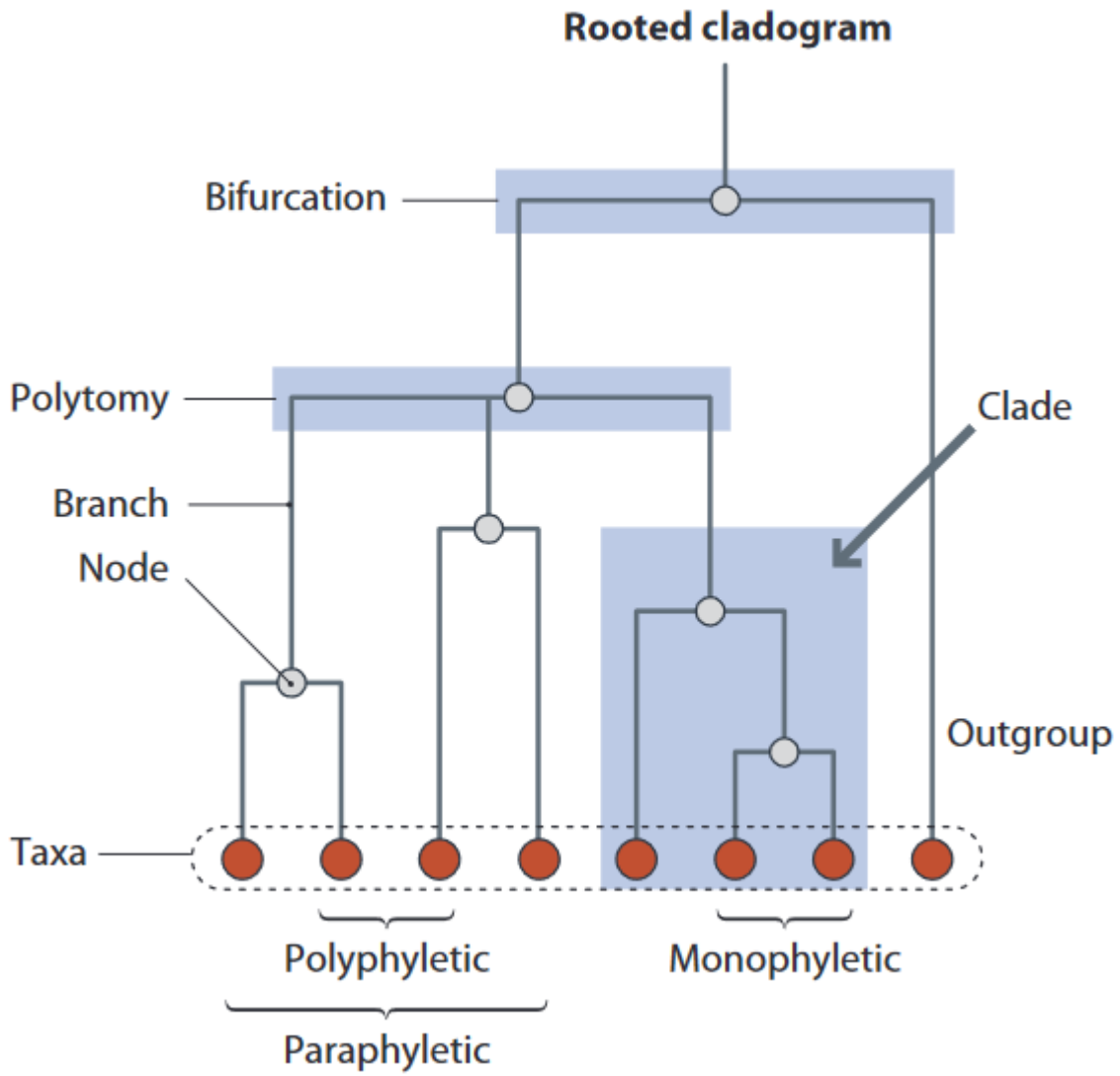

Figure 2. *Terminology for phylogenetic trees. The terminology is partially different from the conventional terms used for trees in computer science. Nodes that are no taxa represent ancestral entities, usually genomes. (Source: 1 p183)*

It is important that these gene trees do not necessarily have to reflect the "true" family tree (genealogy) of its entities, because it is constructed from statistical approaches based on genetic data (1 p185). There are several methods for constructing these trees, for example genetic distance matrices or maximum parsimony, which all have different use cases (1 p184). Though these will not be discussed further.

The lack of recombination in the Y-chromosome and the mtDNA allows reliable reconstruction of phylogenetic trees for both markers (1 p182). The global phylogenies have been reconstructed for both uniparental markers (1 p604-607). These trees are great for visualizing the relationships of human populations (1 p182). For every entity of the tree all ancestral genomes that contributed to its state are represented. Every ancestral node in the tree can also be assigned to a polymorphism event, that distinguishes it and all its descendants genetically from the rest of the tree (1 p601). The sets of polymorphisms on a single DNA molecule that tend to be passed on together are called haplotypes, derived from "haploid genotype" (6). Similar haplotypes that share a common ancestor are called a haplogroup and can be defined for both, the Y-chromosome and the mtDNA (6). The haplogroup definition can be done at different resolutions (i.e. at different points in time) (Figure 3).

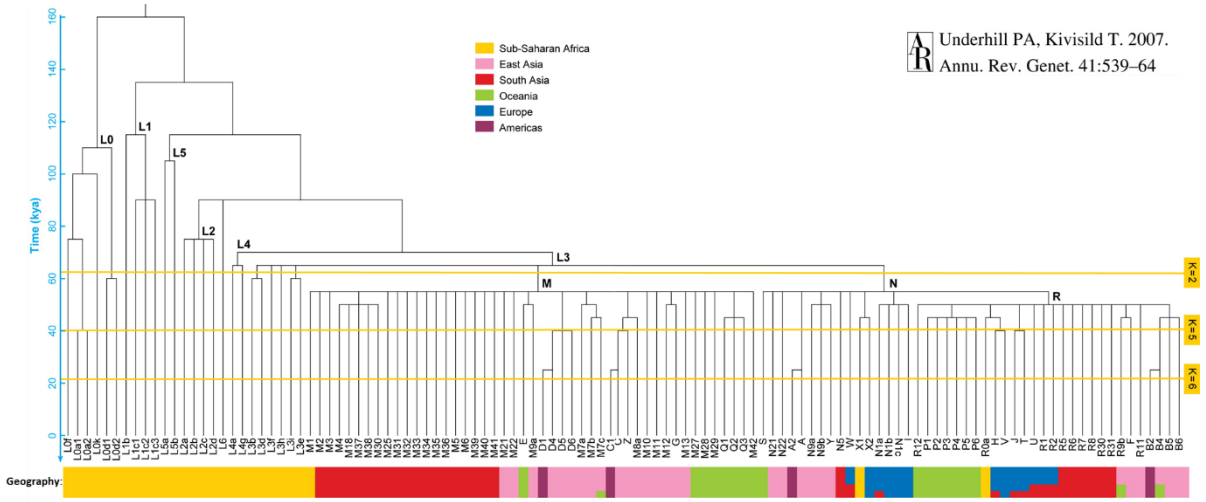

Figure 3. *The global mtDNA phylogeny and different timeframes of haplogroup definition. The color bars at the bottom reflect the geographical distribution of the corresponding haplogroup. The yellow horizontal lines reflect points in time at which continental differences arise in the human populations. Different K values describe the number of distinguishable geographic regions in which different haplogroups occur. For example, at K = 2 the differentiation between African and the non-African haplogroups can be seen, while at K = 6 the differentiation of haplogroups in all six major geographic regions inhabited by humans can be seen, including the haplogroups specific to America, which were inhabited the last since the anatomically modern humans migrated out-of-Africa. (Source 2)*

There is a distinct nomenclature for both mtDNA (7) and Y-chromosomal (8) haplogroups. The nomenclature is based on the cladistic structure of the phylogeny (1 p601). The major clades are usually named with capital letters (Example: R), and all subsequent lineages (sub-haplogroups or minor-haplogroups) in addition to the capital letter have a sequence of alternating numbers and lowercase letters (Example: R1b1b2) (1 p601). The nomenclature of the haplogroups (i.e. the length of the name) is not necessarily reflecting the time of occurrence of the corresponding haplogroup. Many haplogroups can have specific geographical distributions, or they can be specific to certain populations. Thus, analyses based on haplogroup compositions are very useful when comparing different populations (1 p604-607; Figure 3).

1.3 Sex-biased admixture

Admixture describes the process of populations meeting and exchanging migrants or merging into one population (1 p443). How much each population contributed to the admixed population is defined as the admixture proportion (1 p450). Depending on the circumstances it might be possible to quantitatively estimate how much each population contributed to the genetic diversity of the admixed population, but this signal might be altered by genetic drift, selection, mutation, or other gene flow (1 p443). There are several methods to detect admixture and estimate admixture proportions, which can use different types of genetic markers (1 p453). Estimates of admixture based on different markers can lead to different estimations for the same population (for example Colombian population) (9).

