Universidad Autónoma de Querétaro
Facultad de Ingeniería
Maestría en Ingeniería Matemática

Methodology to perform demographic inferences using ancient and modern DNA samples

Tesis

Que como parte de los requisitos para obtener el Grado de
Maestro en Ingeniería Matemática

Presenta
Liliana Chavaje Ávila

Dirigida por:
Dr. Vicente Diego Ortega Del Vecchyo

Codirigida por:
Dr. Roberto Augusto Gómez Loenzo

Dr. Vicente Diego Ortega Del Vecchyo
Presidente

Dr. Roberto Augusto Gómez Loenzo
Secretario

M. en C. Luisa Ramírez Granados
Vocal

Dra. Angélica Rosario Jiménez Sánchez
Suplente

Dr. Víctor Antonio Aguilar Arteaga
Suplente

Centro Universitario, Querétaro, Qro a 31 de enero de 2023.

A

Aitana, Emilio y Andrés

por estar siempre a mi lado.

AGRADECIMIENTOS

A mis padres por su apoyo y amor incondicional.

Al Dr. Diego Ortega por haber compartido su conocimiento conmigo y apoyarme durante toda la investigación y elaboración de la tesis. Le agradezco su paciencia y su determinación a no dejarme claudicar.

A mis sinodales, al Dr. Víctor Aguilar, al Dr. Roberto Gómez, a la Dra. Angélica Jiménez y a la Dra. Luisa Ramírez, por su apoyo y guía. Además agradezco a Roberto que en su papel como coordinador de la Maestría, siempre estuvo dispuesto a ayudarme a resolver cualquier eventualidad.

A mis profesores, por todas sus valiosas enseñanzas. Particularmente, quiero agradecer a mis profesoras Luisa Ramírez, Patricia Spíndola, Angélica Jiménez, Esperanza Trenado y Laura Franco por darme un ejemplo a seguir y continuar fomentando mi amor a las Matemáticas.

A la Universidad Autónoma de Querétaro por darme la oportunidad de pertenecer a su facultad de Ingeniería y darme una formación sólida.

Al LIIGH - UNAM (Laboratorio Internacional de Investigación sobre el Genoma Humano) por haberme permitido colaborar en su proyecto y recibirme como miembro de su equipo.

Al Laboratorio de Super Cómputo de la UNAM por su apoyo con sus recursos tecnológicos.

Al Conacyt por la beca que me brindó para realizar los estudios de Maestría.

A mis compañeros de generación por compartir sus conocimientos, recursos y ocurrencias.

INDEX

INDEX OF FIGURES

INDEX OF TABLES

Resumen

La secuenciación de genomas antiguos nos permite trazar mejor nuestra historia mediante el estudio conjunto de genomas de individuos del presente y del pasado. El análisis de estos genomas requiere de la aplicación y desarrollo de métodos estadísticos y computacionales que nos permitan inferir la historia que mejor explica los patrones de diversidad genética observada en los genomas estudiados. Un estadístico de diversidad genética muy utilizado para la inferencia de la historia es el espectro de frecuencias por sitio. Este estadístico ha sido ampliamente utilizado para inferir el pasado empleando genomas modernos. Sin embargo, todavía no se ha analizado a profundidad la posibilidad de utilizar este estadístico para inferir la historia demográfica en el pasado mediante el análisis conjunto de genomas modernos y antiguos. En esta tesis realizo un análisis computacional para evaluar si el espectro de frecuencias por sitio es informativo para inferir un parámetro demográfico muy informativo sobre la historia en el pasado, el tiempo de divergencia entre poblaciones. Para llevar esto a cabo, se desarrolló una metodología que permite calcular el espectro de frecuencias por sitio a través de un conjunto de genomas simulados. Las simulaciones que se llevan a cabo muestran que el espectro de frecuencias por sitio es sensible a cambios del tiempo de divergencia entre poblaciones y del tamaño de muestra en una población antigua. Los análisis muestran que el tiempo de divergencia es ligeramente subestimado bajo simulaciones con tiempos de divergencia que son cercanos al tiempo de divergencia entre distintas poblaciones en América. Se encuentra que un aumento en el tamaño de muestra mejora los estimados del tiempo de divergencia. Los resultados muestran que es conveniente realizar simulaciones para validar si, dado un cierto diseño de estudio con tamaños específicos de muestra en una población antigua, es posible realizar estimaciones certeras de parámetros demográficos de interés como el tiempo de divergencia.

Palabras clave: Simulaciones, Inferencia, Genética de poblaciones, Historia demográfica.

Abstract

The genome sequencing of ancient genomes allows us to understand our history based on the joint study of present-day and ancient genomes. The analysis of those genomes requires the application and development of statistical and computational methods that allow us to infer the past history that better explains genetic diversity patterns on the studied genomes. A widely used statistic for the study of past population history is the site frequency spectrum. This statistic is used to elucidate the past using present-day genomes. However, we currently do not know if this statistic can be used to infer the past demographic history through the joint analysis of ancient and present-day genomes. In this dissertation I perform a computational analysis to evaluate if the site frequency spectrum is informative to analyze a very informative demographic parameter about the past history of different individuals, the divergence time between populations. To do this, I develop a methodology that allows me to calculate the site frequency spectrum in a set of simulated genomes. The simulations performed show that the site frequency spectrum depends on divergence time changes between populations and to changes in the sample size in an ancient population. The analysis shows that the divergence time is slightly underestimated in simulations where the divergence time employed is close to the divergence time between different populations in America. It is found that increases in the sample size in an ancient population improves estimates of the divergence time. The results show that it is convenient to perform simulations in order to validate if, given a certain study design with specific sample sizes in an ancient population, it is possible to perform accurate estimations of parameters of interest such as the divergence time between populations.

Keywords: Simulations, Inference, Population Genetics, Demographic history

Introduction

The information contained in ancient genomic DNA allows us to obtain new perspectives on human history based on inferences from the genomic comparisons between ancient and modern individuals. Ancient DNA has led to the discovery of previously unknown populations of archaic hominins (Carlhoff et al., 2021). It has also allowed us to discover information about the migration patterns between different populations across the world (Figure 1) (Nielsen, Akey, & Jakobsson, 2017). For example, it is commonly accepted that America's ancestral population started from migration events across the Bering land bridge (Meltzer, 2009). However, the details are not clear, and the analysis of genomic data has provided evidence about the history and divergence of ancient Beringians and ancient Native Americans (Moreno-Mayar, Potter, et al., 2018). Additionally, ancient DNA gave evidence that Native Americans descend from at least three migration events from Asian populations (Reich et al., 2012). More surprisingly, genomic evidence shows that some South American individuals descend in part from a population that has a close association to present-day Australians and other populations from that geographic area (Skoglund et al., 2015).

The study of genomics to infer past demographic history from ancient samples has taken an important role during the last two decades (Figure 2). Although genomic material has been obtained since the last 30 years (Pääbo, 1985), the introduction of Next-generation sequencing (NGS) methods permits DNA sequencing at a large scale, increasing throughput dramatically (Schuster, 2008). Next-generation sequencing can allow us to obtain genomic information from 300 kb using 10 ng of DNA, a significant number compared to Sanger sequencing that would provide 1 kb for that DNA quantity (Illumina, 2021b). Additionally, improvements in the NGS technology have dramatically reduced the cost of DNA sequencing in the last 20 years. As an example, the cost of sequencing a human genome has been reduced from $340,000 USD in 2008 to $4,200 USD in 2015 (Muir et al., 2016).

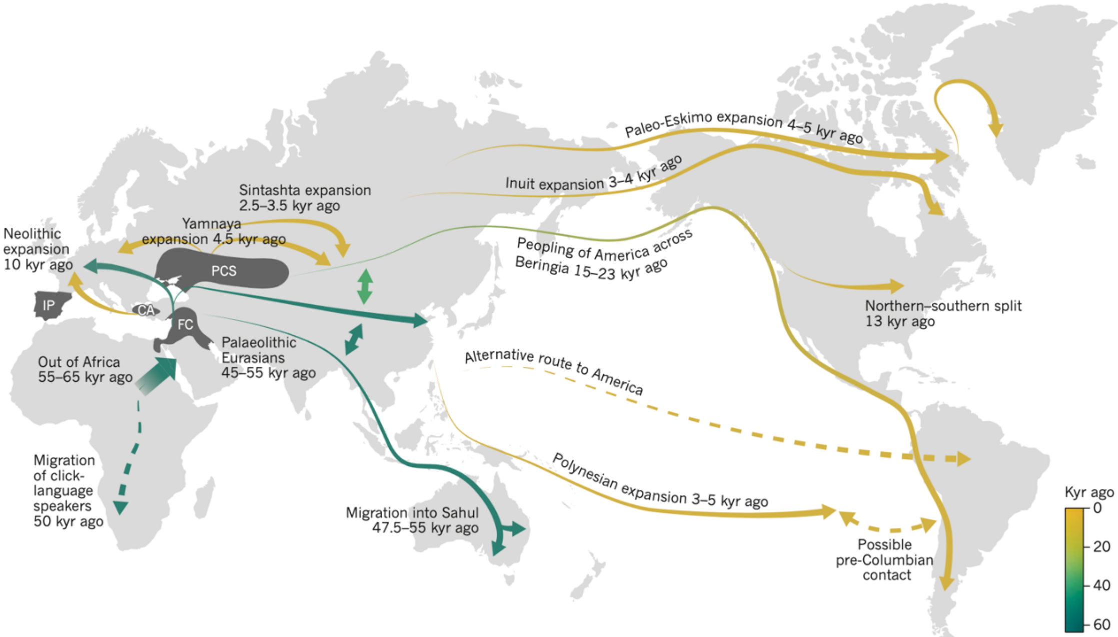


Figure 1.- Migration of humans throughout the world (Figure taken from Nielsen et al., 2017)

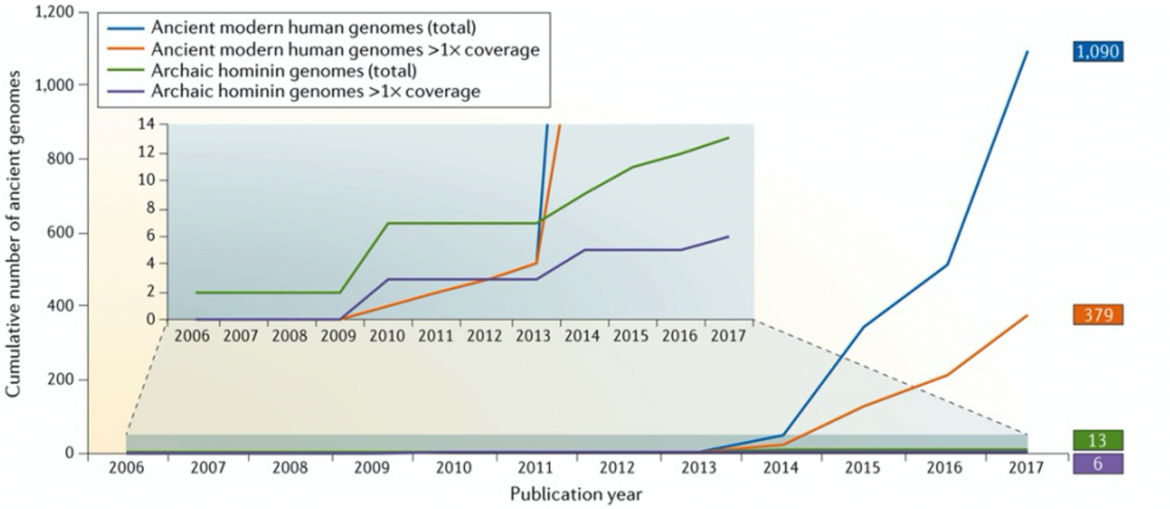


Figure 2.- Cumulative number of sequenced ancient individuals, including modern humans and archaic hominins (Figure taken from Marciniak and Perry, 2017).

Ancient DNA characteristics

Ancient DNA (aDNA) has characteristics that make its study challenging. These include fragmentation, low coverage, degradation, and contamination (Günther & Jakobsson, 2019). Fragmentation is a caused by depurination by hydrolysis, where water breaks chemical bonds in the DNA structure, and by β-elimination (Figure 3). The result of this reaction is that the DNA strand is broken into small fragments (Lindhal, 1993). On the other hand, coverage is defined by how many times a position of the genome is read on average by the NGS sequencing machine. A low coverage takes place when each genomic position is read on average less than one time by the NGS sequencing machine. Low coverage and fragmentation are two factors that make a bad combination in terms of determining the genetic content of an ancient sample (Figure 4). Older samples yield lower coverages when sequenced, preventing the determination of the genomic DNA sequence correctly. The low coverage is due to DNA fragmentation, which degrades DNA sequences over time, and, ultimately, does not allow the NGS sequencing machine to reconstruct the DNA sequences.

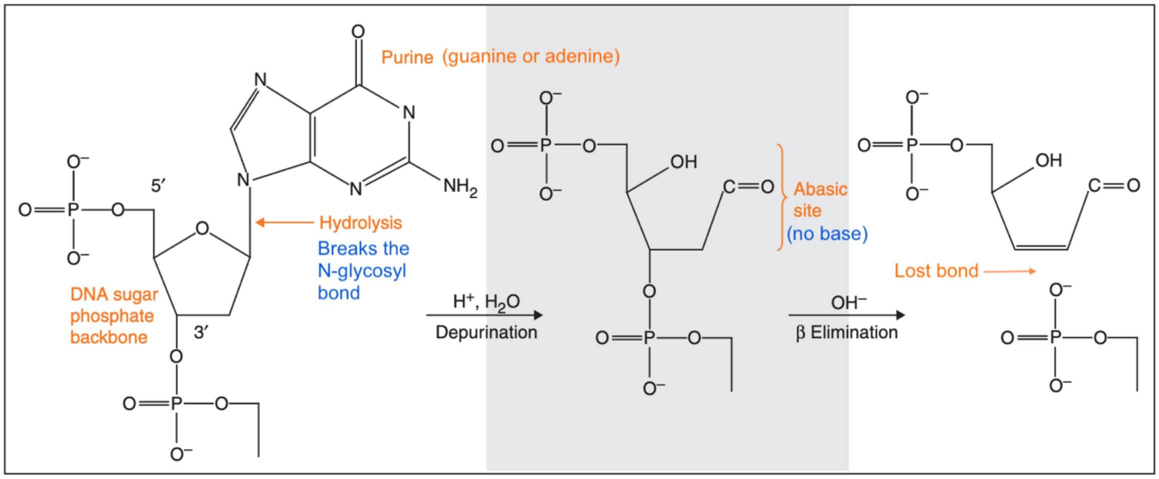

Purine (guanine or adenine)
5′
3′
DNA sugar phosphate backbone
Hydrolysis
Breaks the N-glycosyl bond
H+, H2O
Depurination
Abasic site
(no base)
OH−
β Elimination
Lost bond

Figure 3.- Ancient DNA damage.
Ancient DNA can be damaged by chemical reactions. The first one is depurination by hydrolysis, where a water molecule breaks a bond (N-glycosyl) with a purine. From the four bases of a DNA strand, adenines (A) and guanines (G) are purines, while cytosines (C) and thymines (T) are pyrimidines. After losing its base, another chemical reaction (β elimination), degrades the contents even further**.** This causes fragmentation (Figure taken from Dabney, Meyer, & Pääbo, 2013).

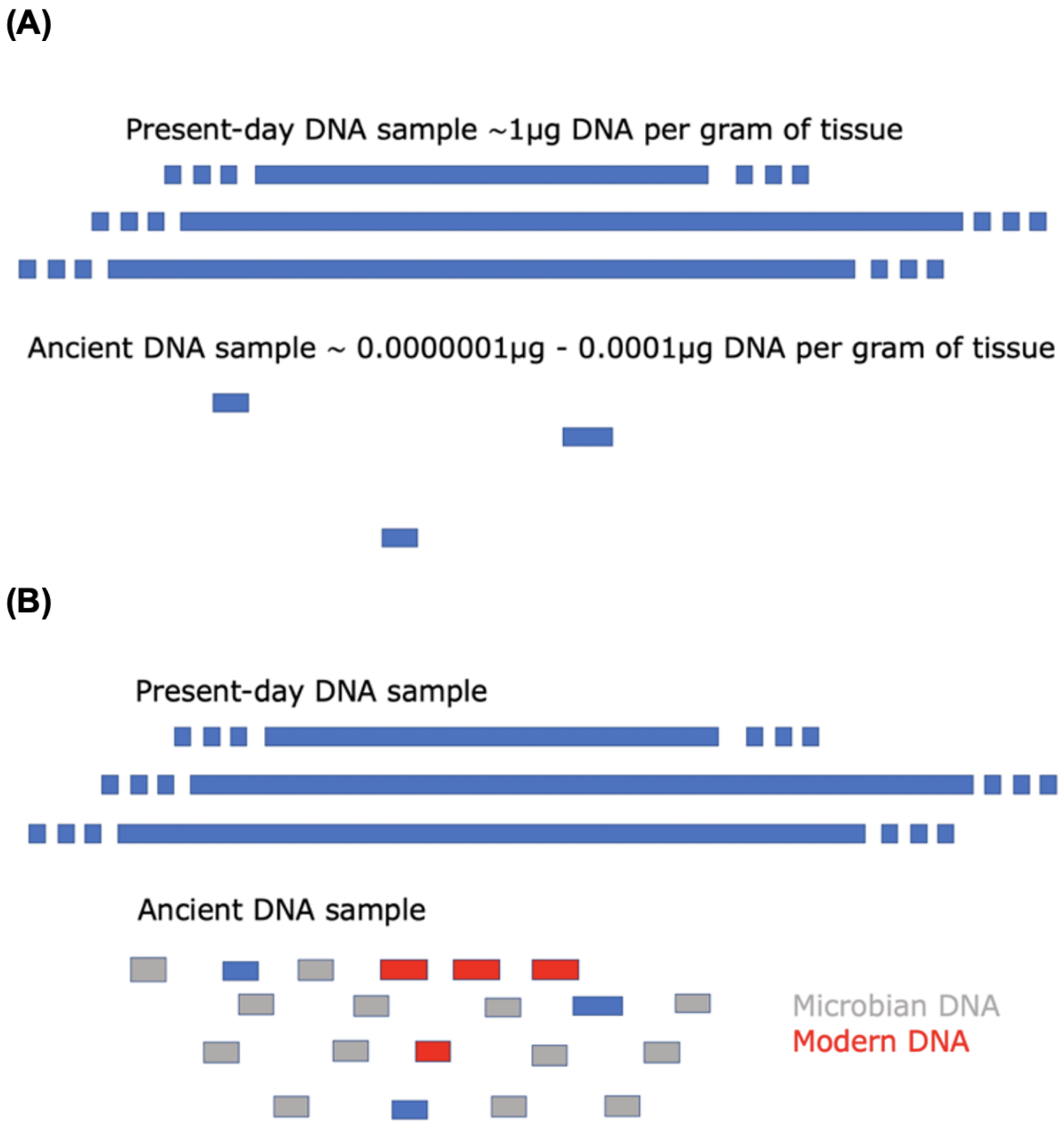

(A)
Present-day DNA sample ~1µg DNA per gram of tissue
Ancient DNA sample ~ 0.0000001µg - 0.0001µg DNA per gram of tissue
(B)
Present-day DNA sample
Ancient DNA sample
Microbian DNA
Modern DNA

Figure 4.- Comparison between modern and ancient DNA samples.
(A) The reads obtained for the sequenced DNA samples are represented in blue. The diagram shows a modern-day sample with high coverage and an ancient sample with low coverage and fragmentation. The amount of data extracted from a fresh sample is high, even from a small sample. Also, the reads obtained will be long and abundant. In comparison, the amount of data obtained from an ancient DNA sample is considerably less and the reads very short. (B) Ancient DNA samples can have high levels of contamination due to environmental and handling factors. Grey squares represent microbial contamination. Ancient DNA is usually found in fossils that are exposed to heat, humidity, and other environmental factors for a long time, making it prone to contamination. Ancient DNA samples can also be easily contaminated by other individuals during extraction. Contamination by modern DNA is represented by the red rectangles (Figure taken from Pääbo, 2018).