When sex-specific markers (e.g. mtDNA and MSY) give very different estimates this may indicate a sex-biased admixture (Figure 4; 1 p464). This can be due to different admixing proportions for the separate sexes from the ancestral populations, or due to directional mating patterns between admixing populations (1 p464). Sex-biased admixture and sex-specific migrations are not unusual in human history, especially in the more recent one (e.g. 10, 11, 12). One of the most known historic events that led to sex-biased admixture is the European colonization which resulted in male-biased migration and admixture between European males and different indigenous groups across the world (1 p464). Detecting such sex-biased admixture events is possible when comparing the genetic markers with different inheritance patterns (1 p465).

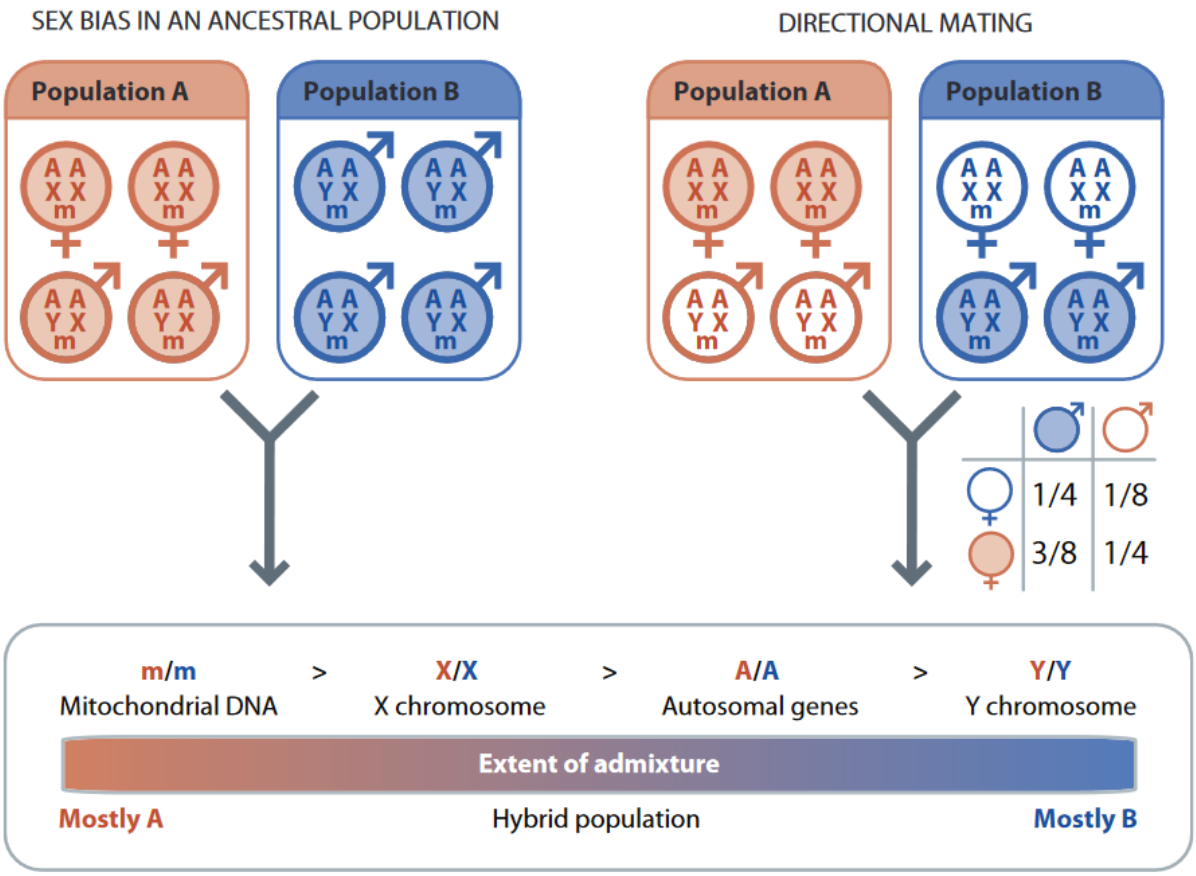

Figure 4. *Example of sex-biased admixture processes and the resulting admixture estimates for different genetic markers. Shown are two scenarios of admixture that would result in similar admixture estimates for different genetic markers. In the first scenario population B contributes only males (a so-called sex-specific admixture), and in the second scenario the different sexes in both populations mate in differing frequencies. Using markers with different inheritance patterns results in different estimates in regards of admixture proportions. Comparing the estimates of different markers enables inferences of sex-biased gene flow. (Source: 1 p464)*

The Figure 4 depicts two different admixture scenarios, which would result in the same admixture estimates. In both scenarios, population B contributes more males to the admixed population. The mtDNA being inherited strictly maternally would give an estimation heavily skewed towards population A. The exclusively paternal inherited Y-chromosome, in contrast, would estimate a higher contribution of population B. Since in both scenarios the number of mating individuals from populations A and B is the same, the autosomal estimations are expected to be in the middle between population A and B. The X-chromosome with its 2-to-1 female-to-male biased inheritance pattern is skewed towards population A which contributes the majority of the mtDNA in the admixed population.

To quantitatively estimate sex-biased admixture one can resort to different methods, which all compare admixture proportions inferred from two or more genetic markers with different inheritance patterns (1 p466).

1.4 Simulations in population genetics

Statistical methods in human population genetics usually use genomic data to reconstruct the genetic history of a given population. Often the quantitative aspects of such methods are based on estimations of demographic parameters, such as effective population size or migration rates. When checking the accuracy of a statistical methodology simulations can be a very useful, and often a necessary step. In a simulator one can precisely set all the demographic parameters, and quantify their influence on the genetic data. This helps to validate and assess analytical methods (13).

There are two commonly used types of simulations in human population genetics: forward simulations, and coalescent simulations. Coalescent simulations take the genomic data of a present population as input and run backwards in time (i.e. they simulate the past that generated the given data). The main advantage of coalescent simulation is that it runs fast and with great computational efficiency (14). On the other hand, they break down for large DNA regions, because the resulting recombination patterns become too complex (14). In contrast, forward simulations start at a defined point and from there simulate forward in time. This allows modeling large portions of the genome, as well as complicated demographic scenarios, at the cost of a longer runtime (14). The extra computational demand is due to the fact that forward simulators have to simulate genomes that will not necessarily contribute to the individuals of the final generation.

1.5 Aims

The frequencies of uniparental markers have been commonly used to detect sex-biased admixture (e.g. 15). Yet, there is no clear guidance on what the limitations of such approach are, and how to avoid over-interpretation of such signal. The main aim of this study is to test the limitations of the estimates of sex-biased gene flow based on the differences in frequencies of mtDNA- and MSY on the population level. To do so, I simulated the mtDNA and the MSY using the forward-simulator SLiM (16) with predefined parameters for the time of divergence between admixing populations, time of admixture, and intensity of sex-biased admixture. After defining haplogroup frequencies per population, and comparing observed and expected frequencies, I was able to determine the circumstances under which calculations of sex-biased gene flow using the frequencies of uniparental markers could be considered trustworthy. This study provides recommendations on how to analyze and interpret sex-biased admixture events in the human history using uniparental markers. Here I show that inferred sex-biases can only be considered accurate for recent admixture events. Parameters such as the degree of divergence between the source populations, and the time at which haplogroups are defined, are of less importance.

2 Materials and methods

2.1 Simulation setup

To test the limitations of uniparental markers in detecting sex-biases I simulated the mtDNA and Y-chromosome (MSY) across three populations using the forward simulator SLiM version 3.5 (16). To enable the desired crossing between individuals of different populations I used a non-Wright-Fischer model, but I implemented it to resemble a Wright-Fisher model, similarly to the recipe 16.15 from the SLiM manual (17). I defined fixed population sizes and non-overlapping generations in the model. This is achieved by removing all individuals with an age greater than zero at the start of every generation (See early()-callback in Code snippet 1). Fixed population sizes are implemented in the reproduction()-callback, where for the desired population size parents within each population are drawn randomly to generate one offspring individual with a randomized sex (Code snippet 1). The addRecombinant()-command ensures that every offspring in the simulation inherited the mtDNA in genome 1 from its mother and the Y-chromosome in genome 2 from its father regardless of sex. Both, the mtDNA and the Y-chromosome, were tracked with specific marker mutations.

```
reproduction() {
	for (s in sim.subpopulations){
		for (i in seqLen(Ne)){
			mom = s.sampleIndividuals(1, sex = "F");
			dad = s.sampleIndividuals(1, sex = "M");
			s.addRecombinant(mom.genome1, NULL, NULL, dad.genome2, NULL, NULL);
		}
	}
	self.active = 0;
}

early() {
	sim.subpopulations.individuals[sim.subpopulations.individuals.age > 0].fitnessScaling = 0;
}
```

Code snippet 1. *SLiM code for ensuring the correct reproduction pattern and non-overlapping generations in the simulations.*

To ensure that there is a most common-recent ancestor (MCRA) for every individual at the end of the simulation the basic model (Figure 5) simulated a 100 000 generation long burn-in period for an initial ancestral population with a fixed population size of N = 5000 (27).