Degradation is a process that changes the DNA sequences over time. Hydrolysis can cause post-mortem damage by breaking the amino bond (NH_2) in cytosines present in a DNA strand. When deamination of cytosines occur, these bases are converted to uracils (U). (Figure 5). As uracils are not present in DNA samples, sequencing machines misinterprets them as thymines (T) (Briggs et al., 2007). C to T misincorporations can result in incorrect inferences about the genotypes present in each position of the genome (Ho, Heupink, Rambaut, & Shapiro, 2007).

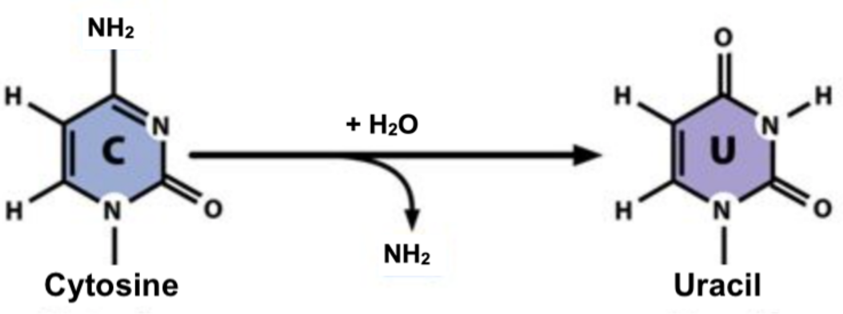


Figure 5.- Deamination of cytosines (C) to uracils (U).
Molecules of water (H_2O) can break the bond with NH_2 in cytosines and form new ones to convert into uracil. This causes sequencing errors because uracils are not usually present in DNA, causing a false thymine (T) incorporation. Nowadays, the abundance of thymines at the ends of a strand helps authenticate a sample as ancient (Figure taken from Briggs et al., 2007).

Finally, contamination is the mixture of the DNA sample with genomic material from other individuals and even other species. Microbes, multiple handling and even dust contamination can overwhelm ancient DNA samples, distorting the contents of the sequences (Richards, Sykes, & Hedges, 1995). (Figure 4B).

Working with fossils and obtaining ancient DNA from them is a complicated process that must be thoroughly planned because they are often unique and very fragile. Extraction can also be difficult because the specimen can get destroyed, losing valuable material forever. Also, as mentioned above, samples are easily contaminated (Orlando et al., 2021).

The challenges of extracting and sequencing an ancient sample is an area of constant new developments. Apart from this topic, the analysis of these samples requires analysis that consider the features of ancient DNA explained in this section (fragmentation, low coverage, degradation, and contamination). Currently the field is working on creating statistical methods that model these features. A review of the methods to analyze ancient DNA data to infer past demographic history will be taken in chapter II.

Motivation

The number of sequenced genomes of ancient modern humans and archaic hominins in the Americas is still very low compared to other regions in the world. Figure 6 shows that until the publication of the data, the number of genomes in the Americas was 64, compared to 882 in Europe and 85 in the Middle East. Also, the samples are considerably less ancient than in other regions. Overall, the region is understudied and there is a need for more research on archeological sites and caves.

The oldest culture known in the Americas is the Clovis culture. The most ancient archeological evidence of Clovis people was estimated to be 13,000 to 12,600 calendar years BP (before present) and is located in North America (Rasmussen et al., 2014). However, evidence of human activity in a pre-Clovis era has been found in different sites in the region. These findings include tools, bones, and footprints (Table 1). In Mexico, researchers have found evidence of inhabitants from previous periods known to the region (Acosta Ochoa, 2010; Ardelean et al., 2020; Chatters et al., 2014; Des Lauriers, Davis, Turnbull, Southon, & Taylor, 2017; González et al., 2014; Sanchez et al., 2014). These findings could be useful to understand better the past history of ancient populations in the region (Table 2).

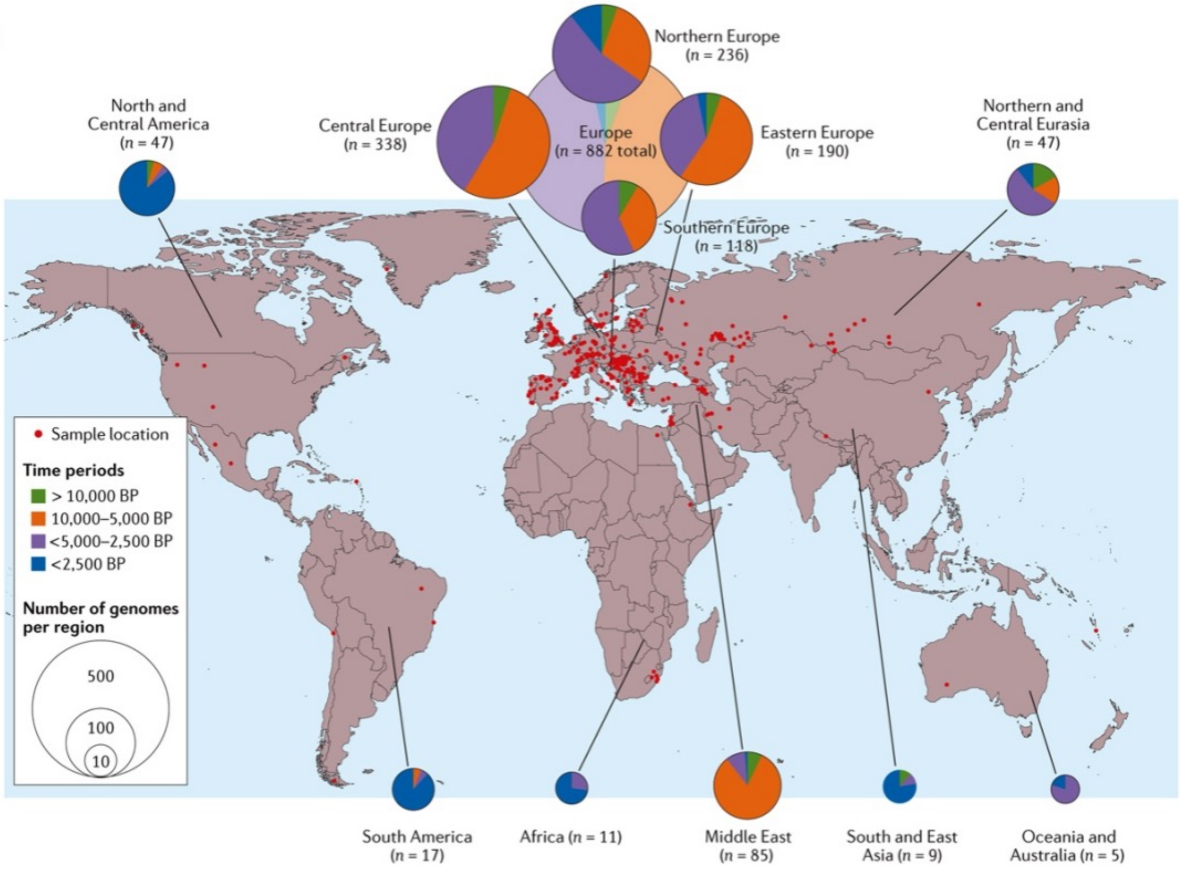


Figure 6.- Genome samples across the world (Figure taken from Marciniak & Perry, 2017).
This diagram shows the number of genomes of ancient samples according to location and age of the samples. Time periods are expressed in years BP (before present) that is an approximation established with radiocarbon dating (Taylor, 1985). Regions outside of Europe are clearly understudied.

The reconstruction of past human history requires an interdisciplinary approach that considers both Archaeology and Genetics. However, genetic analysis that use ancient DNA require statistical methods that model its features. The aim of this project is to contribute in the development of methods that infer our history using ancient DNA. Particularly, to continue developing methods for inferences of population parameters such as the divergence time between populations. Advances in this area will hopefully provide researchers in Mexico with more resources for the study of recently discovered ancient fossils.

Table 1.- Published evidence findings from Pre-Clovis Era in the Americas.

Approximate number of years ago	Evidence found	Site name	Location	References
13,200 - 15,500	Projectile points, blades, and other tools	The Buttermilk Creek Complex	Texas, USA	(Waters et al., 2011)
16,560 - 15,280	Stone artifacts and bones	Cooper's Ferry	Idaho, USA	(Davis et al., 2019)
23,000 - 21,000	Human footprints	White Sands National Park	New Mexico, USA	(Bennett et al., 2021)
26,500 - 19,000	Stone tools	Chiquihuite Cave	Zacatecas, Mexico	(Ardelean et al., 2020)

Table 2.- Examples of published evidence findings from Mexico.

Approximate number of years ago	Evidence found	Importance	Site name and location	References
11,300 - 10700	Fishhooks	Possible alternative Pacific coastal route from Asia.	Isla de Cedros, Baja California	(Des Lauriers et al., 2017)
12,500	Stone tools and arrow points	Evidence suggests a more complex behavior of inhabitants than previously known.	Santa Marta Cave, Chiapas	(Acosta Ochoa, 2010)
13,000 and 9,000	Human skeletons	Largest databases on bones of early humans in Mexico	Different caves and cenotes in Quintana Roo.	(González et al., 2014) (Chatters et al., 2014)
13,390	Bones, stone and bone artifacts	Oldest and southern most site from Clovis era.	El Fin del Mundo, Sonora	(Sanchez et al., 2014)
26,500 - 19,000	Stone tools	Oldest Pre-Clovis era site	Chiquihuite cave, Zacatecas	(Ardelean et al., 2020)

Thesis structure

Chapter 1 introduces the impact of research involving ancient DNA studies in our knowledge of ancient populations. It also highlights the growth of human and archaic hominin genomes thanks to the evolution in sequencing techniques and lower costs. This chapter also describes the problem and motivation that led to this project.

Chapter 2 includes a description of preliminary concepts to understand the project and a review of different statistical techniques that have been used for genomic analysis.

Chapter 3 and 4 contains the hypothesis and the objectives of this project, respectively.

Chapter 5 describes the methods that I followed to test the hypothesis and fulfill the objectives from chapters 3 and 4. The methodology is based on a shell script pipeline that will simulate ancient and modern-day DNA samples. It will also be the main vehicle to infer their joint site frequency spectrum and the divergence time between the two populations.

Chapter 6 presents the results obtained after following the methodology presented in chapter 5. This chapter also includes an analysis of these results.

Finally, in chapter 7 we describe the conclusions of this project.

Background

General concepts

This section introduces definitions that are relevant to this research project, assuming that most readers are not familiar with terminology related to genomic studies.

DNA sequencing

The nucleotide bases in DNA samples can be determined by a process called sequencing. In a human DNA strand, there are four possible nucleotides: A, C, G and T representing adenine, guanine, cytosine, and thymine respectively. After sequencing, this genomic information is presented as a list of letters (nucleotide bases) in a specific order. To establish the correct base, it is necessary to sequence a region multiple times and obtain several reads. The reads obtained after sequencing are compared with a reference. The average number of reads that align with the same base in a particular position is the coverage. A high coverage helps to determine confidently the bases of a DNA strand. To sequence a human whole genome, a coverage of 30· - 50· per base pair is recommended (Illumina, 2021a) (Figure 7).

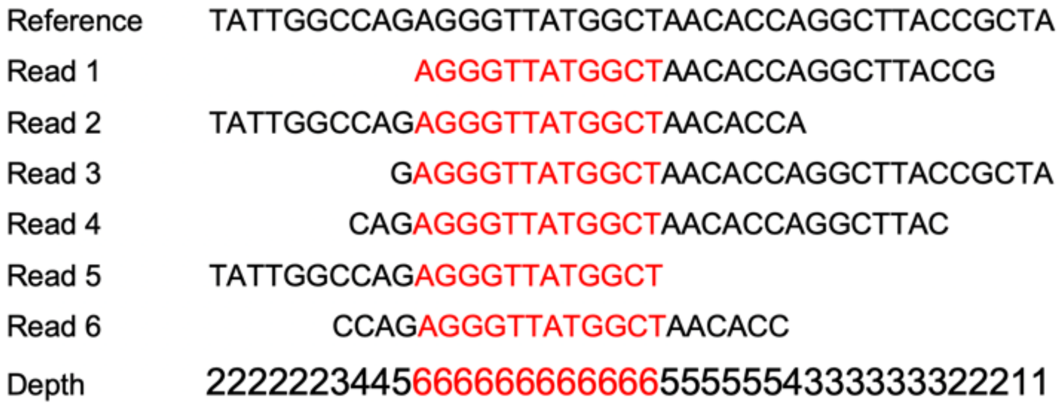

Reference TATTGGCCAGAGGGTTATGGCTAACACCAGGCTTACCGCTA
Read 1 AGGGTTATGGCTAACACCAGGCTTACCG
Read 2 TATTGGCCAGAGGGTTATGGCTAACACCA
Read 3 GAGGGTTATGGCTAACACCAGGCTTACCGCTA
Read 4 CAGAGGGTTATGGCTAACACCAGGCTTAC
Read 5 TATTGGCCAGAGGGTTATGGCT
Read 6 CCAGAGGGTTATGGCTAACACC
Depth 22222234456666666666655555543333332211

Figure 7.- Example of sequencing reads alignment and the depth.
The average depth across every position of the genome is the coverage of the sample. A coverage of 30× or more is recommended to correctly infer the genotype present in each position of a human genome.

Site Frequency Spectrum

The unfolded allele frequency spectrum or unfolded site frequency spectrum (SFS) represents the number of genomic positions that contain a certain number of derived alleles in a particular sample. The alleles are the nucleotide bases that are present in a sample of individuals for each position of the genome. The possible bases in each position of the genome are adenine, thymine, guanine, or cytosine. Segregating sites are the places in the genome where the bases differ between samples. The ancestral allele is determined by the ancestral state, present in the most recent common ancestor of all the samples collected for that particular segregating site (Nielsen & Slatkin, 2013). The derived allele is the other possible base that is not present in the ancestral state. If the ancestral allele is known it is possible to calculate the unfolded SFS, which is the frequency distribution of the positions with a certain number of derived alleles in the segregating sites of a given sample.

For example, given the following samples and ancestral state from the same population:

```
Ancestral state  ATTGGCCACAGGGTTATGGCTAACACCAGGCTTACCGCTT
Sample 1         ATTGCCCAGAGGGTTATGGCTTACACCAGGCTTACCGCTA
Sample 2         ATAGGCCAGAGGGTTATGGCTAACACCAGGCTAACCGCTT
Sample 3         ATTGGCCAGAGGGTTATGGCTTACACCAGGCTTACGGCTA
Sample 4         ATAGGCCAGAGGGTTATGGCTAACACGAGGCTAACCGCTA
Sample 5         ATTGGCCAGAGGGTTTTGGCTAACACGAGGCTTACCGCTT
```

The segregating sites are used to estimate the allele frequency spectrum:

```
Ancestral state  ATTGGCCACAGGGTTATGGCTAACACCAGGCTTACCGCTT
Sample 1         ____C___G____________T_________________A
Sample 2         __A_____G_______________________A_______
Sample 3         ________G____________T_____________G___A
Sample 4         __A_____G_________________G_____A______A
Sample 5         ________G______T__________G_____________
Derived alleles:   2 1   5      1     2    2     2  1   3
```

Since the ancestral state is available for this population, then the unfolded site frequency spectrum can be represented with a histogram (Figure 8).

Derived allele frequency	Number of sites
1	3
2	4
3	1
4	0
5	1

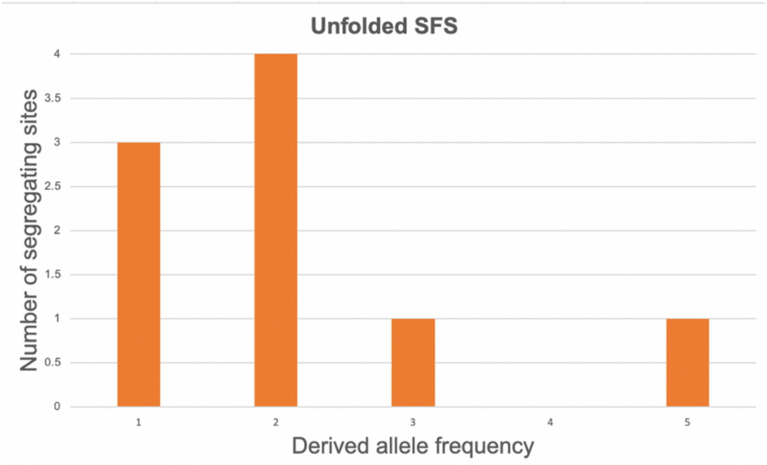
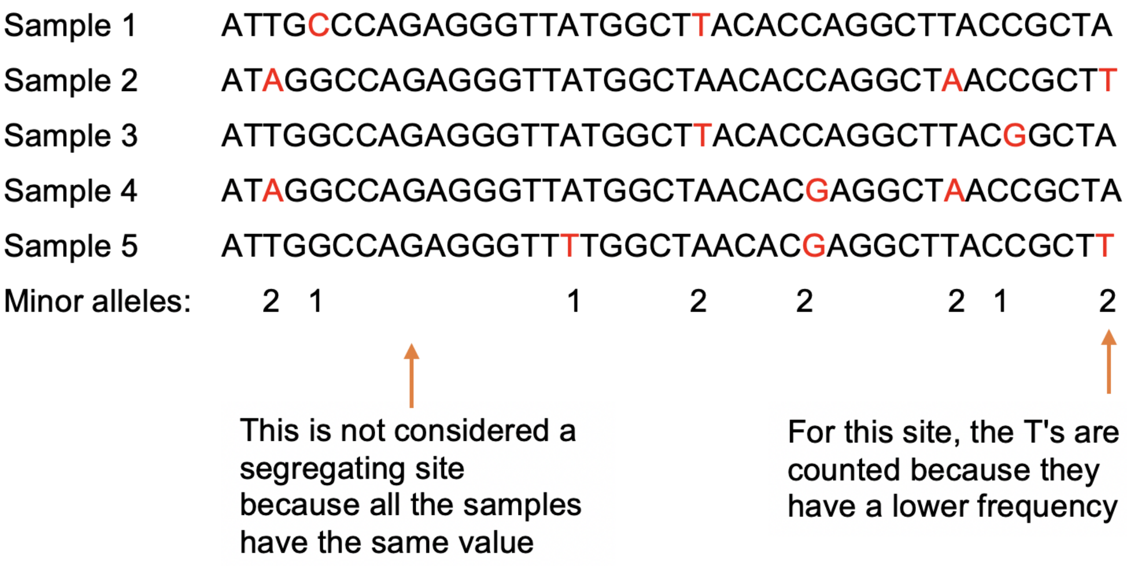


Figure 8.- Unfolded Site Frequency spectrum.
This unfolded sfs summarizes the segregating (different) sites compared to the ancestral state provided for the samples presented above. The number of derived alleles counts how many samples have a different nucleotide base (represented by letters A, C, G and T) to each position in the ancestral state. The Unfolded sfs graph shows the number of times the derived alleles have a count of 1, 2, 3, 4 or 5. For example, the first orange bar shows that there are 3 sites where the number of derived alleles is 1.

If the ancestral state is unknown, using the same samples from above, the site frequency spectrum can count the minor allele frequencies for all segregating sites:



```
Sample 1        ATTGCCCAGAGGGTTATGGCTTACACCAGGCTTACCGCTA
Sample 2        ATAGGCCAGAGGGTTATGGCTAACACCAGGCTAACCGCTT
Sample 3        ATTGGCCAGAGGGTTATGGCTTACACCAGGCTTACGGCTA
Sample 4        ATAGGCCAGAGGGTTATGGCTAACACGAGGCTAACCGCTA
Sample 5        ATTGGCCAGAGGGTTTTGGCTAACACGAGGCTTACCGCTT
Minor alleles:    2 1          1     2    2     2  1   2
```

This is not considered a segregating site because all the samples have the same value

For this site, the T's are counted because they have a lower frequency

See Figure 9 for the graphic representation of the folded sfs for these samples.

Minor allele frequency	Number of sites
1	3
2	5
3	0
4	0
5	0

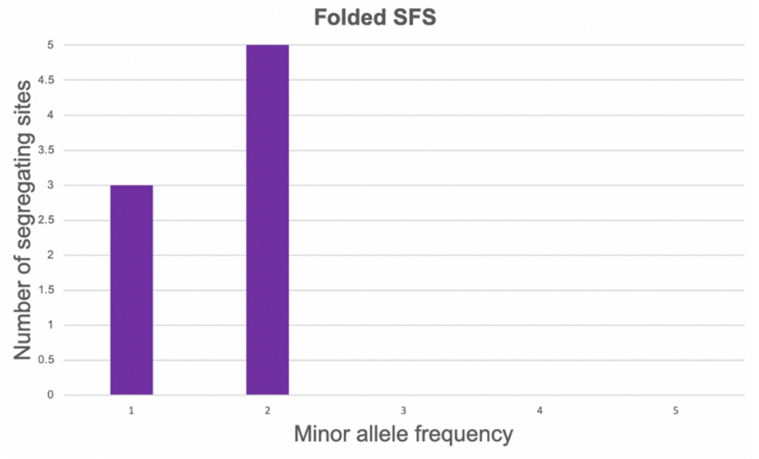


Figure 9.- Folded site frequency spectrum.
This folded sfs represents the number of segregating sites which have specific number count of minor (less frequent) alleles for the samples presented above. For the unfolded sfs, all the samples are compared between them. From the sites that have differences, the alleles that have less quantities are the ones counted. In the graph, the first purple bar shows that there are 3 sites where the number of minor alleles is 1.

Joint site frequency spectrum

The previous examples showed the allele frequencies for samples from the same population. However, to compare more populations, a joint frequency spectrum is needed. The 2D joint frequency spectrum represents the proportions of derived (unfolded) or minor (folded) alleles between populations. For two populations, it can be represented by a heat diagram. For more populations, it can be challenging to visualize it.

To build an unfolded joint frequency spectrum, two or more samples are compared to their ancestral state. The number of derived alleles is counted for each population (Figure 10).

```
Population 1 samples:

Ancestral state ATTGGCCACAGGGTTATGGCTAACACCAGGCTTACCGCTT
Sample 1        ATTGCCCAGAGGGTTATGGCTTACACCAGGCTTACCGCTA
Sample 2        ATAGGCCAGAGGGTTATGGCTAACACCAGGCTAACCGCTT
Sample 3        ATTGGCCAGAGGGTTATGGCTTACACCAGGCTTACGGCTA
Sample 4        ATAGGCCAGAGGGTTATGGCTAACACGAGGCTAACCGCTA
Sample 5        ATTGGCCAGAGGGTTTTGGCTAACACGAGGCTTACCGCTT
Derived alleles:  2 1   5   0  1     2    2     2  1  03

Population 2 samples:

Ancestral state ATTGGCCACAGGGTTATGGCTAACACCAGGCTTACCGCTT
Sample 1        ATTGGCCAGAGGGTTATGGCTTACACCACGCTTACGGCTA
Sample 2        ATTGGCCACAGGGTTTTGGCTTACACCAGGCTAACCGCTT
Sample 3        ATAGGCCAGAGGGTTATGGCTTACACCAGGCTAACGGCTT
Sample 4        ATTGGCCACAGCGTTATGGCTTACACGAGGCTAACCGCTT
Sample 5        ATTGGCCAGAGGGTTTTGGCTTACACGAGGCTTACCGGTT
Derived alleles:  1 0   3  1   2     5    1     3  2 1 1
```

Figure 10.- Comparison of samples from two different populations to an ancestral state.
Each of the samples is compared to the ancestral state to determine the number of derived alleles for each segregating site. This process is the same as the one shown previously for the unfolded sfs for one population. The difference is that the results will be compared between populations.

Table 3 summarizes the number of segregating sites according to frequencies between the two populations. The most common representation for two populations is a heat map instead of a table (Figure 11 and 12). Heat maps provide visual representations between different demographic scenarios (Figure 13).

Table 3.- Comparison of derived allele frequencies between populations.
This table shows the segregating site number with a certain derived allele count. For example, the first row shows the number of segregating sites in which population 2 has 0 derived alleles and the number of segregating sites in which population 1 has 0, 1, 2, 3, 4, or 5 derived alleles respectively.

		Population 2					
		None (0)	Singletons (1)	Doubletons (2)	Tripletons (3)	4	5
Population 1	None (0)	0	2	0	0	0	0
	Singletons (1)	1	0	2	0	0	0
	Doubletons (2)	0	2	0	1	0	1
	Tripletons (3)	0	1	0	0	0	0
	4	0	0	0	0	0	0
	5	0	0	0	1	0	0

Joint site frequency spectrum for populations 1 and 2 samples

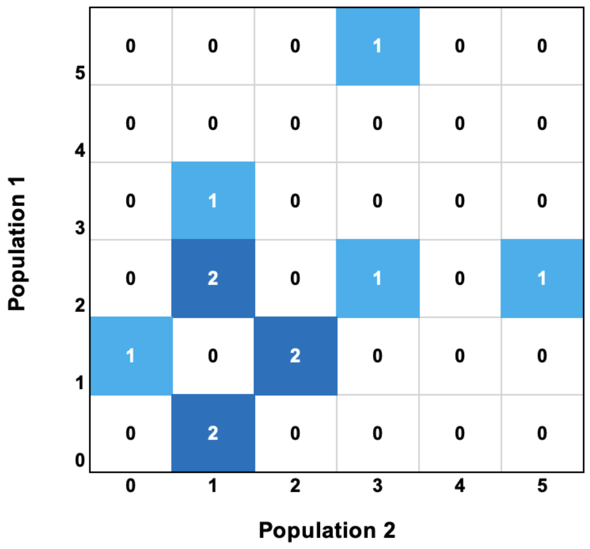


Figure 11. - Joint site frequency spectrum for two populations from example.

This heatmap shows the unfolded joint frequency spectrum from the samples presented above for Population 1 and 2. Each square shows the number of sites in which Population 1 has a specific number of derived alleles compared to Population 2 and vice versa. The axes show the number of alleles for the specified population. The heatmap represents numbers with different colors to visually identify them easily

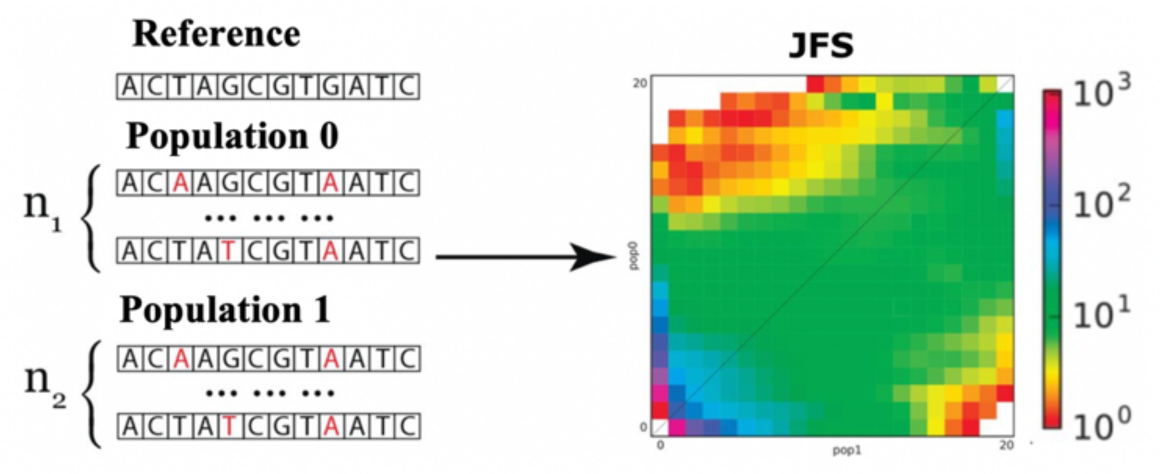


Figure 12.- Joint frequency spectrum for two populations.
Usually, joint site frequency spectra are more complex than the one shown from the previous example. This heatmap is an example of a more common situation when the samples are much larger (Figure taken from Noskova, Ulyantsev, Koepfli, O'brien, & Dobrynin, 2020). There are multiple colors in the diagram and must have a scale of values to interpret them. The values in the heatmap represent the number of derived alleles.

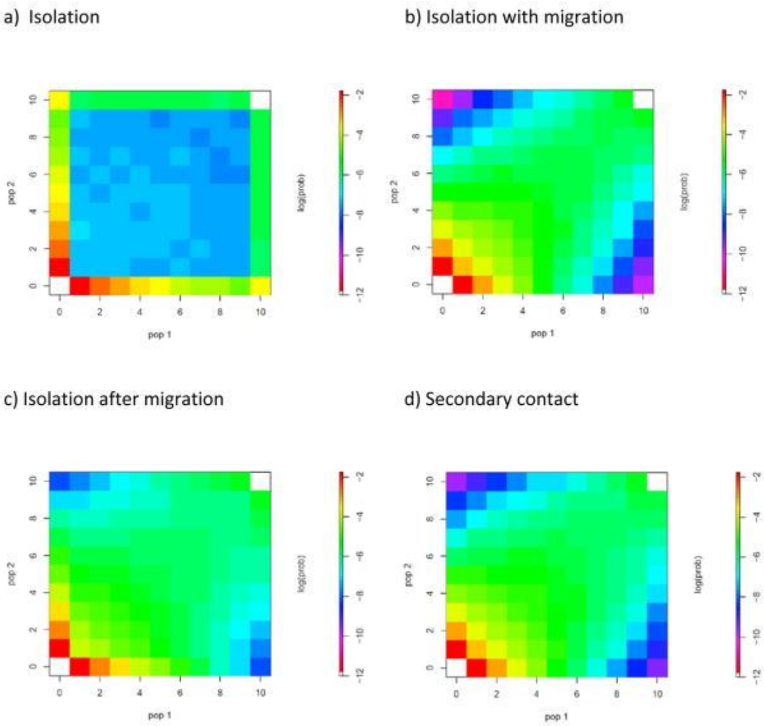


Figure 13.- Expected joint site frequency spectra for different evolutionary scenarios.
(Figure taken from Sousa & Hey, 2013). These jsfs have a different distribution according to the migration patterns between two populations. In a), the evolution between two populations occurred with them being isolated from each other. Therefore, there is little or none geneflow between them. The other scenarios, have more geneflow.

All representations of the site frequency spectrum (folded site frequency spectrum, unfolded site frequency spectrum, joint unfolded site frequency spectrum and joint folded site frequency spectrum) can be used to infer past population history as in Excoffier et al (2013). The main idea of the methodologies that use the site frequency spectrum to infer past population history is to find the past history that produces genetic data with a site frequency spectrum similar to what is seen in the analyzed data. In my thesis I will use a folded site frequency spectrum between two populations (an ancient population and a modern population), which creates a matrix stating the number of positions with i copies of the less frequent allele in the ancient population and j copies of the less frequent allele in a modern population.

Hardy-Weinberg Equilibrium

The Hardy-Weinberg equilibrium model is based on the assumption that individuals mate with each other randomly in a population. The outcome is that the genetic variation will remain constant across generations. If the allele frequencies are equal among the population and there are only two alleles possible in a site, then the probability of having a specific genotype is the same as having a random independent experiment, such as flipping two equal coins.

If there were only two possible alleles, A and a for a specific locus, the possible genotypes would be *AA*, *Aa* and *aa*. If the probability of having an allele A is denoted by $p(A) = p$ and for allele a by $p(a) = q$. Considering all the possible genotypes then:

$$p^2 + 2pq + q^2 = 1$$

There are several factors that disrupt genetic equilibrium. Some of the causes include inbreeding, natural selection, mutations, genetic flow between populations, non-random mating, and population structure (Mayo, 2008). The Hardy-Weinberg equilibrium is an ideal model, frequently used as a starting point for other models. Disruptions in the Hardy-Weinberg equilibrium changes the site frequency spectrum which, in turn, changes the estimates of past population history.

Divergence time

Divergence is when a population splits into two or more populations (Nielsen & Slatkin, 2000). The original population is referred to as ancestral. One main parameter of interest in a divergence model between two populations is the divergence time. The divergence time is the estimated time from the split to the present, and is commonly expressed in number of generations, rather than years (Edwards & Beerli, 2000).

The divergence time is useful for understanding the origins of populations over the world and the evolution of humans. The study of Neanderthal fossils in Sima de los Huesos site in Spain, concluded that Neanderthals and modern humans had a much larger divergence time than previously thought. The new estimated time is between 550,000 and 765,000 years ago (Gómez-Robles, 2019). Previous estimations determined the divergence time to be between 260,000 years ago to 350,000 years ago (Schlebusch et al., 2017).

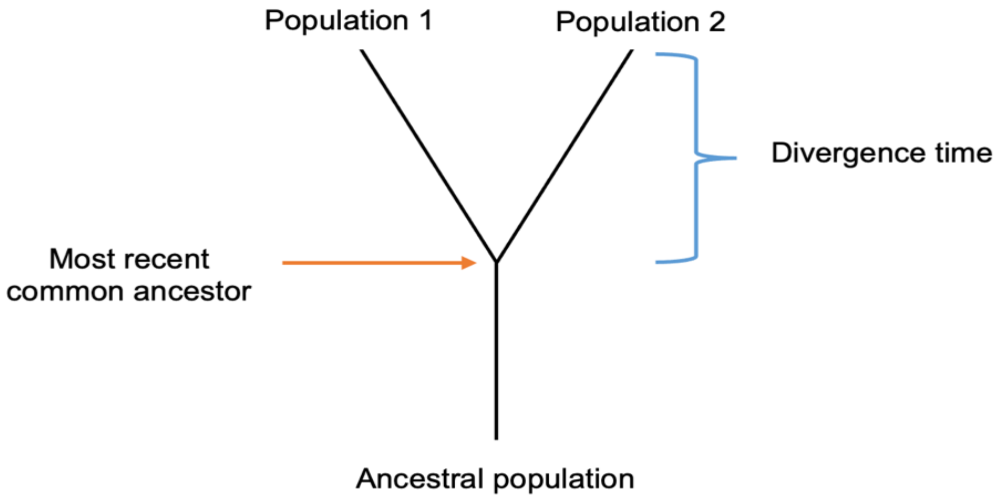


Figure 14.- Divergence of an ancestral population split into two populations

Review of statistical analysis methods

Research based on DNA samples has several steps such as extraction, sequencing, pre-processing, and analysis. There have been considerable advances for each of these stages. This review focuses on the analysis based on samples that have been either already sequenced or simulated using software programs.