After the burn-in a population split of the initial ancestral population into two populations of equal size N = 5000 was set up (Figure 5: divergence). After the population split different scenarios were set up. The setups differed in the time between the splitting event and the subsequent admixture event (Tdiff), which was scheduled either 100, 300 or 3000 generations after the split (Figure 5: admixture). These scenarios allow me to investigate how the divergence time between admixing populations is affecting the power of detecting the sex-biased admixture between them.

Then, for each of those three different scenarios, I simulated two different strengths of sex-biased admixture between population 1 (P1) and population 2 (P2) measured as proportion of P1 males in P3 (p1MaleRatio). For the no-sex-biased scenario I set a ratio of 50:50 (i.e. 50% of P3 males are migrants from P1 and 50% from P2). For the strong sex-bias scenario I set a ratio of 80:20 (i.e. 80% of P3 males are migrants from P1 and 20% from P2). Just like the other two populations, the admixed

third population was set to have the same population size of N = 5000. Then for each of those six different simulation scenarios (3 scenarios for Tdiff x 2 scenarios for sex-bias intensity) I recorded the output on generation 10, 100, 200, 300, 400 and 500 after the admixture event (Tadmx). Each of the six different simulation setups was simulated for 100 times with random seeds, resulting in 3600 output files (containing both mtDNA and MSY).

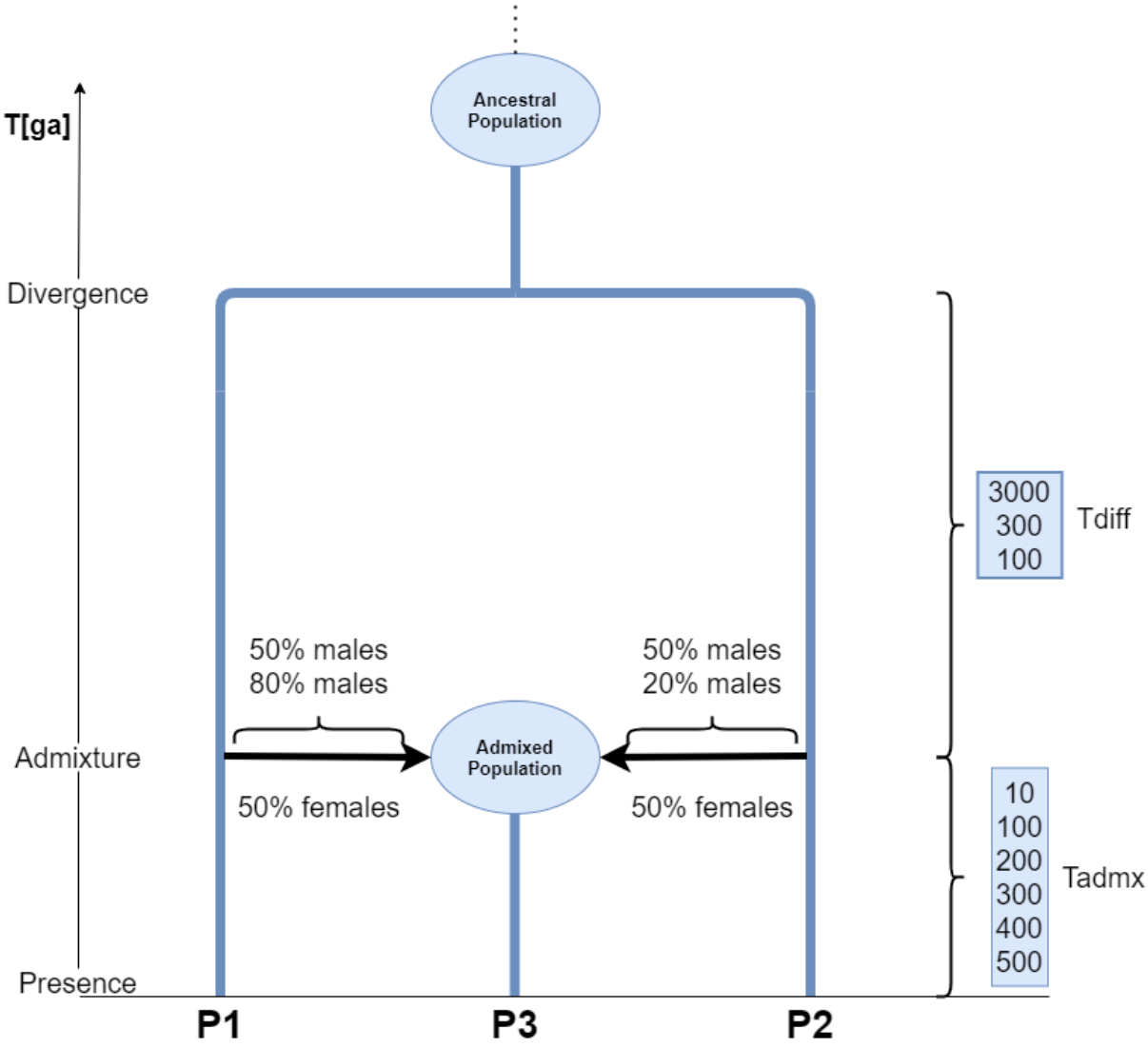

Figure 5. *SLiM simulation setup. The simulation starts with a 100000 generation burn-in period of an ancestral population. At the time point of divergence the ancestral population splits into the populations P1 and P2. Tdiff describes the time between divergence and admixture and varies between scenarios (100, 300, and 3000). Tadmx defines at how many generations after admixture output was generated. Admixture events with no sex-bias (50% of P3 males and 50% of P3 females come from P1) or strong sex-bias (80% of P3 males and 50% of P3 females come from P1) were simulated for each combination of Tdiff and Tadmx.*

Altogether, this set of simulations allowed me to monitor the impact of the time of divergence and the time of admixture on the detection of sex-biased admixture using comparisons of uniparental markers.

2.2 Tree-sequence recording

I implemented SLiM's build in tree-sequence recording function (18), to record the mating patterns of the simulated populations and later reconstruct the genetic trees for the mtDNA and the MSY across all populations. The tree-sequence recording also reduced the computational time needed for

simulations by removing a need for simulating neutral mutations. The resulting output trees were used to investigate sex-biased processes.

The SLiM tree sequences consist of trees that record the ancestry at every position of a genome. Simulating the non-recombining mtDNA and MSY results in one tree for each of the markers. After simplification, the trees only contain the nodes of ancestral genomes that are necessary to reconstruct the genetic history for the genomes of individuals in the last generation (i.e. present individuals). This means that every pair of present individuals will have their MRCA preserved in the trees, as well as the MRCA of the entire simulation.

2.3 Tree analysis

The analysis of the trees was performed in python (19). For the processing of the trees, I used the *tskit* (20) and *pyslim* (21) API. Other packages used were *pandas* (25) and *numpy* (26). By simplifying the tree sequence to contain only genomes of female individuals, which contain the mtDNA marker mutation specified in SLiM script and only genomes of male individuals, which contain the Y-chromosome marker mutation specified in SLiM script, it is possible to define separate genealogical trees for the mtDNA and the MSY (Code snippet 2).

```
#Importing and simplifying tree sequence

    treeseq_raw = pyslim.load(input)
    all_nodes = treeseq_raw.samples()
    keep_nodes_F = []
    keep_nodes_M = []

    for u in all_nodes:
        if treeseq_raw.mutation_at(u,0) != -
1 and treeseq_raw.individual(treeseq_raw.node(u).individual).metadata["sex"] =
= pyslim.INDIVIDUAL_TYPE_FEMALE:
            keep_nodes_F.append(u)
        if treeseq_raw.mutation_at(u,100) != -
1 and treeseq_raw.individual(treeseq_raw.node(u).individual).metadata["sex"] =
= pyslim.INDIVIDUAL_TYPE_MALE:
            keep_nodes_M.append(u)
        else:
            pass

    treeseq_mtDNA = treeseq_raw.simplify(keep_nodes_F)
    treeseq_YChrom = treeseq_raw.simplify(keep_nodes_M)

    if(treeseq_mtDNA.num_trees != 1): raise ValueError("more than one tree!")
    if(treeseq_YChrom.num_trees != 1): raise ValueError("more than one tree!")

    tree_mtDNA = treeseq_mtDNA.first()
    tree_YChrom = treeseq_YChrom.first()
```

Code snippet 2. *Python script that performs iteration through all genomes of the tree sequence and discards non-relevant genomes and separates mtDNA and Y-chromosome genealogies.*

On the resulting trees I implemented an algorithm that defines haplogroups on the phylogenetic level similar to how it is done in (2). A cutoff time at which to define haplogroups was determined based on a preliminary runs for each scenario. I choose a time point that minimizes the number of monophyletic haplogroups specific to population 3 and the number of uninformative haplogroups that occur in all three populations. The nodes of the trees of those preliminary runs were grouped based on the populations to which their leaves belonged (i.e. the node was assigned a category "13" if all leaves of the node were either belonging to population 1 or 3). Then I plotted the proportions of leaves across different categories of nodes (Figure 7) using python script. Based on close inspection of such plots for all scenarios I choose the cutoff time to be the average of Tdiff and Tadmx.