Genotype callers based on Bayesian probability models

After a DNA sample is sequenced, the base of each site is called with a quality score based on the sequencing error rate. Then, the sequence is aligned to a reference genome and recalibrated. Two important processes take place afterwards. The first one is SNP calling, the process of determining which sites have a different allele from the reference sequence. The other one is genotype calling, the establishment of a genotype. Human beings are diploid, meaning that cells have paired chromosomes, one from each parent. Genotypes are the two alleles contained at a specific genomic site.

Genotype calling for ancient DNA on raw data is difficult because most have low coverages. Older genotype callers require coverages higher than 20· and very high-quality scores (Nielsen, Paul, Albrechtsen, & Song, 2011), making them unsuitable for ancient sample analysis. Another theoretical framework to perform genotype calling is to use genotype likelihoods. Genotype likelihoods use the Bayes' Theorem to determine which genotype has the highest probability at a site. The genotype likelihood can be represented as $p(X_i|G)$, where X_i are the alleles read at a site for all individuals and G is a genotype. The highest probability $p(X_i|G)$ will determine the more probable genotype at that position.

BCFtools is one of the most used genotype callers based on genotype likelihoods. BCFtools is part of the SAMtools package (H. Li et al., 2009), used in this research project for data alignment. BCFtools calculates genotype likelihoods considering the possible alleles in each position, qualities of the reads and base calling, and errors during alignments. The package determines the genotype based on the genome with the highest probability assuming Hardy-Weinberg equilibrium (Danecek et al., 2021). Other software packages that use genotype likelihoods for genotype calling that are commonly used are SOAP2 (R. Li et al., 2009) as well as the Genome Analysis Toolkit (GATK) (McKenna et al., 2010) and. Correct genotype calling on ancient samples is challenging due to their low coverage (Figure 4) but is always used with modern samples. In ancient samples it is common to only sample one allele from each position in the genome based on one read and use that information to perform further analysis. In this work we will assume that the ancient DNA information is perfectly known and see if the inferences are perfectly known instead of relying on information from one read in each position of the genome which could give imperfect inferences.

Methods for structure analysis

Population structure occurs when individuals located in places where a geographical division exists, such as remote islands or mountains. Individuals located in these areas are more likely to mate, creating subgroups with different levels of genetic similarity within the main population. (Novembre & Ramachandran, 2011). These subgroups can be used to define the populations that will be used for demographic analysis and inferences of parameters such as the divergence time. There are two main approaches to analyze population structure, Principal component analysis (PCA) and Admixture proportion inference.

Structure (Pritchard, Stephens, & Donnelly, 2000) is a software based on the Pritchard-Stephens-Donelly model for Admixture proportion inference that uses a Bayesian clustering method to infer the proportions of the genomic data that belong to K predefined ancestral clusters. A cluster represents a population, and each population has a set of allele frequencies by locus. Individuals are given a percentage that represents how much of their genome comes from each of the populations (Figure 15). The genotypes are observed from the individual's data while the source population and the population allele frequencies are inferred from a probabilistic distribution. The exact probability distribution is difficult to obtain, therefore, approximations of samples are found using Markov chain Monte Carlo (MCMC) methods (Pritchard et al., 2000).

FRAPPE and ADMIXTURE are other approaches for admixture proportion inferences are FRAPPE and Admixture (Tang, Peng, Wang, & Risch, 2005; Alexander, Novembre, & Lange, 2009). Both software programs calculate population allele frequencies and proportions from K populations. Instead of sampling using MCMC method they use maximum likelihood, speeding the processing times. FRAPPE uses an Expectation-Maximization (EM) algorithm (Tang et al., 2005). ADMIXTURE implemented a block relaxation algorithm and a quasi-Newton acceleration of convergence to yield accurate results in less time than structure (Alexander et al., 2009). Figure 16 shows an example of an ADMIXTURE run for 938 unrelated individuals form different regions and K = 6 populations.

A popular method to analyze population structure is principal component analysis (PCA). This method reduces the genomic information down to a set of meaningful components. The genetic information of a set of individuals, which is represented as a large matrix where the rows are the genomic information of each individual and the columns are genetic markers such as SNPs. The reduction of data is done by calculating the eigenvectors and eigenvalues of the covariances of the columns and then finding the most significant components. The result is the generation of coordinates for visual representation (Novembre & Ramachandran, 2011). Populations with lower levels of admixture are usually clustered in a specific location. Admixed individuals are located in areas between the clusters. PCA was first introduced by Menozzi et al. (1978). Subsequently, SMARTPCA, a software package based on the Cavalli approach for fast and precise results was developed by Patterson et al. (2016) (Figure 17).

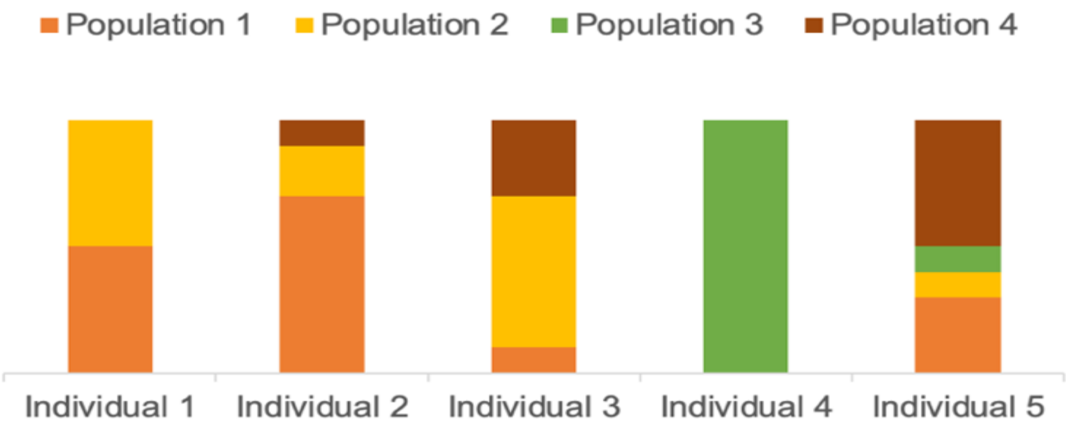


Figure 15.- Admixture proportions in sampled individuals.

The bars show proportions of the individual's genomic data associated with clusters representing populations. The results are useful to determine if there is substantial evidence of structure in the samples. For each individual, there is a vector of length K describing the proportions of their genetic ancestry. Given the observed genetic information, a matrix with these vectors is inferred using software programs with admixture proportion inference models

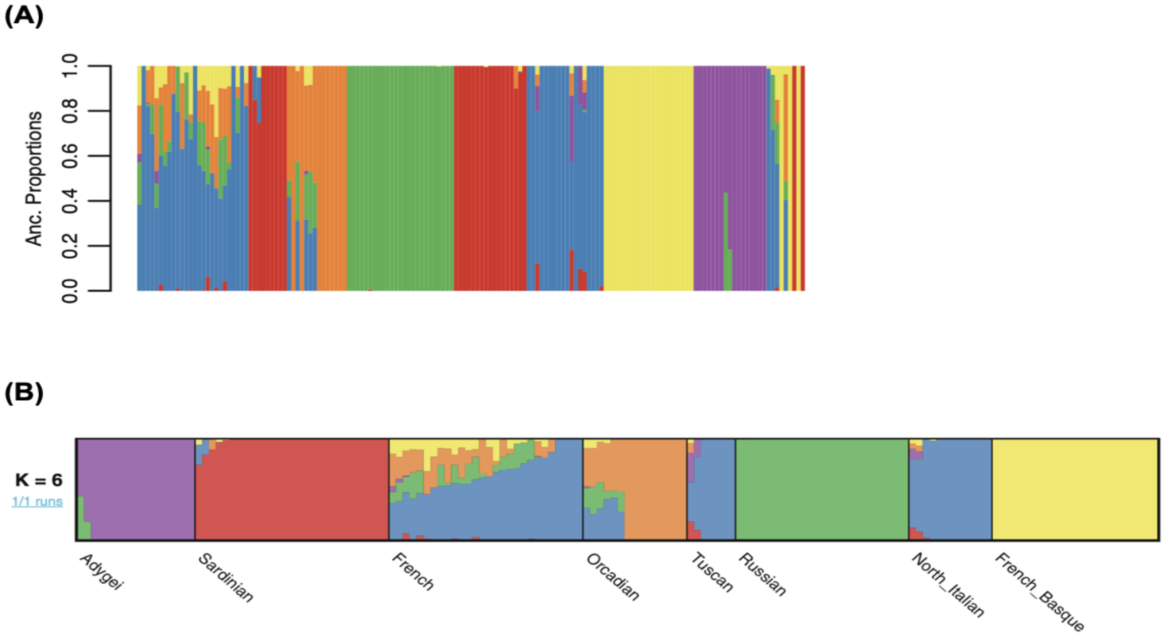


Figure 16.- Example of admixture proportions using ADMIXTURE (Figure taken from Liu, Shringarpure, Lange, & Novembre, 2020).

(A) Admixture proportions obtained from ADMIXTURE software run using 938 samples from unrelated individuals. Each bar represents an individual and the admixture proportion from 6 populations. (B) This diagram is a plot from the same run form (A) using pong (Figure taken from Behr, Liu, Liu-Fang, Nakka, & Ramachandran, 2016), a visualization tool. The samples are organized by region and by level of admixture.

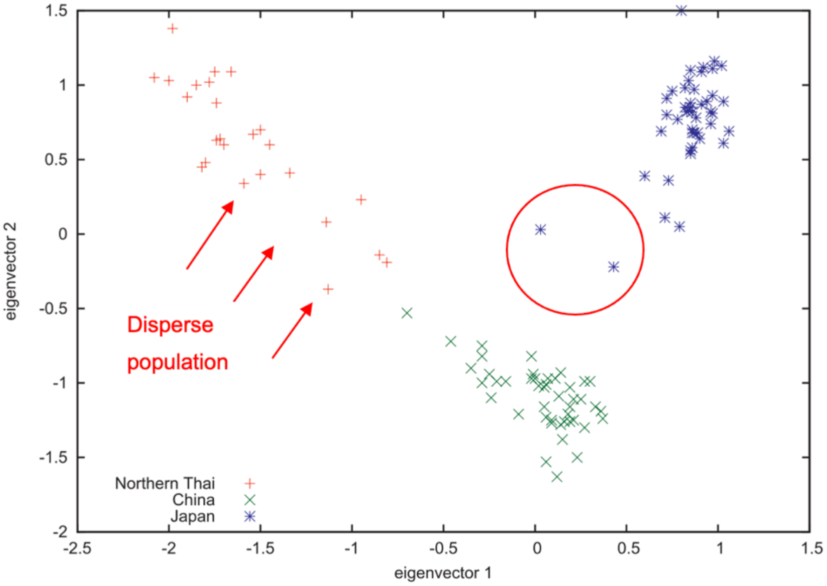


Figure 17.- SmartPCA plot from three East Asian Populations (Figure taken from Patterson et al., 2006).
This is an example of a plot from the two first eigenvectors for population samples from Thailand, China, and Japan. The circled area shows some admixture from Japanese samples. Samples from Thailand are more dispersed as they present more gene flow with the Chinese population. Japanese and Chinese samples are clustered a lot closer.

Useful statistics to analyze past demographic history

Apart from the site frequency spectrum, there are other statistics that can be used to perform demographic inferences that I will mention here. The D-statistic is one of those statistic and can be used to develop a method to determine if there is admixture between individuals from different species. To calculate if samples from different populations share gene flow, it is necessary to have 4 samples from populations with a tree-like relationship (Green et al., 2010). Two of the samples (H1 and H2) are from populations from the same species. The third (H3) and fourth (H4) population samples must have a more distant relationship with H1 and H2 (Figure 18A). Green et al. (2010) introduced D-statistics to determine if there was admixture between Neandertals and present-day individuals from different geographic areas (Figure 18B).

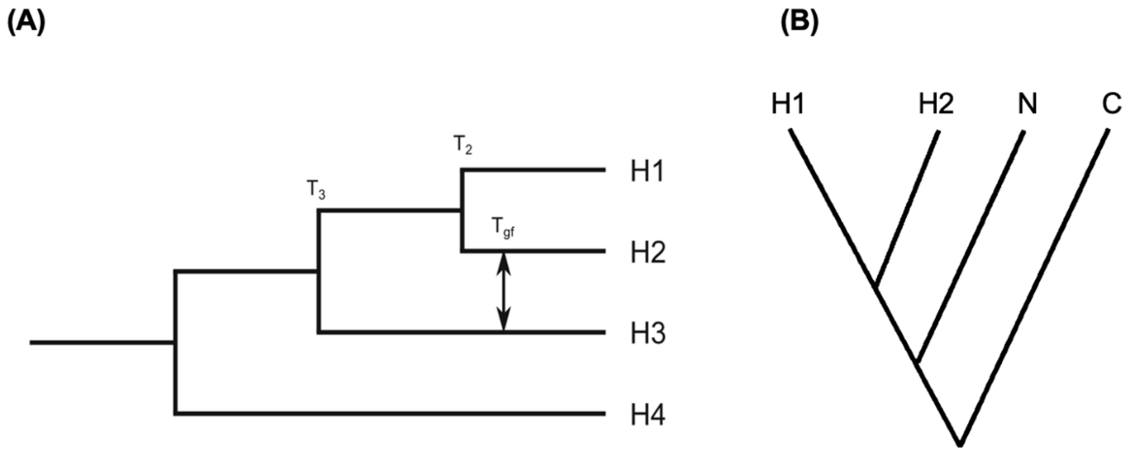


Figure 18.- Samples in a population tree for D-statistics.
(A) H1, H2, H3 and H4 are samples with different levels of relatedness and are used to calculate D-statistic. D can detect geneflow between H1 and H3 or H2 and H3 (depicted in the diagram). H4 is an outgroup, or a species very far related with the rest of the samples. T_2, T_3 and T_{gf} is the time passed since these species separated (Figure taken from Zheng & Janke, 2018). (B) (Figure taken from Slatkin, 2016) This population tree represents the model that Green et al. (2010) used to prove admixture between Neandertals and humans. N and C were samples from Neanderthal and Chimpanzee while H1 and H2 were samples from modern-day populations from different geographic areas. The analysis included comparisons between populations from Asia, Europe, and Africa.

Using samples with the structure shown in Figure 18A, we calculate $D(H1, H2, H3, H4)$ to determine if there is gene flow between H1 and H3 or H2 and H3. For diploid individuals (such as humans and hominins), from the two alleles at each site, one is randomly chosen and later used for counts. H4 is the outgroup population and determines the ancestral allele. Equation 2 shows the formula to calculate *D*. For example, to determine if there was admixture between H2 and H3, N_{ABBA} would be site number where H1 and H4 have an ancestral allele and H2 and H3 have a derived allele. Therefore, N_{BABA} would be the site number where H2 and H4 have an ancestral allele and H1 and H3 have a derived allele. If there had been no admixture between H3 and both H1 and H2, then $D(H1, H2, H3, H4)$ is expected to be zero. If there was admixture between H3 and H2, then *D* would be positive (Slatkin, 2016).

$$D(H1, H2, H3, H4) = \frac{N_{ABBA}(H1, H2, H3, H4) - N_{BABA}(H1, H2, H3, H4)}{N_{ABBA}(H1, H2, H3, H4) + N_{BABA}(H1, H2, H3, H4)}$$

F-statistics are methods used to quantify genetic drift in a population phylogeny (population tree). Reich et al introduced these measures to study Indian populations and their relationship to two local ancient populations. Patterson et all summarized f-statistics and introduced ADMIXTOOLS, a software package that incorporates these and other methods for admixture analysis.

There are four f-statics, f_2, f_3 and f_4. f_2 measures the time separating two populations. It is defined as the average cross loci of *(a - b)* where a and *b* represent allele frequencies on populations *A* and *B* (Slatkin, 2016). The expected value of f_2 *(A, B)* can be calculated as:

$$F_2(A, B) = E[(a-b)(a-b)]$$

$$F_2(A, B) = E[(a-b)^2]$$

f_3 is a three-population test that can provide unambiguous evidence of admixture and quantify genetic drift using an outgroup (Orlando et al., 2021). For the SNP allele frequencies of the A, B and C populations, calculating the average of *(c - a) (c - b)* over SNPs, would result in a negative value if population C is admixed from A and B (Reich, Thangaraj, Patterson, Price, & Singh, 2009).

The expected value of f_3 (C; A, B) to test admixture of C from A and B would be:

$$F_3(C; A, B) = E[(c-a)(c-b)]$$

f_4 is a four-population method that measures the length of the branch that connects two pairs of populations. In Figure 19, F_4 is the estimation of the length of the red branch connecting populations A and B with C and D.

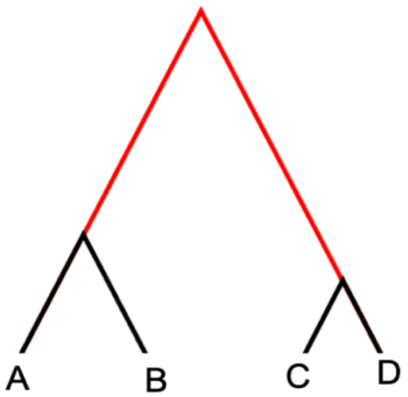


Figure 19.- Population tree to illustrate branch (Figure taken from Peter, 2016).
f_4 represents the red branch that connects populations A and B with C and D. The branches on each side don't need to be symmetrical, as depicted on the figure.

The estimated value of f_4 is the average of the SNP allele differences:

$$F_4(A,B;C,D) \;=\; E[(a-b)(c-d)]$$

f_4 is also used to calculate the ancestry proportion for admixed populations. Figure 20 also includes visual examples and calculations of F_2, F_3, and F_4 for the phylogenies shown.