Since the *tskit* trees reflect the genealogy of the simulations without any information about mutation events, it is important to ensure that haplogroups can be genetically distinguished before being defined. Instead of overlaying mutations with a coalescent simulator onto the genealogy trees I probabilistically tested if a branch at the defined time point would have at least one mutation that would make the haplotype of this branch genetically distinguishable from the others. The probability for a branch being mutated was calculated using following formula:

$$P(M) = 1 - (1 - \mu)^{L \times t}$$

The probability of a branch having at least one or more mutations is the complementary probability of a branch having no mutation, so *P(M) = 1 - P(not M).* The probability of no mutation in a branch can be calculated as the complementary probability of a mutation happening, which is the mutation rate (μ) per site per generation to the power of the length of the genome (L) times the time in generations (t) represented by the branch length in a genealogical tree. For the 16500 bp long mtDNA a mutation rate estimate of 6.85×10^{-7} was used (22). For the Y-chromosome, a mutation rate estimate of 3.01×10^{-8} (23) assuming a 900000 bp long sequence was used (24).

The implementation of these probabilistic tests into the haplogroup definition algorithm avoids unnecessary simulation of mutations or overlaying them onto the genealogical tree with an additional coalescent simulator like *msprime* (28). In addition, using the previously described approach, mutations are only generated for the branches, which are relevant for haplogroup definition. This considerably decreases the computational load and runtime of the python script for tree analysis.

The python script for haplogroup definition is shown in Code snippet 3.

```
    #Define Haplogroups

    def mutateEdge(tree: tskit.Tree, node, dataSource: str):
        edgeMutationDict = dict()
        if dataSource == "YChrom":
            L=900000
            m=3.01*10**-8
        elif dataSource == "mtDNA":
            L=16500
            m=6.85*10**-7
        else:
            raise Exception("invalid dataSource argument!")
        return 1-((1-
m)**(L*tree.branch_length(node))) > np.random.uniform(0,1)


    def defineHaplogroup(tree: tskit.Tree, dataSource, time):
        haplogroups = {}
        for u in tree.nodes(root=tree.root, order="timedesc"):
            if u == tree.root or tree.time(tree.parent(u)) <= time:
                pass
            else:
                if mutateEdge(tree, u, dataSource):
                    for i in haplogroups.values():
                        for v in tree.leaves(u):
                            i.discard(v)
                    haplogroups[f"H_{u}"] = set([u for u in tree.leaves(u)])
        haplogroups = {key:val for key, val in haplogroups.items() if len(val)
 != 0}
        return haplogroups
```

Code snippet 3. *Python functions that test whether a branch contains mutations or not and define haplogroups for a specific time point. The defineHaplogroup()-function defines a dictionary with the haplogroups as keys, and the corresponding leaves as values.*

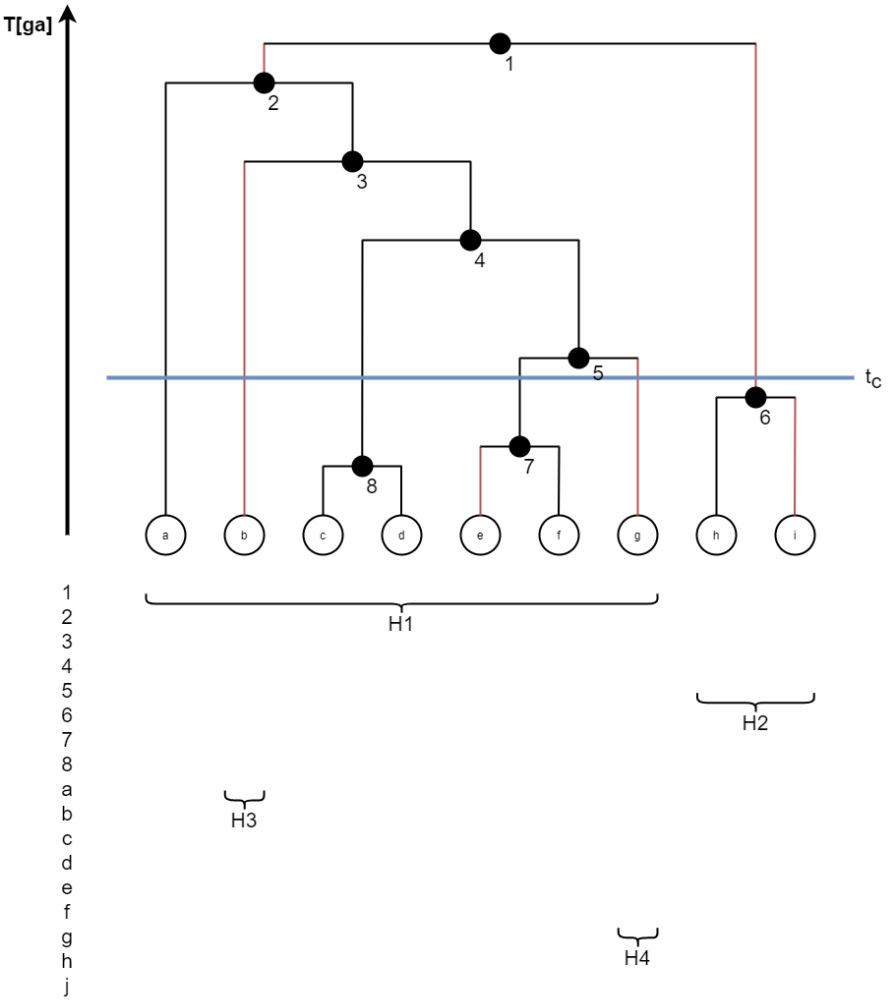

Figure 6. *Explanatory diagram for the haplogroup definition algorithm (from Code snippet 3) on a simple example tree from tskit. The tree is scaled to the time in generations ago (T[ga]). The leaves (labeled from "a" to "i") represent the genomes of living individuals at the end of the simulation. The ancestral nodes are numbered from 1 to 8. The horizontal blue line shows the time at which the haplogroup should be defined (t_c). Branches for which the mutateEdge-function returned True and therefore represent branches that would be genetically distinguishable are shown in red. The steps of iteration at which haplogroups are defined are shown below the tree (from 1 to j).*

In the following subsection I will illustrate how the haplogroup definition algorithm operates on a simple example tree depicted in Figure 7. The algorithm iterates over all nodes of the tree in descending time starting with the oldest (node 1) and ending with the youngest one (last leaf, in this case "i"). When nodes are at the same time, they will be ordered based on their ID defined by SLiM. A node will be disregarded if it is the root of the tree, and if its parent node is below the cutoff time for haplogroup definition (t_c), since in both of those cases no haplogroups should be defined from these nodes. For each node, the mutateEdge-function returns true or false based on random number generation and the probability of mutations based on mutation rate and branch length. If it returns true, the branch leading to this node is considered as genetically distinguishable and can be used to define a haplogroup. Then a haplogroup is defined as key in a dictionary data structure with the value being a set of leaves of the subtree of this node (i.e. these leaves get assigned to the new haplogroup). To avoid assigning leaves into multiple haplogroups the leaves added into a more recent haplogroup will be removed from all older haplogroups.

In the example tree mutated branches are preemptively marked in red to simplify the explanation. Usually this would be tested during iteration with the mutateEdge-function. The iteration would start

with the node 1, which would be discarded due to being the root of the tree. The node 2 would define the first haplogroup H1 which would contain the leaves “a”, “b”, “c”, “d”, “e”, “f”, “g”. Nodes 3, 4 and 5 would be discarded due to their branches being not genetically distinguishable from the original haplotype. Node 6 would define a new haplogroup H2 with leaves “h” and “i". Node “a” would be discarded due to a lack of mutation in the branch leading to it. Node “b” would define a new haplogroup H3, and it would remove leaf “b” from the previously defined H1 haplogroup. Nodes “c”, “d”, “e” and “f” would be discarded due to their parent nodes being below t_c. Node “g” would define haplogroup H4, and it would be removed from H1. Nodes “h” and “i” would also be skipped due to their parents being below the cutoff time. This would result in a dictionary like this:

H1	a, c, d, e, f
H2	h, i
H3	b
H4	g

Since there is the possibility that all leaves in an early defined haplogroup get removed later on (resulting in an empty haplogroup), all such haplogroups are removed at the end of the iteration. The simulated output trees from this study contain 15000 leaves, and reach up to approximately 8000 generations of depth. The definition of haplogroups containing only one leaf does not occur in our data due to a higher t_c. For small trees (like the example tree) it is possible to have leaves that are not assigned to any haplogroup due to a lack of mutation events. This is noticeable in the haplogroup frequencies, and it does not occur on larger trees.

After assigning every individual to a mtDNA and MSY haplogroup, the frequencies of the haplogroups were calculated for every population. The frequencies of both uniparental markers were further used to explore the circumstances under which sex-biased gene flow can be detected.

2.4 Summary statistics

The haplogroup frequency tables of every simulation run were then used as an input for the weighted correspondence analysis (CA) performed using the *CA* function from the “*FactoMineR*” package in R (29). The first hundred dimensions were retained by setting the parameter *ncp = 100.* To ensure that the difference in numbers of haplogroups between mtDNA and MSY would not influence the CA analysis, the weight of each variable (i.e. mtDNA and MSY) was set to be inversely proportional to the number of other variables (i.e. number of haplogroups) of the same marker type, and thus the sum of weights for each kind of marker is the same.

The output of the CA was further used as an input for hierarchical clustering on principal components (HCPC) using the *HCPC* function from the *“FactoMineR v1.42”* package (29) to define two clusters of highly correlated mtDNA and MSY haplogroups characteristic to P1 and P2.

Next, the sums of each marker belonging to one of the two clusters were calculated (i.e. all mtDNA haplogroup frequencies assigned to the first cluster were summed into one value, and all MSY haplogroup frequencies assigned to the first cluster were summed into one value, and same for the mtDNA for MSY haplogroup frequencies assigned to the second cluster). Cluster frequencies calculated this way were further used to calculate the intensity of sex-biased gene flow between the two clusters across populations (measured as difference between mtDNA haplogroup frequencies and MSY haplogroup frequencies within same cluster). The difference between mtDNA and Y-chromosome frequencies within each cluster indicate sex-biased processes (Figure 9, Figure 10). It is expected that mtDNA and MSY haplogroups from P1 are completely assigned to one of the clusters and that all mtDNA and MSY haplogroups from P2 are assigned to the other cluster. On the other side,

haplogroups from P3 are expected to be distributed between the two clusters reflecting the simulated admixture proportion.

Comparing the observed sex-biases and expected sex-biases (set up in the simulations) allowed me to describe under which circumstances the frequencies of uniparental markers accurately reflect sex-biased gene flow. This was done in R using the function *rmse()* from the Metrics package (30) which calculates the root mean square error (rmse). I also provide the percentage of runs for each simulation set up that are within a certain interval (± 1%, 5% and 10%) of the simulated sex-bias.

In addition, I calculated the overall accuracy of detecting sex-bias. For each simulation setup I counted the number of simulations that are 1) between -0.05 and 0.05 (representing simulations where the output does not support an existence of sex-bias); 2) less than -0.05 (representing simulations with output that would be expected for simulations with sex-bias, that show higher contribution of P1 males than P1 females in P3); and 3) more than 0,05 (representing simulations that show higher contribution of P1 females than P1 males in P3, which are unexpected scenarios as true simulations with sex-bias had more P1 males than females in P3). This allows accuracy calculation for the detection of sex-biased admixture based on rates for true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) with the following formula:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

This was done for every combination of the Tadmx and Tdiff parameters.

3 Results and Discussion

3.1 Defining cutoff time for haplogroup definition

To decide the time at which haplogroups should be defined I inspected the plots with the distribution of leaves for each node (Figure 7).

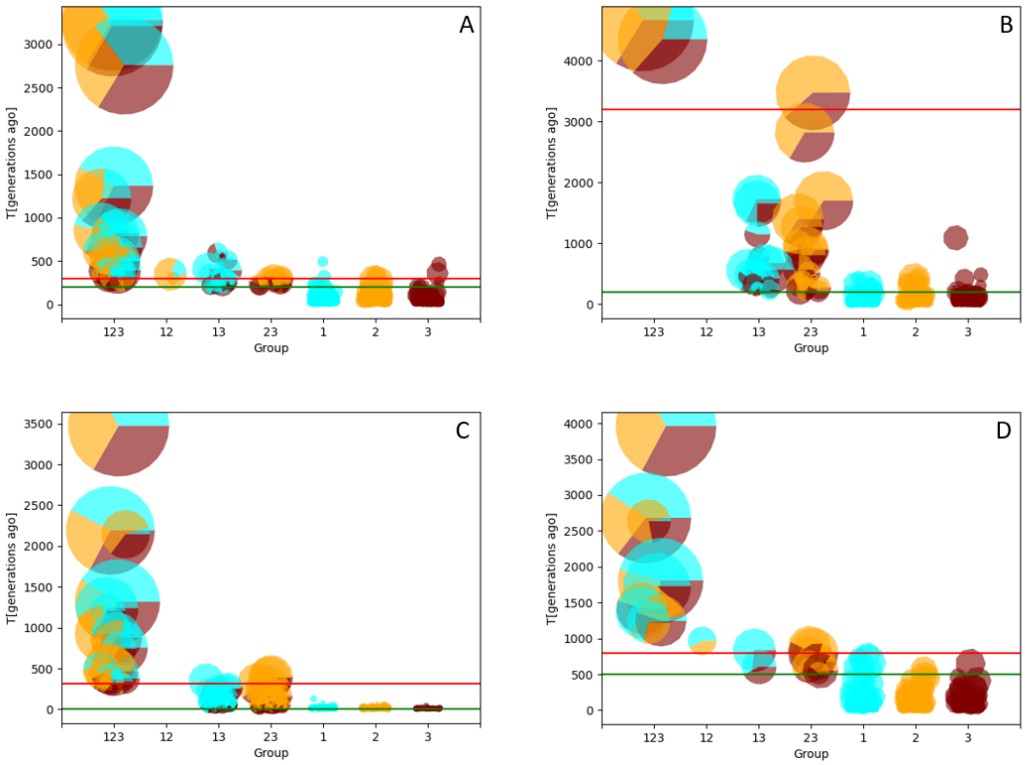

Figure	T diff	T admx
A	100	200
B	3000	200
C	300	10
D	300	500

Figure 7 *Leave distributions across different groups of nodes for scenarios with the maximum and minimum values of simulated Tdiff and Tadmx. The horizontal green line represents the time of admixture (Tadmx) and the horizontal red line represents the time of divergence between P1 and P2 (Tdiff + Tadmx). The size of the plotted pie charts scales with the number of leaves. The size of pieces with different colors corresponds to the frequencies of leaves of a given node that occurs in P1 (blue), P2 (orange), and P3 (brown). Node-groups 1,2 and 3 are monophyletic and population specific, node-groups 12, 13, 23 are nodes that are common between two populations, while node-group 123 is common for all tree populations.*

The distribution of the leaves across different nodes is useful when deciding on the level at which to define haplogroups. To define haplogroups informative in regards of the migration events that link P1 and P2 with P3 one would want to avoid a cutoff time that defines haplogroups which only occur in a single population as well as haplogroups that are very ancestral and occur in all populations at similar frequencies. Based on the careful inspection of such plots across all different simulation scenarios, I decided to define haplogroups between divergence and admixture, as this time point seems to minimize the number of unwanted uninformative haplogroups.

3.2 mtDNA and Y-chromosome frequencies and sex-bias estimates

For clarity of plots and tables Tadmx = 400 is left out, which leads to lower resolution for very ancient times, but does not change the overall conclusions.

The distribution of the resulting frequencies of uniparental markers in P3 after clustering for all runs were plotted for every simulation scenario with strong sex-bias (p1MaleRatio = 80) (Figure 8).