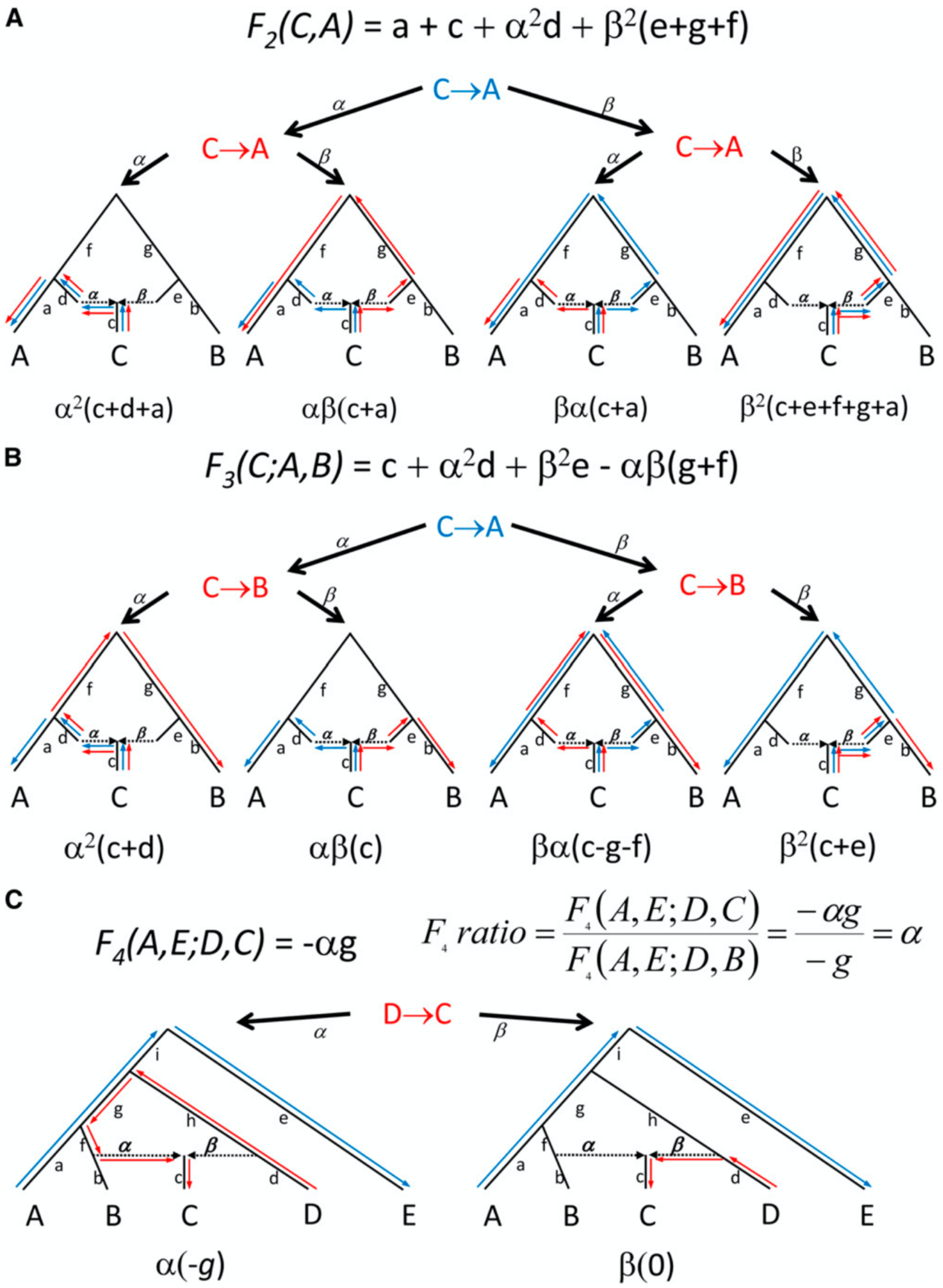

A
$F_2(C,A) = a + c + \alpha^2 d + \beta^2(e+g+f)$
C→A
α
β
C→A
C→A
a
b
c
d
e
f
g
A
C
B
$\alpha^2(c+d+a)$
$\alpha\beta(c+a)$
$\beta\alpha(c+a)$
$\beta^2(c+e+f+g+a)$
B
$F_3(C;A,B) = c + \alpha^2 d + \beta^2 e - \alpha\beta(g+f)$
C→A
C→B
C→B
$\alpha^2(c+d)$
$\alpha\beta(c)$
$\beta\alpha(c-g-f)$
$\beta^2(c+e)$
C
$F_4(A,E;D,C) = -\alpha g$
$F_4 \; ratio = \frac{F_4(A,E;D,C)}{F_4(A,E;D,B)} = \frac{-\alpha g}{-g} = \alpha$
D→C
h
i
D
E
$\alpha(-g)$
$\beta(0)$

Figure 20.- Calculations of *f*-statistics (Figure taken from Patterson et al., 2012).
Each of the examples shows the procedures to calculate f2, f3, f4, and α for the given population trees. The procedures take all the possible routes to get from one point to the other.

Another statistic commonly used to analyze past population demography is Wright's $F_{ST.}$ In nature it is common to have some population structure because individuals that live close by have a higher probability to mate (Nielsen & Slatkin, 2013). Wright's F_{ST} is the most common measure for quantifying population differentiation within and among populations due to genetic structure. F_{ST} represents the allele correlation within a subpopulation compared to the whole sample (Holsinger & Weir, 2009). A F_{ST} value < 0.05 means that there is negligible or no differentiation between populations. In contrast, F_{ST} > 0.25 implies very great differentiation (Wright, 1951). A drawback of calculating F_{ST} values is that the definition in Wright's paper is ambiguous. F_{ST} is defined as the allele correlation of from one population compared to the whole sample. However, different sources interpreted the "total population" in various ways (Cockerham, 1969; Nei, 1973), leading to confusion about the correct estimation of F_{ST}. Bhatia, Patterson, Sankararaman, & Price (2013) clarified how to calculate F_{ST} and gave some guidelines.

Simulation and inference methods

Researchers working with ancient DNA face several challenges, as described in the introduction. Simulations provide an important aid for research on theories and techniques involving genetics. Simulation tools can simulate large number of DNA samples and countless demographic scenarios. Kelleher and Lohse (2020) highlight the importance of comparing observed and simulation results and describe it "as an important sanity check for both".

This project relies on simulation software heavily as it is a necessary step to prove the utility of the method before using real data. This section describes three software tools: msprime, ANGSD and fastsimcoal2. msprime and fastsimcoal2 were used for simulations and inferences in this project. The method section has detailed information of their use.

The coalescent model was introduced by Kingman (Kingman, 1982) to describe the genealogical history of a sample. It is commonly used to simulate different population scenarios for data analysis and population parameters inferences (Rosenberg & Nordborg, 2002). The coalescent process consists in determining the closest ancestor of a group of samples in previous generations. When two lineages meet, it is called a coalescent event. If a sample has *n* individuals, then there are *n* - 1 coalescent events. All lineages meet eventually, finding the most common ancestor (MRCA) for the individuals in the sample (Wakeley, 2009) (Figure 21).

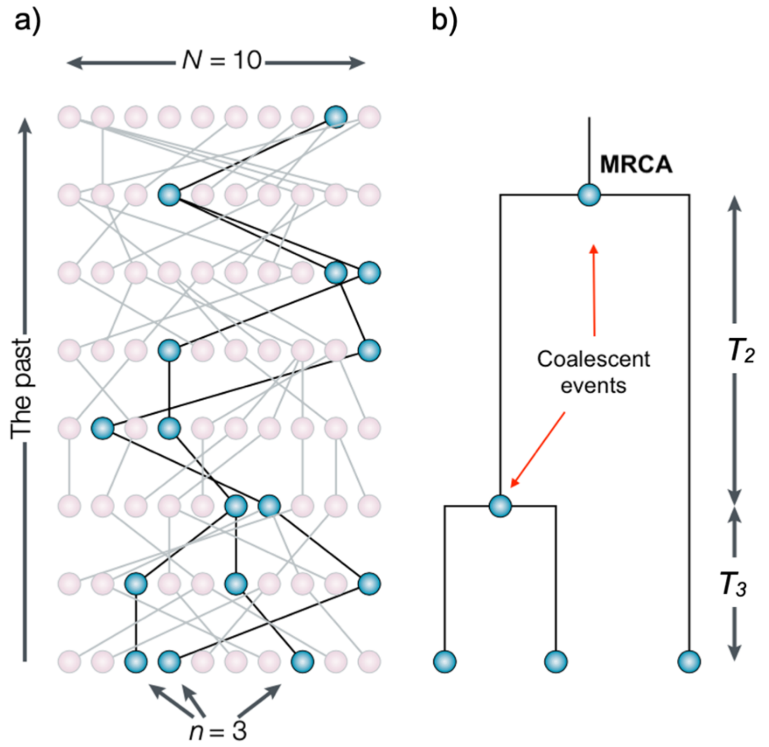


Figure 21.- Principle of the coalescent model (Figure taken from Rosenberg & Nordborg, 2002).
a) Genealogy for a population of ten individuals. Highlighted in black is the sample genealogy of n = 3, in each past generation the ancestors are depicted with a blue circle. In this case, all the samples coalesce 7 generations ago. b) Coalescent genealogies are commonly graphed as a simple tree showing only the coalescent events. The two upper branches meet at the MRCA defined as the most recent common ancestor. Here T2 and T3 are times between coalescent events.

msprime is a Python library that simulates genealogies using the coalescent model. It can simulate large quantities of data using different demographic scenarios. The main advantages are high customization level, very efficient processing, and the integration with other software tools for genomic analysis. msprime simulates the ancestral history as coalescent trees (Figure 22). The results of these simulations are expressed in a data structure called succinct tree sequence (Kelleher & Lohse, 2020). Succinct tree sequences provide full ancestral histories with compact data storage (Kelleher, Etheridge, & McVean, 2016).

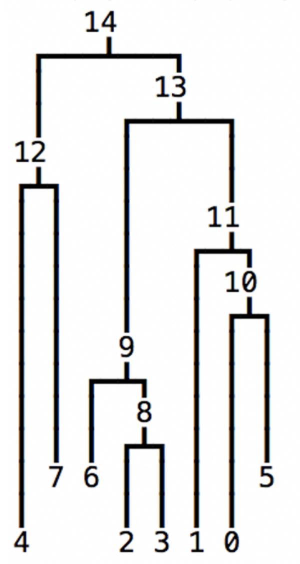


Figure 22.- Simulation of coalescent tree using msprime.
This coalescent tree shows branches of different lengths, depicting their generation. The samples at the lowest level (0-4) are present-day, while samples 5-7 are from a previous generation. Numbers 8 - 14 are coalescent events. Number 14 represents the most recent common ancestor.

ANGSD stands for Analysis of Next Generation Sequencing Data, a software tool for population genetic data analysis. It can make error estimates, calculate different summary statistics, D-statistics, genotype likelihoods and call SNPs and genotypes among other population estimators. An important feature of ANGSD is the capacity to estimate the SFS (site frequency spectrum) and the joint site frequency spectra for two or more populations (Korneliussen, Albrechtsen, & Nielsen, 2014).

A two-population joint site frequency spectrum (2D-SFS) contains the site number with specific counts of derived alleles in both populations. The frequencies are the elements of a matrix $(2n_1 + 1) \times (2n_2 + 1)$, where n_1 and n_2 are samples of individuals from population 1 and 2 respectively. The matrix structure can be expressed as:

$$\gamma = \begin{pmatrix} \gamma_{00} & \gamma_{01} & \cdots & \gamma_{02n_2} & \gamma_{02n_2+1} \\ \gamma_{10} & \gamma_{11} & \cdots & \gamma_{12n_2} & \gamma_{12n_2+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma_{2n_10} & \gamma_{2n_11} & \cdots & \gamma_{2n_12n_2} & \gamma_{2n_12n_2+1} \\ \gamma_{2n_1+10} & \gamma_{2n_1+11} & \cdots & \gamma_{2n_1+12n_2} & \gamma_{2n_1+12n_2+1} \end{pmatrix}$$

For example, if $\gamma_{10} = 3$, then there are three sites where population 1 samples have 1 derived allele and population 2 has zero.

The calculation of a 2D-SFS based on low or medium coverage NGS data is likely to be biased, based on analysis of a 1D-SFS (Han, Sinsheimer, & Novembre, 2014). Instead, ANGSD uses a maximum likelihood approach to infer the SFS which is a matrix γ. The likelihood for a site *s* given the SFS γ is:

$$L(X \mid \gamma) = \prod_{s=0}^{N} L(X_s \mid \gamma) = \prod_{s=0}^{N} \sum_{i=0}^{2n_1} \sum_{j=0}^{2n_2} \gamma_{ij}\, p(X_s^1 | D^1 = i)\, p(X_s^2 | D^2 = j)$$

Where $p(X_s^1 | D^1 = i)$ is the likelihood of having the sequencing data X_s^1 in the site s in population 1 given that there are *i* derived alleles in population 1 (Korneliussen, Moltke, Albrechtsen, & Nielsen, 2013). The maximum likelihood value of $L(X | \gamma)$ is found via an expectation-maximization algorithm (Korneliussen et al., 2014). In ancient samples with a low coverage, the use of ANGSD is preferred to correctly estimate the site frequency spectrum. In this project, we will assume that the site frequency spectrum is estimated perfectly, and we will focus the work on the power of this statistic to estimate past demographic history.

fastsimcoal2 is a software program that uses a site frequency spectrum (SFS), to estimate demographic parameters of a given demographic model. fastsimcoal2 also uses the coalescent model and can make estimations for various demographic scenarios. Given an SFS, it estimates the expected SFS using a composite likelihood method. Demographic parameters are inferred from the estimated SFS (Excoffier et al., 2021). For a single population with a sample size of n and a SFS equal to $X = \{m_1, \cdots, m_{n-1}\}$, the composite likelihood of a given model with its set of parameters θ is:

$$CL = Pr(X|\theta) \propto {P_0}^{L-S}(1 - P_0)^S \prod_{i=1}^{n-1} {\hat{p}_i}^{m_i}$$

X is the data

S is the polymorphic site number,

L is the number of positions analyzed

P_0 is the probability of a site displaying no variation.

$\hat{p}_i$ is the estimated probability of a given derived allele frequency *i*

The composite likelihood can be extended for more populations. For a joint SFS for two populations, as used in this work, the estimation is done using the following:

$$CL_{12} \propto {P_0}^{L-S}(1 - P_0)^S \prod_{i=0}^{n_1} \prod_{j=0}^{n_2} {\hat{p}_{ij}}^{m_{ij}}$$

Both $\hat{p}_i$ and ${\hat{p}_{ij}}^{m_{ij}}$ are calculated via a large number of simulations (100,000 in this work) under a particular demographic model that contains a certain set of parameters θ. fastsimcoal2 uses a greedy algorithm to maximize its likelihood. The algorithm starts at a random set of values for the parameters and tries to optimize the values of the parameters given the starting set of values for the parameters. The starting set of values for the parameters determines the local optima set of values for the parameters. In practice, you must start at many different points and then keep the parameter values where you reached a higher likelihood value. In my work I started from 100 different parameter values and picked the parameter values where the likelihood had a highest value as my estimate for the parameter. In each run, the parameter values are optimized using a conditional maximization algorithm where we used 40 cycles of optimization as recommended by the authors (Excoffier et al., 2013).

Hypothesis

The joint site frequency spectrum (jsfs) is a statistic that provides useful information to infer demographic parameters. This statistic will be sufficient to infer the divergence time between populations, an important demographic parameter to understand the relationship between different populations.

A two-population (2D) joint site frequency spectrum from ancient and modern genomic samples from the same region will allow us to infer an unbiased divergence time between those populations.

Objectives

- To analyze if the joint site frequency spectrum contains sufficient information to make inferences of demographic history, particularly of the divergence time between populations.

.

Specific objectives

- To determine if the number of samples plays a role in the correct inference of the divergence time.
- To compare the joint site frequency spectra obtained in simulations with different sample sizes and divergence times between populations .

Methods

Pipeline overview

We developed a Bash pipeline to analyze if the joint site frequency spectrum contains sufficient information to infer the divergence time, a demographic parameter that is of high interest in the studies of aDNA. Our bash pipeline calls several Python and Shell scripts. It also uses previously developed software that is commonly used for genomic data research. The main software tools used are msprime, Seq-Gen, and fastsimcoal2. msprime, developed by Kelleher & Lohse (Baumdicker et al., 2022; Kelleher & Lohse, 2020), is a very efficient coalescent simulator that uses the coalescent model to generate a set of coalescent trees that reflect the relationships between a set of genomes. msprime was used to simulate ancient and modern genomic samples. Then, Seq-Gen uses the coalescent trees to generate a genomic region with nucleotide bases for both the present day and ancient samples (Rambaut & Grassly, 1997). Then, we developed a python script to estimate the jsfs, which is a highly informative summary statistic that is applied to infer the past demographic history of many samples. Finally, we used fastsimcoal2 (Laurent Excoffier et al., 2021) to estimate the divergence time of two populations using the jsfs calculated before.

Computational Pipeline

The process consists of several steps that are represented in Figure 23 and are described in the following sections.

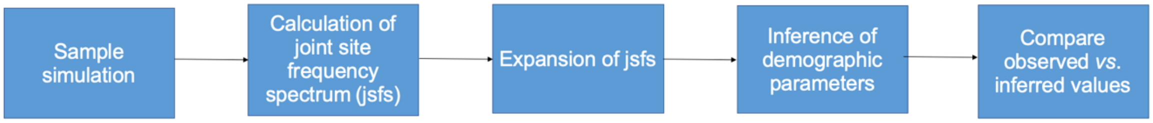

Sample simulation
Calculation of joint site frequency spectrum (jsfs)
Expansion of jsfs
Inference of demographic parameters
Compare observed *vs.* inferred values

Figure 23.- Methods stages
The diagram shows the main stages that represent the steps to implement the proposed method.

Sample simulation

The samples are generated using two python scripts. The first one is a user customizable script in which it is possible to create different demographic scenarios. Some of the input variables that are possible to modify are: effective population size, sample number, the sequence length, recombination rate, mutation rate, number of modern samples, number of ancient samples and time difference between the modern and ancient samples. The second python script calls the first one to use it as input for sample simulation.