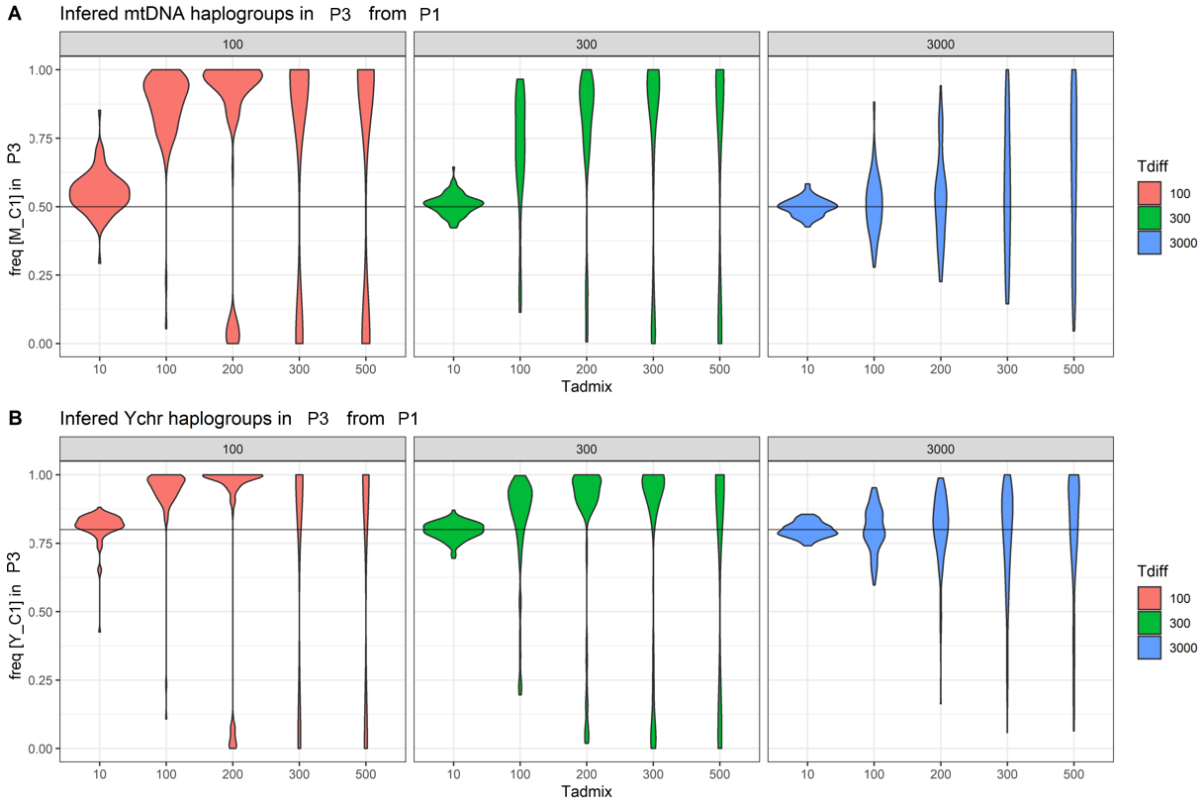

Figure 8 *Violin plots of mtDNA (A) and Y chromosome (B) haplogroup frequencies of P3 in the cluster associated with P1 for all scenarios with sex-bias (p1MaleRatio = 80). Each scenario has 100 simulation runs. The horizontal lines represent the expected frequency based on the simulation setup*

For Tadmx = 10 generations the frequencies distribute around the expected value for every Tdiff (Figure 8). This suggests that with recent admixture events the frequencies of uniparental markers reflect the sex-biased gene flow. For Tadmx = 100 generations the distributions already show a strong spread due to a long period of genetic drift. As a consequence of drift, the increasing Tadmx is leading to the skewed distributions towards the extreme values of 0 and 1 (i.e. fixation or elimination of certain haplogroups). For example, a haplogroup that occurs at a frequency of 0.8 has a higher probability of increasing in frequency, and eventual fixation when compared to haplogroups with lower frequencies. For this reason, the Y-chromosome haplogroups associated with P1 starting with a frequency of 0.8 in P3 have a higher likelihood of reaching fixation in P3 than the haplogroups associated with P2. The distributions show that there are still cases despite strong sex-bias, where the haplotypes associated with P2 reach fixation.

Tdiff has a weaker, yet noticeable, impact on the haplogroup frequencies when compared to Tadmx. For higher Tdiff values the distributions are less skewed and spread less broadly than those with lower values of Tdiff. This might be due to higher likelihood of complete lineage sorting between P1 and P2 and thus, easier classification of haplogroups to either cluster associated with P1 or P2.

The difference of mtDNA frequencies and Y-chromosome frequencies within the cluster associated with P1 can be used as a measure of the sex-bias. When summarized over all simulation setups, this measure supports the previous results based on frequencies of the separate genetic markers (Figure 9).

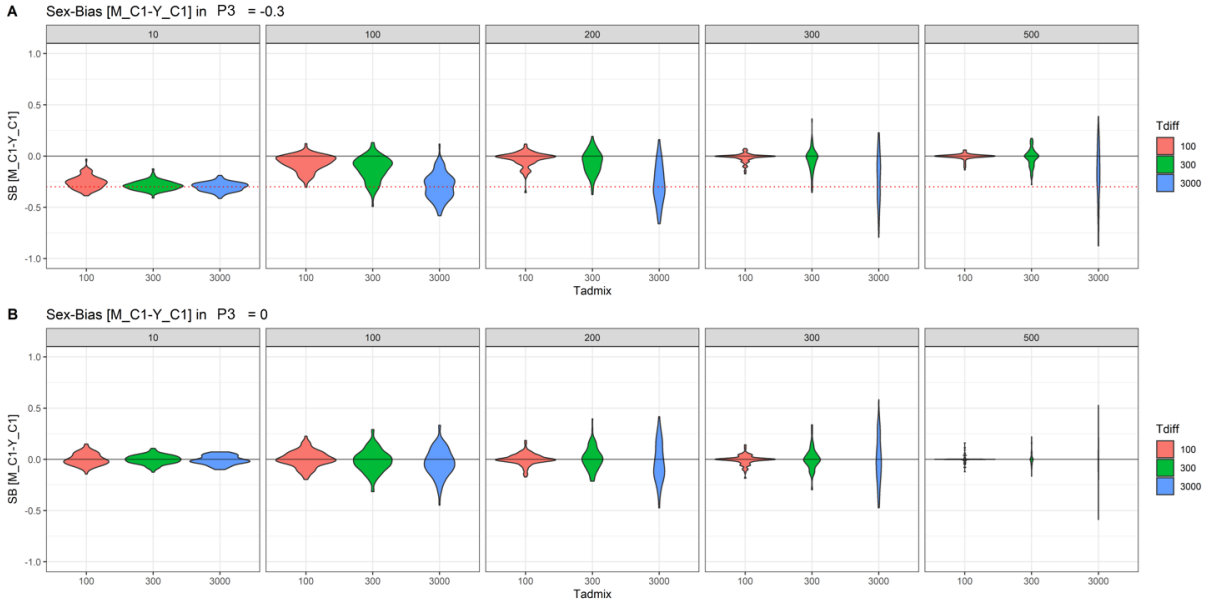

Figure 9 *Distributions of inferred sex-biases across different simulation setups. The expected sex-biases are calculated as difference of mtDNA and Y-chromosome frequencies in the cluster associated with P1. The dotted red line represents the actual sex-bias for SB = 80 (horizontal line at -0.3, which would represent 0.5 frequency for the mtDNA minus 0.8 frequency for the Y-chromosome). Distributions are plotted for every scenario with (A) and without (B) sex-bias.*

For Tadmx = 10 distributions are narrowly spread around the simulated sex-bias. This indicates that the inferred sex-bias based on frequencies of uniparental markers can be considered trustworthy for very recent admixture events. For Tadmx = 100 distributions already spread more broadly and start to shift towards not detecting sex-bias even for the simulations with strong sex-bias (Figure 9 A). This process continues for increasing Tadmx. For Tdiff = 3000 the distributions seem to stay centered around the simulated sex-bias longer through time. Still with increasing Tadmx the distributions already show that it is unlikely to infer the correct sex-bias even for deeply divergent populations. In summary the distributions of inferred sex-biases suggest that uniparental markers and their respective frequencies might only be considered reliable when admixture is recent. More deeply divergent population might still be able to give good estimates for sex-biased gene flow when admixture was less recent.

3.3 Accuracy calculations

To gauge the accuracy of the methodology of using the frequencies of uniparental markers to infer sex-biases the rmse values for all scenarios are provided in Table 1 for strong sex-bias and in Table 2 for no-sex-bias. In addition, the percentage of inferred sex-bias within the intervals of ± 0.01, 0.05 and 0.1 of the simulated sex-bias are shown.

Table 1. *Statistical evaluations of inferred sex-bias versus true sex-bias for all simulation runs with strong sex-bias (p1MaleRatio = 80).*

Tdiff	SB	Tadmx	rmse	within ±0.01	within ±0.05	within ±0.1
100	80	10	0.07627051	11	52	84
100	80	100	0.2447511	1	1	7
100	80	200	0.27976272	0	0	1
100	80	300	0.28844745	0	0	0
100	80	500	0.29733446	0	0	0
300	80	10	0.0448843	21	76	96
300	80	100	0.21004645	4	13	23
300	80	200	0.26007441	1	3	12
300	80	300	0.28462702	1	4	8
300	80	500	0.2970983	0	1	2
3000	80	10	0.04216681	15	79	98
3000	80	100	0.12616382	2	25	61
3000	80	200	0.18728201	8	22	38
3000	80	300	0.24756779	3	9	27
3000	80	500	0.29909258	2	18	22

Table 2. *Statistical evaluations of inferred sex-bias versus true sex-bias for all simulation runs with no sex-bias (p1MaleRatio=50).*

Tdiff	SB	Tadmx	rmse	within ±0.01	within ±0.05	within ±0.1
100	50	10	0.05672751	11	58	94
100	50	100	0.07850949	14	55	80
100	50	200	0.05529042	29	74	92
100	50	300	0.04749168	40	79	93
100	50	500	0.03546731	65	89	96
300	50	10	0.04304242	18	81	98
300	50	100	0.10938286	7	37	66
300	50	200	0.10907745	12	39	67
300	50	300	0.10487555	17	48	71
300	50	500	0.06566963	36	67	87
3000	50	10	0.04374048	18	70	99
3000	50	100	0.13701631	6	29	56
3000	50	200	0.18869531	3	19	32
3000	50	300	0.22513702	0	20	32
3000	50	500	0.24880661	1	15	35

The rmse values of Table 1 confirm the assumption that inferring sex-biased gene flow correctly is less precise for more ancient admixture events. The estimates of a Tadmx of 100 generations and more already lead to very high rmse-values representing how the methodology of inferring sex-bias based on haplogroup frequencies of uniparental markers can lead to false conclusions for more ancient

admixtures. Inferred sex-bias signals for admixture events 100 generations ago or older may not be considered trustworthy.

For higher Tdiff values the same Tadmx values seem to infer sex-bias more precisely. This suggests that for more divergent populations inferred sex-biases seem to be reflective of the true sex-bias for longer time after admixture. For example, for a Tadmx of 100 and deeply divergent populations represented by Tdiff = 3000 61% of inferred sex-biases lie within ±10% of the true sex-bias, whereas for a Tdiff of 100 and the same Tadmx = 100 it is only 7%.

Overall the findings presented here indicate that when quantifying sex-biases in admixture events one has to cautiously take the time of admixture, and the degree of divergence between source populations into consideration to avoid over-interpreting signals of sex-biased gene flow.

The overall accuracy in detecting sex-bias for the different Tadmx and Tdiff combinations is shown in Table 3.

Table 3. *Accuracy calculation for the detection of sex-bias in regards of the different Tdiff and Tadmx parameters.*

Tdiff	Tadmx	Acc
100	10	0.785
100	100	0.535
100	200	0.505
100	300	0.47
100	500	0.475
300	10	0.91
300	100	0.56
300	200	0.455
300	300	0.445
300	500	0.48
3000	10	0.855
3000	100	0.64
3000	200	0.525
3000	300	0.475
3000	500	0.42

The accuracy data indicates if sex-bias was detected correctly and does not indicate if sex-bias was quantified correctly. The results shown here suggest that accurate detection of sex-bias can only be done for recent admixtures events. Accuracy decreases rapidly after Tadmx = 100. For higher Tdiff values this decrease seems to be delayed, as seen in the quantitative results in Table 1 and 2. Qualitative detection of sex-biases in admixture, thus require cautious consideration of the time of admixture, and the degree of divergence between source populations.

Important to mention is that the time of haplogroup definition was set to be the average of Tadmx and Tdiff. This means that haplogroups are defined earlier for the simulation scenarios with Tdiff = 3000. Therefore, any observed effects of Tdiff might instead reflect the time at which haplogroups were defined, and not necessarily the effect of Tdiff itself. The time of haplogroup definition might have an important impact on sex-bias estimates and should be further investigated.

3.4 Visualization of intensity of sex-biased admixture

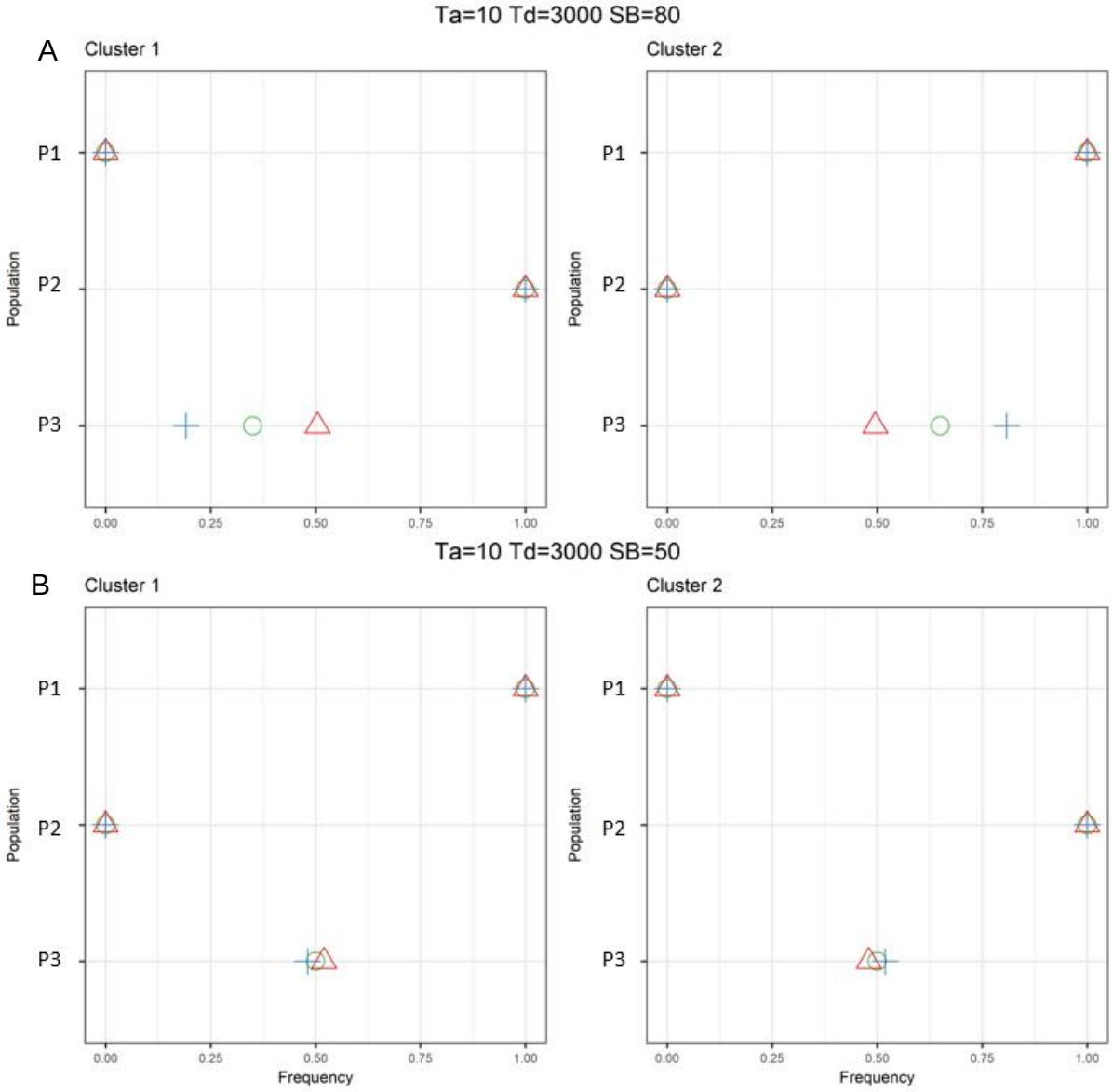

Figure 10 *Visualization of the intensity of sex-bias across populations for the two clusters of uniparental lineages. Each plot shows the sums of haplogroup frequencies for the mtDNA (red triangles) and MSY (blue plus signs) associated with cluster 1 or cluster 2. In addition, the expected autosomal frequencies for the described scenarios (Tadmx = 10, Tdiff = 3000 with no and strong sex-bias) were marked with a green circle.*

The differences in admixture proportions inferred from different genetic markers are indicative of sex-biased processes. Thus, if a population is found to have higher estimates of e.g. MSY haplogroups associated with P1 and lower estimates of mtDNA haplogroups associated with the same population, then we can assume that such population could have experienced sex-biased gene flow, in which more males than females contributed to the admixed population. If all markers associated with the source population are indicating the same value of admixture proportions in the admixed population, then we can assume that the admixed population did not experience sex-biased gene flow from the source population. To illustrate, here I show example plots for simulation scenarios with Tadmx=10 and Tdiff=3000 with (Figure 10A) and without sex-biased gene flow (Figure 10B). The frequencies of mtDNA and MSY within each cluster were calculated as sums of frequencies of mtDNA and MSY

haplogroups belonging to the same clusters inferred after running CA and HCPC on the haplogroup frequency tables. In Figure 10A population P1 is associated with cluster 2 and P2 is associated with cluster 1. Therefore, the frequencies of haplogroups associated with cluster 2 in P3 may reflect the admixture proportion from P1 to P3. The observed differences in Figure 10A are in accordance with the simulated sex-bias (i.e. 50% of females in P3 are from P1 and 80% of males in P3 are from P1). The same is true for the example plot of no-sex-bias (Figure 10B).