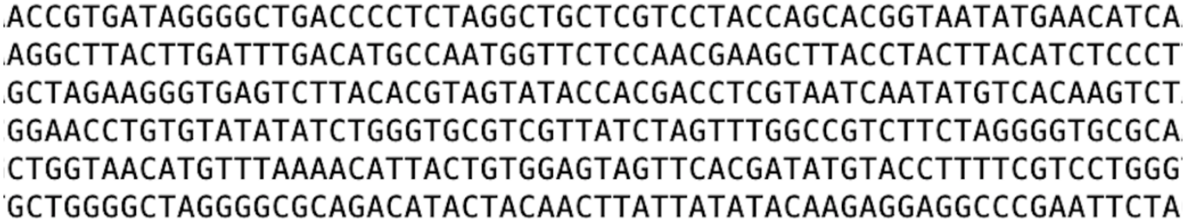
The main reasons to use simulated samples are, to tailor them for a specific scenario and to be able to know their precise contents. For this project, we simulated scenarios in which we varied two variables, time of divergence and number of ancient samples (Table 4).

The times of divergences are 0, 288 and 600 generations ago. The generation time is approximately 25 years. Time of divergence = 0 is when the ancient and modern samples are from the same population. The second case, Time of divergence = 288 generations ago is approximately 7,200 years of divergence between the population of the modern samples and the population of the ancient sample. This divergence time has previously been estimated as the split time between some northern and southern ancient populations from Mexico (Ávila-Arcos et al., 2020). The last time of divergence I used was 600, used for populations that split 15,000 years ago. This is an estimation of the divergence time between northern and southern North American populations (Moreno-Mayar, Vinner, et al., 2018). The ancient sample comes from 20 generations ago (500 years) in all the simulations performed.

The simulation script takes the parameter inputs and uses msprime to simulate coalescent trees that represent possible genealogies in a particular region based on the given values. After obtaining the trees, the Seq-Gen program converts the coalescent trees into sets of nucleotide sequences in a region and delivers them in FASTA format files (Rambaut & Grassly, 1997). FASTA format is a text file that contains letters representing the four possible bases that form DNA (Figure 24).

Table 4.- Simulation scenarios for different sample sizes and divergence time

Scenario	Time of divergence	Number of ancient diploid samples	Number of modern diploid samples	Outputs
0	0	5	30	10 ancient sequences 60 modern sequences
1	0	25	30	50 ancient sequences 60 modern sequences
2	0	50	30	100 ancient sequences 60 modern sequences
3	288	5	30	10 ancient sequences 60 modern sequences
4	288	25	30	50 ancient sequences 60 modern sequences
5	288	50	30	100 ancient sequences 60 modern sequences
6	600	5	30	10 ancient sequences 60 modern sequences
7	600	25	30	50 ancient sequences 60 modern sequences
8	600	50	30	100 ancient sequences 60 modern sequences



```
ACCGTGATAGGGGCTGACCCCTCTAGGCTGCTCGTCCTACCAGCACGGTAATATGAACATCA
AGGCTTACTTGATTTGACATGCCAATGGTTCTCCAACGAAGCTTACCTACTTACATCTCCCT
GCTAGAAGGGTGAGTCTTACACGTAGTATACCACGACCTCGTAATCAATATGTCACAAGTCT
GGAACCTGTGTATATATCTGGGTGCGTCGTTATCTAGTTTGGCCGTCTTCTAGGGGTGCGCA
CTGGTAACATGTTTAAAACATTACTGTGGAGTAGTTCACGATATGTACCTTTTCGTCCTGGG
GCTGGGGCTAGGGGCGCAGACATACTACAACTTATTATATACAAGAGGAGGCCCGAATTCTA
```

Figure 24.- Simulated DNS sequences in FASTA format
Seq-Gen simulates the genomic sequences from the coalescent tree from msprime. Seq-Gen provides a file that contains letters that simulate the four nucleotide bases that can be present in a genome.

Each of the simulated scenarios is run 100 times and each one is saved in a different file. The output of this step are sequences with 1,000,000 bases that represent the DNA samples of modern individuals and ancient individuals according to each scenario (Figure 25).

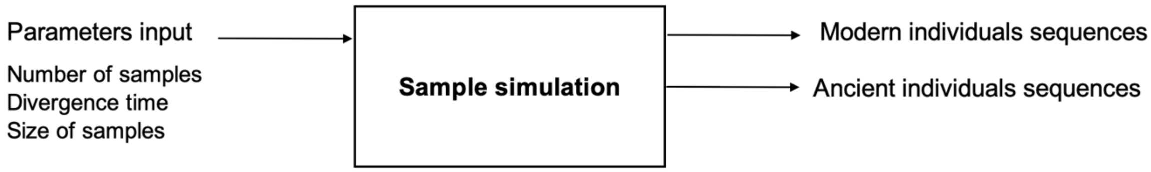


Figure 25.- Inputs/Outputs from sample simulation stage
The input for the sample simulation script are the demographic parameters for a specific scenario. These are included in a separate Python script, exclusively used for input selection. The outputs for this process are the simulated DNA sequences that will be used to calculate the joint site frequency spectrum.

Calculation of joint frequency spectrum (jsfs)

The folded jsfs is calculated using a python script that counts the minor allele frequencies in a population for all the sites that have different alleles, also known as segregating sites. Each segregating site can have one of two possible nucleotide bases, which are added across all the sequences of a population. In this case, we will consider the modern samples one population and the ancient samples another population. The base with a lower count in each site in each population is the minor allele for that population in that site.

The accumulated number of sites with minor alleles for the ancient and modern sequences is saved in a list. The list has the shape of a matrix m x n, where m is the chromosome number of the present-day samples and n is the chromosome number for the ancient samples. The matrix is the folded joint site frequency spectrum. The output from this stage (Figure 26) is a matrix that represents the folded joint site frequency spectrum from the given samples (Figure 27). This process is repeated for each scenario and each run.

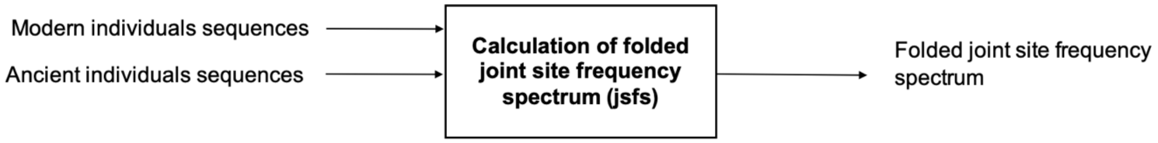
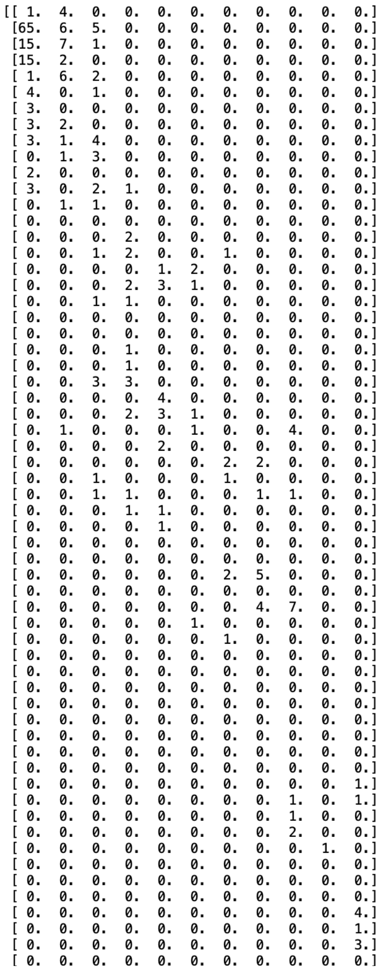

Modern individuals sequences
Ancient individuals sequences
Calculation of folded joint site frequency spectrum (jsfs)
Folded joint site frequency spectrum

Figure 26.- Inputs/Outputs from calculation of joint site frequency stage
The input for the sample simulation script are the simulated sequences representing ancient and present-day DNA samples. The output is the jsfs calculated from the sequences. The matrix is useful to provide a visual representation of the relationship between the samples provided.



```
[[ 1.  4.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [65.  6.  5.  0.  0.  0.  0.  0.  0.  0.  0.]
 [15.  7.  1.  0.  0.  0.  0.  0.  0.  0.  0.]
 [15.  2.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 1.  6.  2.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 4.  0.  1.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 3.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 3.  2.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 3.  1.  4.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  1.  3.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 2.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 3.  0.  2.  1.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  1.  1.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  2.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  1.  2.  0.  0.  1.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  1.  2.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  2.  3.  1.  0.  0.  0.  0.  0.]
 [ 0.  0.  1.  1.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  1.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  1.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  3.  3.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  4.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  2.  3.  1.  0.  0.  0.  0.  0.]
 [ 0.  1.  0.  0.  0.  1.  0.  0.  4.  0.  0.]
 [ 0.  0.  0.  0.  2.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  2.  2.  0.  0.  0.]
 [ 0.  0.  1.  0.  0.  0.  1.  0.  0.  0.  0.]
 [ 0.  0.  1.  1.  0.  0.  0.  1.  1.  0.  0.]
 [ 0.  0.  0.  1.  1.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  1.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  2.  5.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  4.  7.  0.  0.]
 [ 0.  0.  0.  0.  0.  1.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  1.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  1.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  1.  0.  1.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  1.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  2.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  1.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  4.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  1.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  3.]
 [ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
```

Figure 27.- Example of a joint site frequency spectrum (jsfs) from a pipeline run.
This matrix is the jsfs of the samples obtained from a simulation run. The number in each cell is the site number with a minor allele count. Here the rows have the counts for modern samples and the columns for ancient samples. Counts start at zero. Therefore, the first row are all the sites in which modern samples have 0 minor alleles and the number of segregating sites in which ancient samples have 0, 1, 2, 3, etc. minor alleles, according to the column.

Expansion of joint site frequency spectra

The folded jsfs calculated in last step are for samples with 1,000,000 bases. This is a small number to make demographic analysis. Gronau, Hubisz, Gulko, Danko, & Siepel (2011) estimated ancestral population sizes, divergence time and migration rates from inferences using 37,574 regions of 1,000 nucleotide bases that are not affected by natural selection. Based on this number, we calculated 20 joint site frequency spectra for 38,000,000 bases by sampling with replacement all the 100 jsfs found for each of the demographic scenarios (Figure 28).

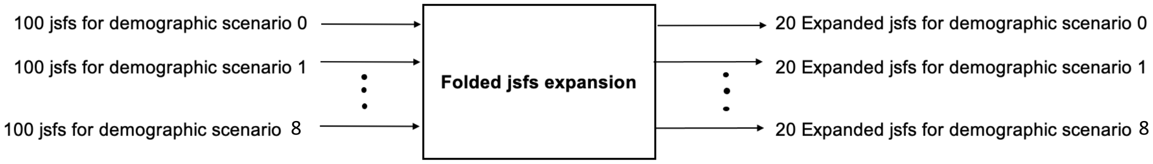

100 jsfs for demographic scenario 0
100 jsfs for demographic scenario 1
100 jsfs for demographic scenario 8
Folded jsfs expansion
20 Expanded jsfs for demographic scenario 0
20 Expanded jsfs for demographic scenario 1
20 Expanded jsfs for demographic scenario 8

Figure 28.- Inputs/Outputs from Expansion of joint frequency spectra stage
The inputs for the Perl script are the 100 jsfs calculated for each demographic scenario. Then they are sampled with replacement to generate 20 expanded jsfs of 38,000,000 bases each. The outputs are 20 expanded jsfs per scenario for the 8 demographic scenarios considered (check Table 4 to see the list of the eight demographic scenarios considered).

Exploring differences in aspects of the expanded jsfs as a function of the divergence time and the sample size in the ancient population

We calculated a set of small summary statistics from the 20 expanded jsfs produced for every demographic scenario. Those statistics are:

- The number of nonvariable sites. Those are the sites in the expanded jsfs where there is only one allele present in the modern and the ancient population.
- The number of differences between sites that contain an allele fixed in one population but variable in the other population between pairs of expanded jsfs. This is equal to:

$$Diff = \sum_{i=1}^{\#modern\ samples} |SFS1_{0,i} - SFS2_{0,i}|$$

$$Diff = \sum_{i=1}^{\#ancient\ samples} |SFS1_{i,0} - SFS2_{i,0}|$$

Where $SFS1_{i,j}$ refers to the site number with i allele copies at a lower frequency in the ancient population and j copies of the allele at a smaller frequency in the modern population.

- The difference number between two jsfs is:

$$Diff = \sum_{i=1}^{\#modern\ samples} \sum_{j=1}^{\#ancient\ samples} |SFS1_{j,i} - SFS2_{j,i}|$$

Inference of demographic parameters

We analyzed if the joint frequency spectrum (jsfs) can be used to infer demographic parameters when employing ancient DNA data. We tested this using fastsimcoal2 (Laurent Excoffier et al., 2021) which employs the joint site frequency spectrum to infer demographic parameters under any model. Particularly, we analyzed the inference of the divergence time for the 9 demographic scenarios. We particularly chose to analyze the inference of one demographic parameter of interest, the divergence time, based on data from the jsfs on the 9 demographic scenarios. In our analysis I compared how the number of samples and the time of divergence itself would impact the precision of the fastsimcoal inference.

To infer demographic parameters using fastsimcoal2, we need to determine an evolutionary model scenario that fits best the data including a set of parameters according to a maximum likelihood approach. To do this, we ran the fastsimcoal2 software with each expanded jsfs and the model template file 100 times. fastsimcoal2 uses a greedy algorithm to estimate the demographic parameters, that is why we need to run the software 100 times to find the best parameter estimate after running the program from 100 different starting points. The output of each run of fastsimcoal2 are the maximum likelihood estimates, the expected sfs from the demographic parameters that include the parameter values, and the parameter estimates. We can compare the expected jsfs (input) vs the expected jsfs (output) visually to see the fit of the model (Figure 29).

Each run of fastsimcoal2 employed the following parameters:

```
./fsc2709 -t output0.tpl -n 100000 -e output0.est -M -L 40 -q -c0 -r $i --logprecision 18 -m
```

Where: “-t output0.tpl” represent the demographic model ran. In this case the demographic model represents two populations that became split at a certain divergence time that will be inferred.

“-n 100000” The number of simulations that will be used to create the expected SFS given the demographic model and a certain value for the parameters (in this case the only parameter is the divergence time).

“-e output0.est” contains the list of parameters and its likely range. In this case it is only the divergence time, and we used a range of 0-10000 and that represents the values of the parameters that fastsimcoal2 can use.

“-M” Infer the demographic parameter via a maximum likelihood approach.

“-L 40” Number of iterations used in the ECM optimization scheme adopted by fastsimcoal2.

“-q” Do not print too many messages to the terminal.

“-c0” Let fastsimcoal2 choose the best core usage scheme to run the program.

“-r $i” random seed used, where different numbers of $i are used in each run.

“--logprecision 18” number of decimal numbers used when running the program.

“-m” use the folded site frequency spectrum to do the analysis. In this case, the joint folded site frequency spectrum will be used to perform the inferences.

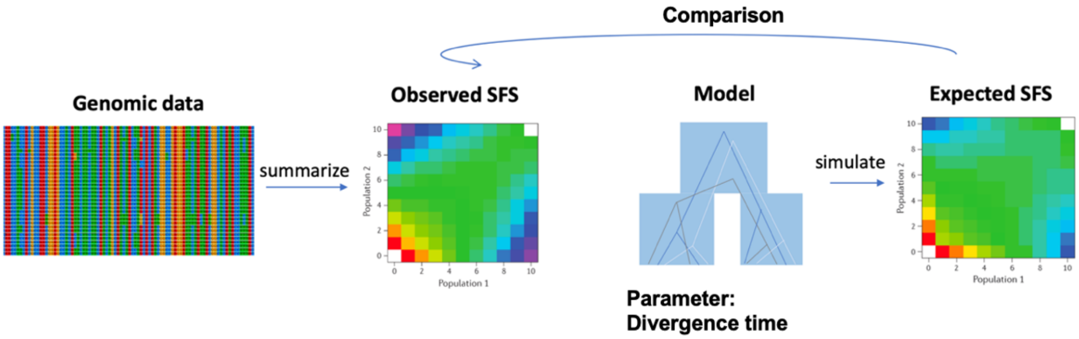


Figure 29.- fastsimcoal2 process to determine best model (Figure taken from a PowerPoint presentation by Meier and Joana)
Genomic data are the DNA sequences that are summarized in the site frequency spectrum. The sfs and the evolutionary model is the input for fastsimcoal2. fastsimcoal2 infers an expected SFS. Both sfs are compared to determine the fit of the model.

Results

Visualization of the jsfs as a function of the divergence time

We inspected the form of the expanded folded joint site frequency spectrum (jSFS) and how it varies due to changes in the divergence time between a present-day population and an ancient population. As a first step, we visualized the expanded joint site frequency spectrum of one simulation made out of 38,000,000 base pairs for three different values of the divergence time when the sample size in the ancient population is equal to 5 individuals (or 10 chromosomes) (Table 5-7). The visual representation of the tables if helpful to analyze broad properties of the jsfs. First, we find that most of the sites in the jsfs are fixed for one allele in both populations in the jsfs generated with the three different divergence times. The number of sites with a 0 minor-allele count in the modern and ancient population is 37977079, 37976043 and 37976043 under the simulations performed with a divergence time of 0, 288 and 600, respectively. Note that the simulated positions number is 38000000. Therefore, the site number with a minor-allele count of 0 in the modern and ancient population, which we will define as nonvariable sites, accounts for approximately 99.9% of all the positions.

Other elements of the jsfs appear to be variable as a function of the divergence time. As an example, when we count the number of positions where the less frequent allele appears once in the ancient population and is absent in the modern population, we see that this count is equal to 739, 1856, and 2835 in the simulations done with a divergence time equal to 0, 288 and 600 generations ago. This provides an additional line of evidence that the jsfs is sensitive to changes in the divergence time. Similar trends can be observed for other elements of the jsfs. In the next three sections I analyze how the number of nonvariable sites, number of sites fixed in one population but variable on the other population, and broad changes on the jsfs changes as a function of the divergence time and the sample number taken in the ancient deme.