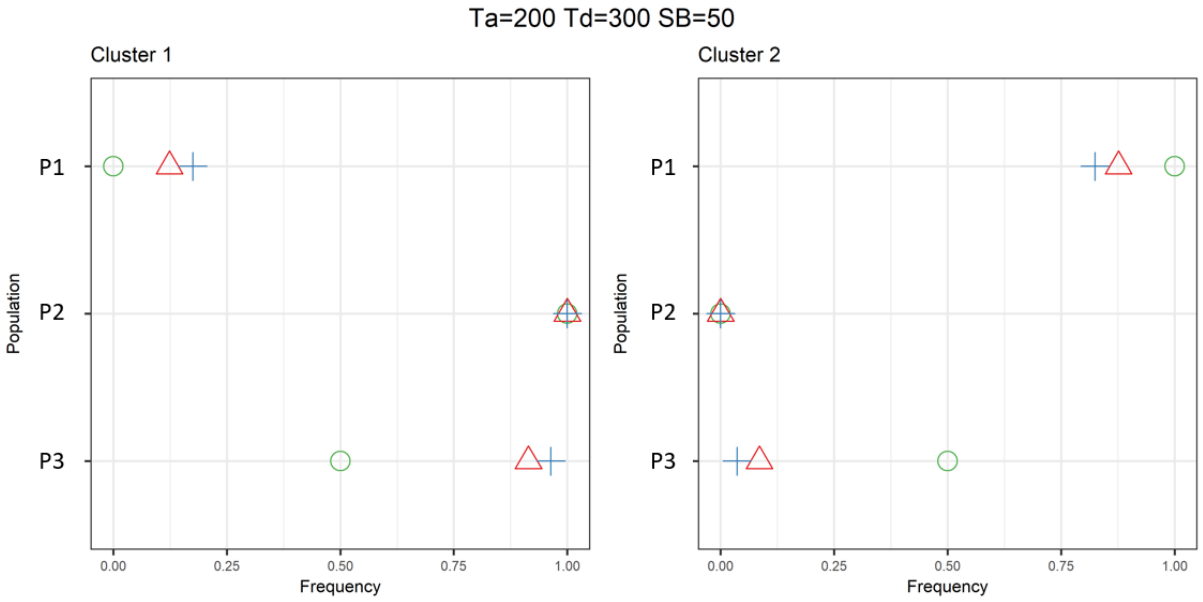

Figure 11 *Example of one simulation run in which P1 and P2 haplogroup frequencies are not fully assigned to their respective clusters.*

In rare occasions the inferences based on CA and HCPC are introducing some biases. The methodology of inferring sex-biases is based on the assumption that each of the source populations should be highly associated with only one of the cluster. This, however, was not always the case. This is illustrated in Figure 11 where P1 has a considerable amount of haplogroups associated with cluster 2 (i.e. the cluster characteristic to P2), and where the majority of P3 haplogroups got associated with P2 (despite the equal contributions of males and females from P1 and P2 in P3 simulated). This could be due to the incomplete lineage sorting between P1 and P2, and the fact that specific haplogroups occurred in both populations. It is also hard to predict how P3 specific haplogroups would cluster (in case of sex-bias they might be attracted towards P1 due to the higher overall sharing caused by the sex-bias when compared with P2).

4 Conclusion

In this thesis, I demonstrate that the detection and quantification of sex-biased admixture using mtDNA and Y chromosome haplogroup frequencies is dependent on the time of admixture, and the time of divergence of the source populations. The comparisons of uniparental lineages might result in signals of sex-biases admixture that are over- or under-estimating the true admixture event, and thus can lead to falls conclusions. This is especially true for the scenarios in which the time of admixture was not recent enough. Further research should focus on exploring other parameters that might play an important role in detecting sex-biased admixture, such as the time at which to define haplogroups. This study provides the framework for further research, and it can be easily expanded to investigate diverse genetic or demographic parameters, their impact on haplogroup frequencies, and the detection of sex-biased processes, not only in human but any sexually reproducing species.

5 Code and data availability

The source code for the project can be found here: https://github.com/JoshuaGensel/SLiM_project

6 Figures, Tables and Code Snippets

6.1 Figures

6.2 Tables

6.3 Code Snippets

7 References

1. Jobling M, Hurles M, Tyler-Smith C. Human evolutionary genetics: origins, peoples & disease. Garland Science; 2013 Jun 25.

2. Underhill PA, Kivisild T. Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. Annu. Rev. Genet.. 2007 Dec 1;41:539-64.

3. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH. Sequence and organization of the human mitochondrial genome. Nature. 1981 Apr;290(5806):457-65.

4. Satoh M, Kuroiwa T. Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell. Experimental cell research. 1991 Sep 1;196(1):137-40.

5. Emery LS, Magnaye KM, Bigham AW, Akey JM, Bamshad MJ. Estimates of continental ancestry vary widely among individuals with the same mtDNA haplogroup. The American Journal of Human Genetics. 2015 Feb 5;96(2):183-93.

6. International Society of Genetic Genealogy, Genetics Glossary 2021. https://isogg.org/wiki/Genetics_Glossary

7. Van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Human mutation. 2009 Feb;30(2):E386-94.

8. Y Chromosome Consortium. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome research. 2002 Feb 1;12(2):339-48.

9. Rojas W, Parra MV, Campo O, Caro MA, Lopera JG, Arias W, Duque C, Naranjo A, García J, Vergara C, Lopera J. Genetic make up and structure of Colombian populations by means of uniparental and biparental DNA markers. American Journal of Physical Anthropology. 2010 Sep;143(1):13-20.

10. Wen B, Xie X, Gao S, Li H, Shi H, Song X, Qian T, Xiao C, Jin J, Su B, Lu D. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. The American Journal of Human Genetics. 2004 May 1;74(5):856-65.

11. Verdu P, Becker NS, Froment A, Georges M, Grugni V, Quintana-Murci L, Hombert JM, Van der Veen L, Le Bomin S, Bahuchet S, Heyer E. Sociocultural behavior, sex-biased admixture, and effective population sizes in Central African Pygmies and non-Pygmies. Molecular biology and evolution. 2013 Apr 1;30(4):918-37.

12. Stefflova K, Dulik MC, Pai AA, Walker AH, Zeigler-Johnson CM, Gueye SM, Schurr TG, Rebbeck TR. Evaluation of group genetic ancestry of populations from Philadelphia and Dakar in the context of sex-biased admixture in the Americas. PloS one. 2009 Nov 25;4(11):e7842.

13. Yuan X, Miller DJ, Zhang J, Herrington D, Wang Y. An overview of population genetic data simulation. Journal of Computational Biology. 2012 Jan 1;19(1):42-54.

14. Liu Y, Athanasiadis G, Weale ME. A survey of genetic simulation software for population and epidemiological studies. Human genomics. 2008 Dec;3(1):1-8.

15. Bajić V, Barbieri C, Hübner A, Güldemann T, Naumann C, Gerlach L, Berthold F, Nakagawa H, Mpoloka SW, Roewer L, Purps J. Genetic structure and sex-biased gene flow in the history of southern African populations. American journal of physical anthropology. 2018 Nov;167(3):656-71.

16. Haller BC, Messer PW. SLiM 3: forward genetic simulations beyond the Wright–Fisher model. Molecular biology and evolution. 2019 Mar 1;36(3):632-7.

17. Haller, B.C., and Messer, P.W. (2016). SLiM: An Evolutionary Simulation Framework. URL: http://benhaller.com/slim/SLiM_Manual.pdf

18. Haller BC, Galloway J, Kelleher J, Messer PW, Ralph PL. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. Molecular ecology resources. 2019 Mar;19(2):552-66.

19. Python software Foundation. Python Language Reference, version 3.8.5. Available at http://www.python.org

20. Tskit Developers, tskit API Reference, version 0.3.4. Available at https://tskit.dev/tskit/docs/stable/introduction.html

21. pySLiM developers revision ced27369, pyslim API Reference, version 0.501. Available at https://pyslim.readthedocs.io/en/latest/index.html

22. Posth C, Renaud G, Mittnik A, Drucker DG, Rougier H, Cupillard C, Valentin F, Thevenet C, Furtwängler A, Wißing C, Francken M. Pleistocene mitochondrial genomes suggest a single major dispersal of non-Africans and a Late Glacial population turnover in Europe. Current Biology. 2016 Mar 21;26(6):827-33.

23. Helgason A, Einarsson AW, Guðmundsdóttir VB, Sigurðsson Á, Gunnarsdóttir ED, Jagadeesan A, Ebenesersdóttir SS, Kong A, Stefánsson K. The Y-chromosome point mutation rate in humans. Nature genetics. 2015 May;47(5):453-7.

24. Barbieri C, Hübner A, Macholdt E, Ni S, Lippold S, Schröder R, Mpoloka SW, Purps J, Roewer L, Stoneking M, Pakendorf B. Refining the Y chromosome phylogeny with southern African sequences. Human genetics. 2016 May 1;135(5):541-53.

25. McKinney W. Data structures for statistical computing in python. InProceedings of the 9th Python in Science Conference 2010 Jun 28 (Vol. 445, pp. 51-56).

26. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R. Array programming with NumPy. Nature. 2020 Sep;585(7825):357-62.

27. Lippold S, Xu H, Ko A, Li M, Renaud G, Butthof A, Schröder R, Stoneking M. Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. Investigative genetics. 2014 Dec;5(1):1-7.

28. Kelleher J, Etheridge AM, McVean G. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS computational biology. 2016 May 4;12(5):e1004842.

29. Lê S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. Journal of statistical software. 2008 Mar 18;25(1):1-8.

30. Hamner B, Frasco M, LeDell E. Metrics: Evaluation metrics for machine learning. R package version 0.1. 2018;4:2018.

Selbstständigkeitserklärung

Hiermit erkläre ich die Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und diese als solche kenntlich gemacht habe.

Name: Leon Joshua Gensel

Matrikelnummer: 5222745

Berlin, 05.07.2021

X 

Leon Joshua Gensel