Table 5.- Folded joint site frequency spectrum in a simulation of 38,000,000 bases using a divergence time of 0 generations between the modern and the ancient population

Modern\Ancient	*0*	*1*	*2*	*3*	*4*	*5*
0	37977079	739	39	5	0	0
1	3804	525	88	9	0	0
2	1566	607	111	15	6	3
3	1013	515	100	26	0	2
4	685	380	138	28	10	4
5	426	452	129	34	14	1
6	326	337	127	71	10	8
7	211	235	231	66	17	6
8	160	271	144	50	31	11
9	141	218	168	71	28	7
10	106	125	172	97	45	22
11	58	158	135	101	45	22
12	94	167	152	105	24	36
13	52	143	176	117	50	22
14	45	120	124	85	45	32
15	43	100	100	118	86	35
16	33	54	135	133	76	16
17	17	67	111	97	124	34
18	16	63	110	104	81	42
19	16	42	100	116	95	45
20	7	43	79	84	83	60
21	5	32	75	104	64	71
22	11	19	69	93	106	54
23	1	19	54	111	131	106
24	1	6	68	61	109	53
25	5	11	22	68	99	66
26	8	19	58	83	118	66
27	11	23	48	98	98	75
28	2	13	44	86	104	73
29	2	22	37	112	104	48
30	6	13	5	41	70	75

Table 6.- Folded joint site frequency spectrum in a simulation of 38,000,000 bases using a divergence time of 288 generations between the modern and the ancient population

Modern\Ancient	*0*	*1*	*2*	*3*	*4*	*5*
0	37976043	1856	375	108	30	2
1	4183	261	116	62	34	2
2	1947	235	63	72	20	3
3	1270	272	130	52	30	8
4	851	235	112	53	33	9
5	602	191	146	80	33	19
6	516	241	87	61	26	6
7	361	186	113	65	41	5
8	311	135	134	88	17	13
9	211	172	109	31	41	14
10	164	161	112	73	47	19
11	144	128	144	64	31	13
12	122	124	152	50	52	16
13	95	81	104	68	44	36
14	114	77	77	66	62	44
15	58	92	115	58	68	38
16	73	64	76	58	59	34
17	97	71	50	77	69	41
18	28	77	57	80	59	35
19	27	50	95	73	55	30
20	26	53	44	66	89	38
21	21	38	55	101	58	39
22	34	37	53	69	97	50
23	30	37	49	64	101	37
24	7	53	73	76	87	63
25	6	26	70	71	101	48
26	13	39	60	77	76	35
27	11	26	53	46	82	28
28	8	12	74	82	94	31
29	4	15	56	34	55	47
30	2	10	13	38	23	18

Table 7.- Folded joint site frequency spectrum in a simulation of 38,000,000 bases using a divergence time of 600 generations between the modern and the ancient population

Modern\Ancient	0	1	2	3	4	5
0	37973398	2835	844	291	117	34
1	4587	123	95	39	29	11
2	2005	141	97	73	30	9
3	1347	157	107	51	61	19
4	891	130	103	42	28	19
5	626	133	119	54	48	22
6	507	107	96	67	45	16
7	454	76	75	70	51	34
8	411	103	90	51	73	20
9	320	135	94	63	50	28
10	240	96	76	38	73	34
11	220	117	136	69	61	23
12	190	112	65	30	52	23
13	165	99	66	53	68	66
14	179	67	79	52	53	28
15	111	75	48	40	84	21
16	140	81	44	66	101	29
17	111	92	54	90	86	30
18	61	35	41	46	60	30
19	47	52	66	71	122	30
20	51	40	72	41	69	53
21	60	76	56	63	75	42
22	70	56	68	68	89	46
23	49	52	47	60	62	29
24	73	53	70	71	70	43
25	46	55	70	66	88	58
26	39	33	59	104	93	39
27	42	39	74	60	55	39
28	18	27	52	75	86	46
29	48	43	49	56	94	25
30	9	40	25	49	49	20

Number of nonvariable sites depending on the sample number and the divergence time

The jSFS is composed of many elements, shown in a multidimensional matrix, that are jointly informative about past population demographic history. One of the informative elements in the jSFS is a cell in the matrix that counts the number of positions in the genome where the alleles are fixed in both populations. We will refer to these positions as nonvariable sites. In our simulations we analyzed the nonvariable site number depending on sample sizes and the past population divergence time. We found that the number of nonvariable sites decreases as a function of the sample size for the three divergence times analyzed (Figure 30). This result is congruent with coalescent theory which states that the branch length of a genealogy increases with the sample number (Wakeley, 2009). The increase in the branch length of the genealogies decreases the nonvariable site number since the mutation number on a sample is equal to the product of the genealogical branch length times the mutation rate. The nonvariable site number is equal to the positions analyzed minus the number of mutations in the individuals simulated.

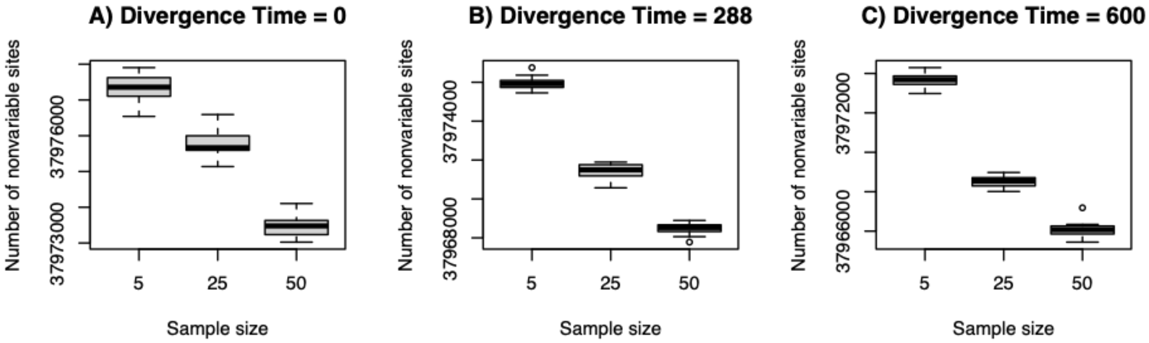


Figure 30.- Number of nonvariable sites depending on the sample size in the ancient population for a divergence time of A) 0 generations, B) 288 generations and C) 600 generations.

We also evaluated the impact of the divergence time on the number of nonvariable sites. We found that the number of nonvariable sites decreases as a function of the divergence time. This observation is also consistent with coalescent theory since an increased divergence time also increases the genealogy branch length. The increased genealogy branch length also increases the number of mutations, and this leads to a decreased number of nonvariable sites. This particular effect is observed for every sample size of the ancient population analyzed (Figure 31).

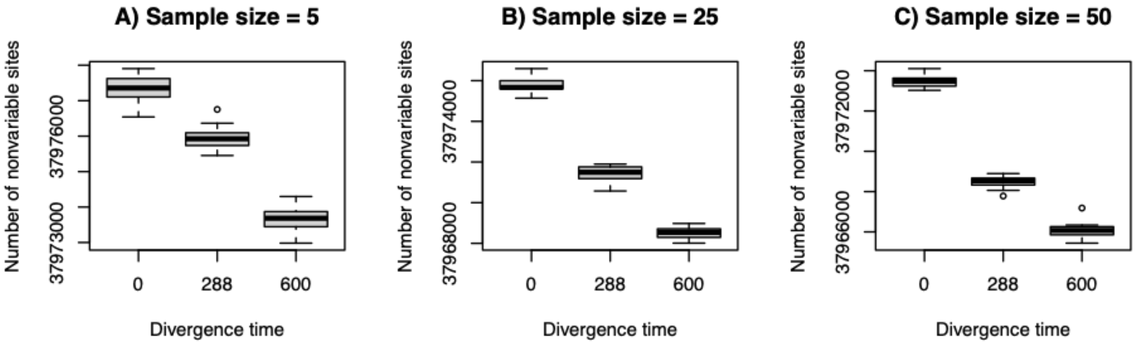


Figure 31.- Number of nonvariable sites depending on the divergence time between the ancient population and the present day population for a sample size of A) 5, B) 25 and C) 50 individuals.

Differences in position numbers on sites with an allele fixed in one population and two alleles present in the other population

Another element of interest in the jSFS are the site number with a fixed allele in one population and two alleles seen in the other population. This particular component of the joint site frequency spectrum is interesting to analyze because it helps us to understand how the divergence time has an impact on the sites that are variable in only one of the two analyzed populations. We find that the sites with the allele is fixed in one population but there are two alleles in the other population are sensible to changes in the divergence time (Figure 32, 33). Particularly, we see clear differences when comparing the number of fixed positions in the modern population and variable in the ancient deme when comparing the jsfs generated with different parameters of the divergence time (Figure 32). We also see that important differences in the count of positions that are variable in the modern population and fixed in the ancient population depending on the divergence time (Figure 33). These results provide additional evidence that there are components in the jsfs that vary as a function of our parameter of interest, the divergence time. This is important because we will use the statistic jsfs to perform inferences of the divergence time.

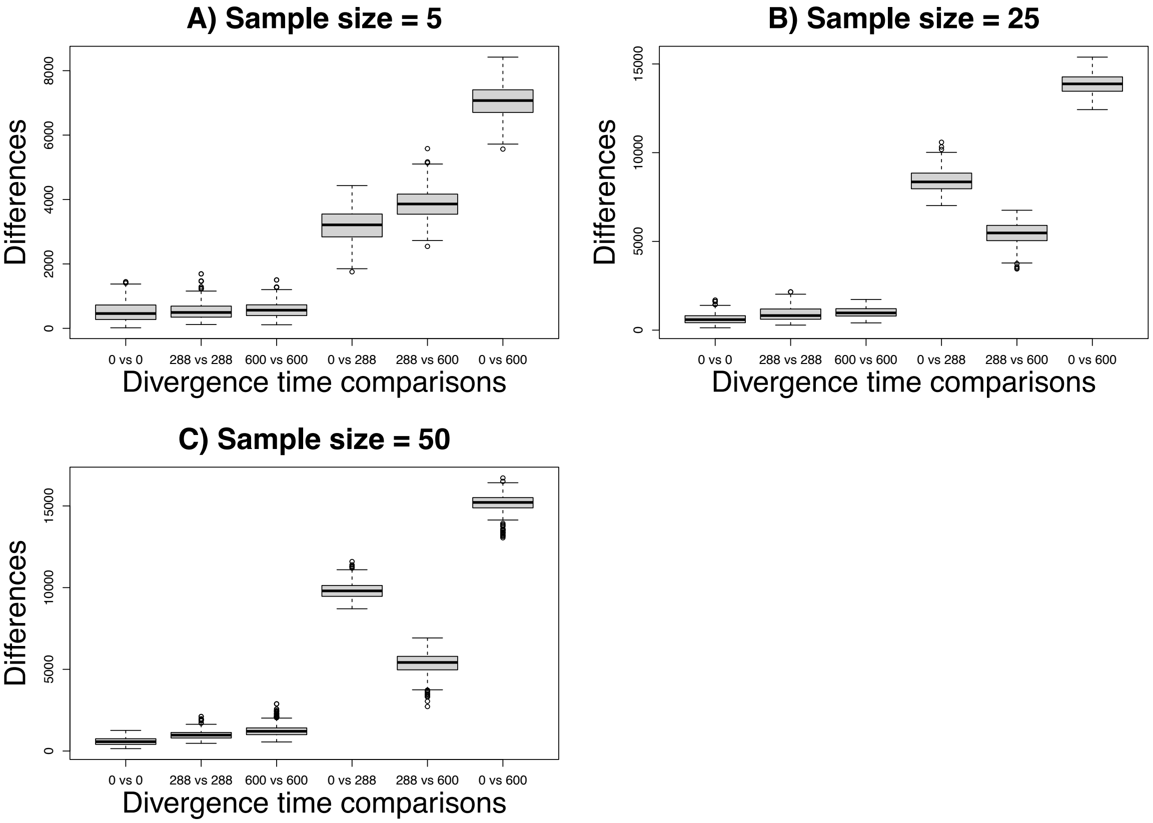


Figure 32.- Differences between jsfs' as a function of the divergence time and the sample size when analyzing fixed positions in the modern population and variable in the ancient deme.

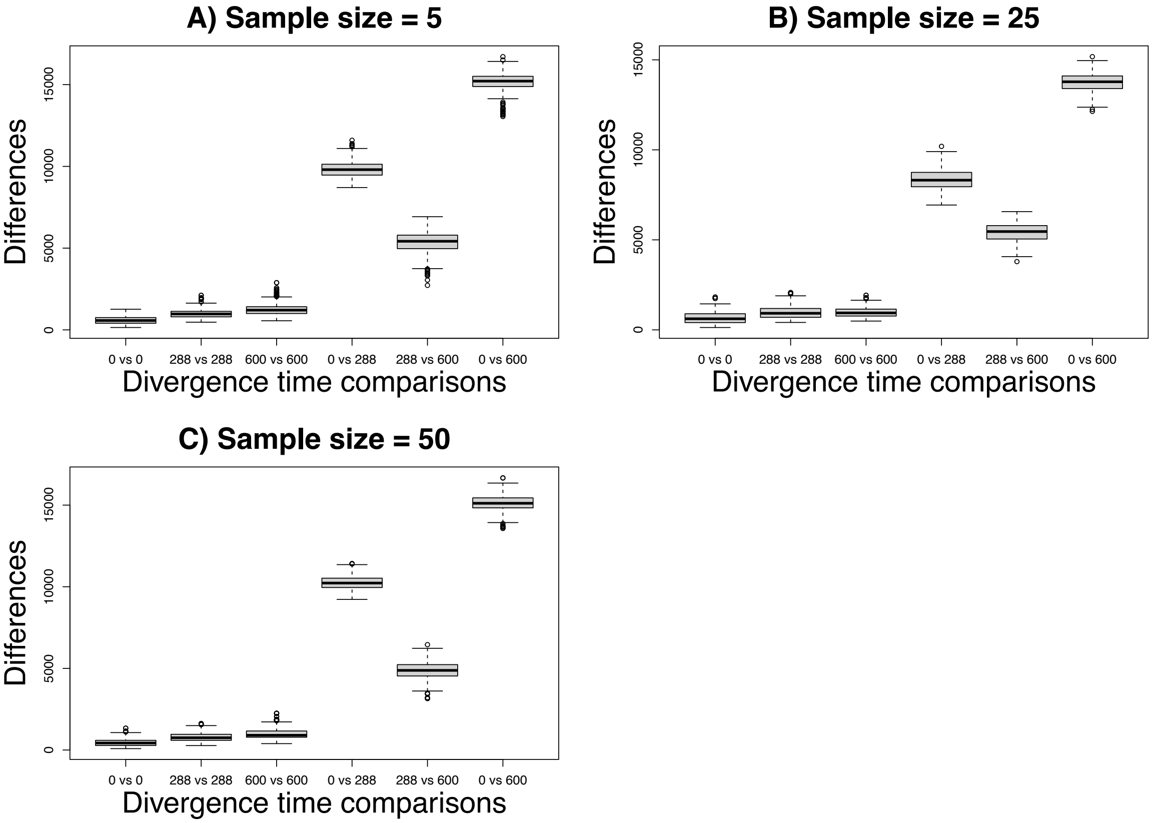


Figure 33.- Differences between jsfs' as a function of the divergence time and the sample size when analyzing positions that are variable in the modern population and fixed in the ancient population.

Differences between the joint site frequency spectrum depending on the divergence time between populations

Apart from the nonvariable site number and the fixed site number in one population but not the other one, we also analyzed if the jsfs is broadly sensitive to changes in the divergence time between an ancient and a present-day population. To do this, we used a metric that captures the number of differences between 20 jsfs generated with different divergence time values. We found that the jsfs display fewer differences when they are simulated under the same divergence time compared to simulations with a different value (Figure 34). The differences between the jsfs' simulated with a different divergence time are important for demographic inferences, since they indicate that this statistic is sensitive to changes in the divergence time parameter which should make this statistic useful for demographic inferences (Figure 34). Interestingly, the differences between jsfs become larger as the sample size increases (Figure 34).

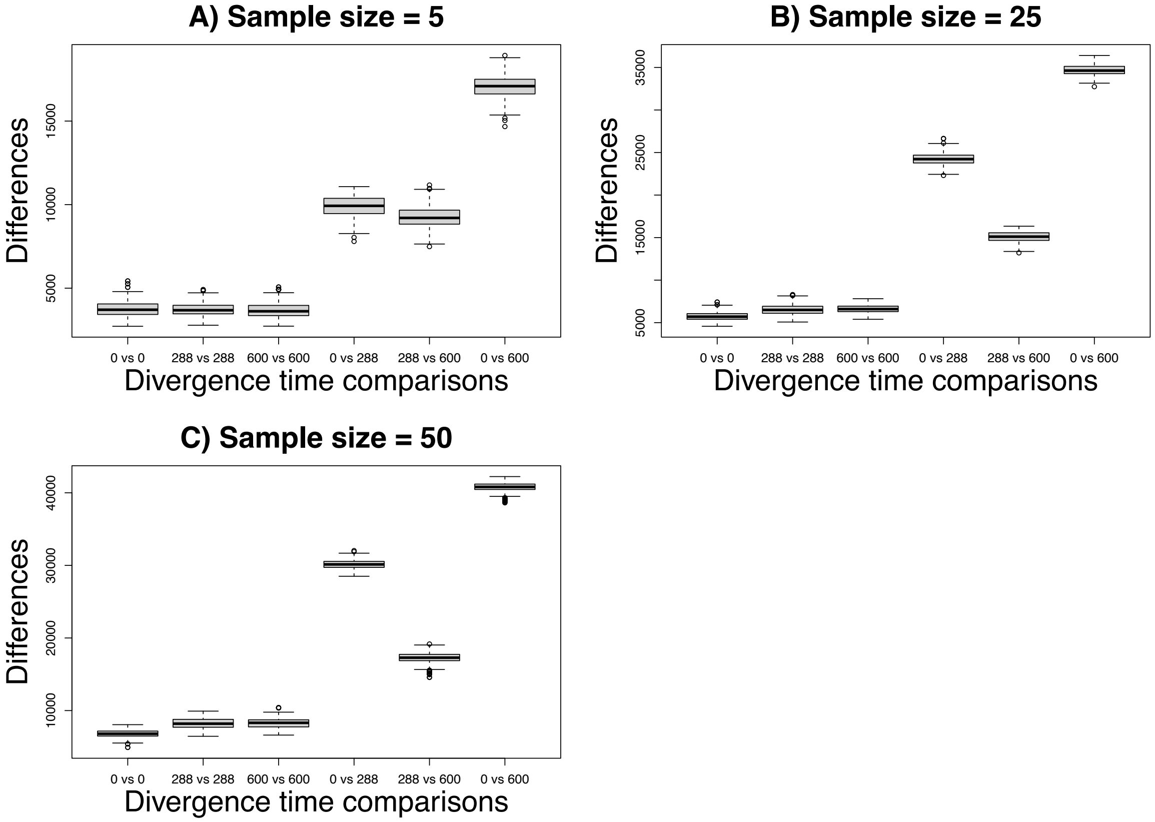


Figure 34.- Differences between jsfs as a function of the divergence time and the sample size

Inference accuracy

We found that the estimates of the divergence time were very accurate when the simulated divergence time was equal to 0 (Figure 35A). Note that estimates of the divergence time between 0 and 20 imply an accurate estimation of the divergence time since the ancient sample is taken from 20 generations ago. The modern lineages can not coalesce the ancient lineages until the ancient samples appear in the genealogy 20 generations ago and, therefore, the genealogies should be identical under the coalescent model for any divergence time between 0 and 20. On the other hand, we found that the estimates were not as accurate when the divergence time was equal to 288 or 600 (Figure35B-35C). Interestingly, we found that the divergence time estimates were closer to the true values when the sample size increased for simulations done under a divergence time equal to 288 or 600 (Figure 35B-35C).

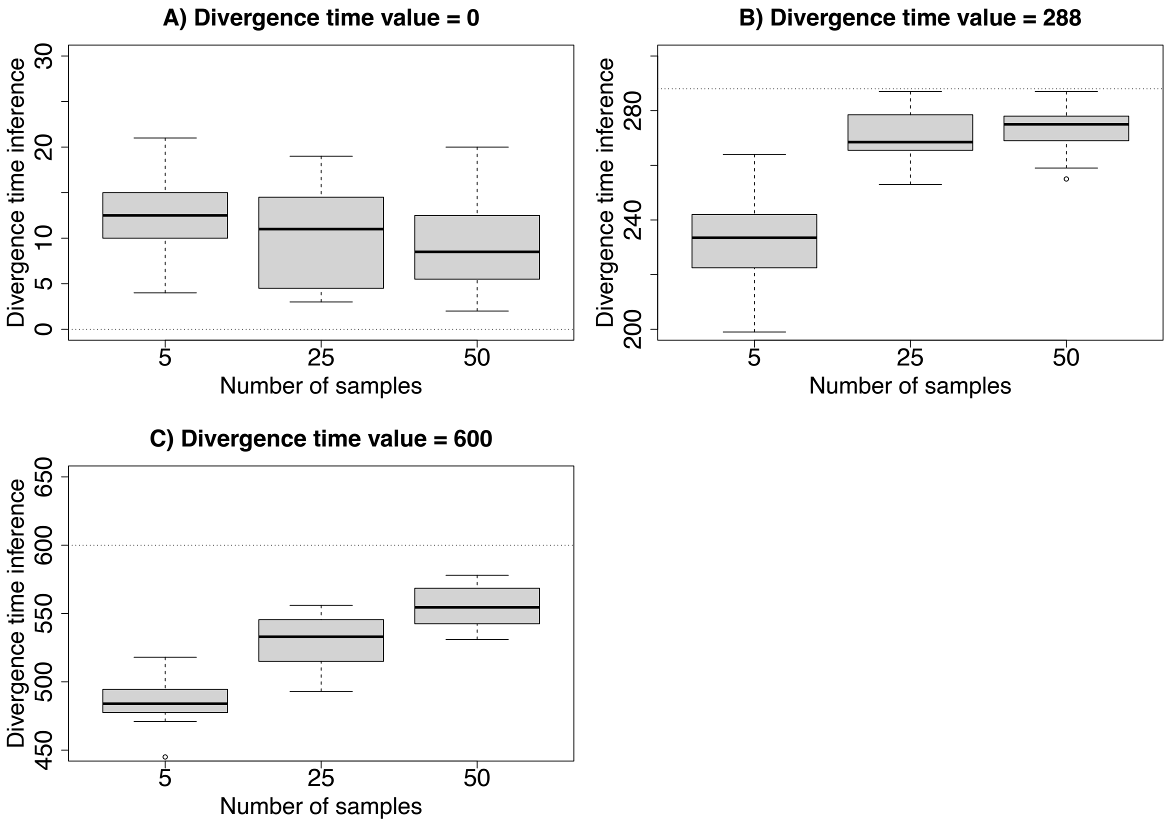


Figure 35.- Inferences of the divergence time depending on the sample size in the ancient population for 3 different sample sizes used in the ancestral population.

Conclusions

The divergence time between populations is a parameter of high interest in population genetics because it informs us about the continuity between populations living in the present and in the past. If this parameter is equal to zero, it indicates that the populations are continuous and that the populations living in the present are direct descendants of a population living in the past. Higher values of this parameter show that the populations are more divergent from each other. In this dissertation we analyzed if this parameter could be inferred using the joint site frequency spectrum. Our results indicate that the joint site frequency spectrum is an informative statistic about the divergence time between two populations.

First, we show that one statistic that is derived from the site frequency spectrum, the number of nonvariable sites, is negatively correlated with the divergence time and also with the sample size. This is an expected result that is congruent result with coalescent theory. The genealogical branch length increases with the divergence time and the sample size, and that should decrease the number of nonvariable sites which is what I find in my study.

Second, we find that the joint site frequency spectrum displays differences when doing simulations under a different divergence time. This is an important point because this is the statistic that we use to perform inferences. Interestingly, the differences found are even largely increased when the sample sizes used become larger.

We analyzed if the divergence time parameter can be estimated from the joint site frequency spectrum. We found that this parameter shows a slight underestimation bias when the divergence time is equal to 288 and 600 generations. Interestingly, the bias is decreased when the sample size increases. This indicates that the sample size should be considered when analyzing the power to estimate a demographic parameter of interest, such as the divergence time between populations.

The joint site frequency spectrum has been broadly used to infer past demographic events using software like dadi or fastsimcoal2 (Gutenkunst et al., 2009; Excoffier et al. 2013). The inferences performed using the site frequency spectrum have been shown to give accurate results on various demographic scenarios (Adrion et al., 2020). However, the site frequency spectrum has been found to show an identical form for very different past demographic histories in some cases using only genomic information from modern populations (Myers et al. 2008). The use of the joint site frequency spectrum to infer past demographic history using jointly modern and ancient populations has not been carefully explored. Broadly, our results show that inferences of demographic parameters of interest using the joint site frequency spectrum when including ancient samples should be analyzed carefully. We suggest performing inferences under the sample size at hand to see if the demographic inferences of one demographic parameter in particular could be accurately estimated given our study design. Our results focus on the analysis of the divergence time, but other demographic parameters could suffer similar biases when performing inferences under the joint site frequency spectrum.

Many ancient samples will be collected and sequenced in the near future and will help us gain important insights about out past population history. However, these inferences should be treated with care since they could be based on statistics that are not sufficient to accurately infer a demographic parameter of interest given the sample size available in the study at hand. Here we present an analysis focused on demographic inferences performed with the joint site frequency spectrum. However, these results could also be found with any other statistic since, in the end, all statistics of diversity are a function of the joint site frequency spectrum (Achaz, 2008). Performing simulations that analyze the power to infer past demographic parameters under a particular study design at hand, with its sample sizes available, should be a crucial step in future analysis to evaluate if we have the necessary data to perform accurate inferences of past demographic history.

References

Acosta Ochoa, G. (2010). Late-Pleistocene / Early Holocene Tropical Foragers of Chiapas, Mexico: Recent Studies. *Current Research in the Pleistocene*, *27*, 1–4.

Achaz, G. (2008) Frequency Spectrum Neutrality Tests: One for all and all for one. Genetics, 183(1), 249-258.

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research, 19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Ardelean, C. F., Becerra-Valdivia, L., Pedersen, M. W., Schwenninger, J.-L., Oviatt, C. G., Macías-Quintero, J. I., … Willerslev, E. (2020). Evidence of human occupation in Mexico around the Last Glacial Maximum. *Nature*, *584*(7819), 87–92. https://doi.org/10.1038/s41586-020-2509-0

Adrion, J. R., Cole, C. B., Dukler, N., Galloway J. G., Gladstein, A. L., Gower, G., Kyriazis, C. C., Ragsdale, A. P., Tsambos, G. (2020) A community-maintained standard library of population genetic models. eLife, 9:e54967.

Ávila-Arcos, M. C., McManus, K. F., Sandoval, K., Rodríguez-Rodríguez, J. E., Villa-Islas, V., Martin, A. R., … Moreno-Estrada, A. (2020). Population history and gene divergence in Native Mexicans inferred from 76 human exomes. *Molecular Biology and Evolution*, *37*(4), 994–1006. https://doi.org/10.1093/molbev/msz282

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., … Ralph, P. L. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, *220*(3).

Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P., & Ramachandran, S. (2016). Pong: Fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics*, *32*(18), 2817–2823. https://doi.org/10.1093/bioinformatics/btw327

Bennett, M. R., Bustos, D., Pigati, J. S., Springer, K. B., Urban, T. M., Holliday, V. T., … Odess, D. (2021). Evidence of humans in North America during the Last Glacial Maximum. *Science*, *373*(6562), 1528–1531.

Bhatia, G., Patterson, N., Sankararaman, S., & Price, A. L. (2013). Estimating and interpreting FST: The impact of rare variants. *Genome Research*, *23*(9), 1514–1521. https://doi.org/10.1101/gr.154831.113

Briggs, A. W., Stenzel, U., Johnson, P. L. F., Green, R. E., Kelso, J., Prüfer, K., … Pääbo, S. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(37), 14616–14621. https://doi.org/10.1073/pnas.0704665104

Carlhoff, S., Akin, D., Nägele, K., Sumantri, I., Nur, M., Skov, L., … Brumm, A. (2021). Genome of a middle Holocene hunter-gatherer from Wallacea. *Nature*, *596*, 543–547.

Chatters, J. C., Kennett, D. J., Asmerom, Y., Kemp, B. M., Polyak, V., Blank, A. N., … Stafford, T. W. (2014). Late Pleistocene human skeleton and mtDNA link Paleoamericans and modern Native Americans. *Science*, *344*(6185), 750–754.

https://doi.org/10.1126/science.1252619

Cockerham, C. C. (1969). Variance of gene frequencies. *Evolution*, *23*(1), 72–84.

Dabney, J., Meyer, M., & Pääbo, S. (2013). Ancient DNA damage. *Cold Spring Harbor Perspectives in Biology*, *5*(7), 1–8. https://doi.org/10.1101/cshperspect.a012567

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., … Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), 1–4. https://doi.org/10.1093/gigascience/giab008

Davis, L. G., Madsen, D. B., Becerra-Valdivia, L., Higham, T., Sisson, D. A., Skinner, S. M., … Buvit, I. (2019). Late Upper Paleolithic occupation at Cooper's Ferry, Idaho, USA, ~ 16,000 years ago. *Science*, *265*(August), 891–897.

Des Lauriers, M. R., Davis, L. G., Turnbull, J., Southon, J. R., & Taylor, R. E. (2017). The Earliest Shell Fishhooks from the Americas Reveal Fishing Technology of Pleistocene Maritime Foragers. *American Antiquity*, *82*(3), 498–516. https://doi.org/10.1017/aaq.2017.13

Edwards, S., & Beerli, P. (2000). Perspective: Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution*, *54*(6), 1839–1854. https://doi.org/10.1111/j.0014-3820.2000.tb01231.x

Excoffier, L, Dupanloup, I., Huerta-Sánchez, E., Sousa, V., & Foll, M. (2013). Robust Demographic Inference from Genomic and SNP Data. *PLos Genetics*. https://doi.org/https://doi.org/10.1371/journal.pgen.1003905

Excoffier, Laurent, Marchi, N., Marques, D. A., Matthey-Doret, R., Gouy, A., & Sousa, V. C. (2021). Fastsimcoal2: Demographic Inference Under Complex Evolutionary Scenarios. *Bioinformatics*, *37*(24), 4882–4885. https://doi.org/10.1093/bioinformatics/btab468

Gómez-Robles, A. (2019). Dental evolutionary rates and its implications for the Neanderthal–modern human divergence. *Science Advances*, *5*(5), 1–10. https://doi.org/10.1126/sciadv.aaw1268

González, A. H., Terrazas, A., Stinnesbeck, W., Benavente, M. E., Avilés, J., Padilla, J. M., … Frey, E. (2014). The First Human Settlers on the Yucatan Peninsula: Evidence from Drowned Caves in the State of Quintana Roo. In K. E. Graf, C. V. Ketron, & M. R. Waters (Eds.), *Paleoamerican Odyssey* (pp. 323–338). Texas A&M University Press.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., … Pääbo, S. (2010). A draft sequence of the neandertal genome. *Science*, *328*(5979), 710–722. https://doi.org/10.1126/science.1188021

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., & Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics*, *43*(10), 1031–1034. https://doi.org/10.1038/ng.937

Günther, T., & Jakobsson, M. (2019). Population Genomic Analyses of DNA from Ancient Remains. *Handbook of Statistical Genomics*, *1*, 295–40. https://doi.org/10.1002/9781119487845.ch10

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., Bustamante, C.D. (2009) Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP

Frequency Data. *PLoS Genetics*, 5(10), e1000695.

Han, E., Sinsheimer, J. S., & Novembre, J. (2014). Characterizing bias in population genetic inferences from low coverage sequencing data. *Molecular Biology and Evolution*, *31*(3), 723–735. https://doi.org/10.1093/molbev/mst229

Ho, S. Y. W., Heupink, T. H., Rambaut, A., & Shapiro, B. (2007). Bayesian estimation of sequence damage in ancient DNA. *Molecular Biology and Evolution*, *24*(6), 1416–1422. https://doi.org/10.1093/molbev/msm062

Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: Defining, estimating, and interpreting FST. *Nature Reviews Genetics*, *10*(9), 639–650. https://doi.org/10.1038/nrg2611

Illumina, I. (2021a). Coverage depth recommendations. Retrieved October 5, 2021, from https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/coverage.html

Illumina, I. (2021b). Key differences between next-generation sequencing and Sanger sequencing. Retrieved from https://www.illumina.com/science/technology/next-generation-sequencing/ngs-vs-sanger-sequencing.html

Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1004842

Kelleher, J., & Lohse, K. (2020). Coalescent Simulation with msprime. In J. Dutheil (Ed.), *Statistical Population Genomics. Methods in Molecular Biology.* (pp. 191–230). https://doi.org/10.1007/978-1-0716-0199-0_9

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Their Applications*, *13*, 235–248.

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, *15*(1), 1–13. https://doi.org/10.1186/s12859-014-0356-4

Korneliussen, T. S., Moltke, I., Albrechtsen, A., & Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics*, *14*(1). https://doi.org/10.1186/1471-2105-14-289

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K., & Wang, J. (2009). SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics*, *25*(15), 1966–1967. https://doi.org/10.1093/bioinformatics/btp336

Lindhal, T. (1993). Instability and decay of the primary structure of DNA. *Nature*, *362*(22 April), 709–715.

Marciniak, S., & Perry, G. H. (2017). Harnessing ancient genomes to study the history of Huan adaptation. *Nature Reviews Genetics*, *18*, 659–674.

Mayo, O. (2008). A Century of Hardy – Weinberg Equilibrium. *Twin Research and Human Genetics*, *11*(3), 249–256.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Meltzer, D. J. (2009). *First Peoples in a New World: Colonizing Ice Age America* (U. of C. Press, Ed.).

Menozzi, P., Piazza, A., & Cavalli-Sforza, L. (1978). Synthetic Maps of Human Gene Frequencies in Europeans. *Science (New York, N.Y.)*, *201*(4358), 786–791. Retrieved from http://www.sciencemag.org/content/201/4358/786.short

Moreno-Mayar, J. V., Potter, B. A., Vinner, L., Steinrücken, M., Rasmussen, S., Terhorst, J., … Willerslev, E. (2018). Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature*, *553*(7687), 203–207. https://doi.org/10.1038/nature25173

Moreno-Mayar, J. V., Vinner, L., de Barros Damgaard, P., de la Fuente, C., Chan, J., Spence, J. P., … Willerslev, E. (2018). Early human dispersals within the Americas. *Science*, *362*(6419). https://doi.org/10.1126/science.aav2621

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D. J., Salichos, L., … Gerstein, M. (2016). The real cost of sequencing: Scaling computation to keep pace with data generation. *Genome Biology*, *17*(1), 1–9. https://doi.org/10.1186/s13059-016-0917-0

Myers, S., Fefferman, C., Patterson, N. (2008) Can one learn history from the allelic spectrum. *Theoretical Population Biology*, 73, 342-348.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America*, *70*(12 (I)), 3321–3323. https://doi.org/10.1073/pnas.70.12.3321

Nielsen, R., Akey, J., & Jakobsson, M. (2017). Tracing the peopling of the world through genomics. *Nature*, (541).

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE*, *7*(7). https://doi.org/10.1371/journal.pone.0037558

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, *12*(6), 443–451. https://doi.org/10.1038/nrg2986

Nielsen, R., & Slatkin, M. (2000). Likelihood analysis of ongoing gene flow and historical association. *Evolution*, *54*(1), 44–50. https://doi.org/10.1111/j.0014-3820.2000.tb00006.x

Nielsen, R., & Slatkin, M. (2013). *An Introduction to Population Genetics: Theory and applications*. Sinauer Associates.

Noskova, E., Ulyantsev, V., Koepfli, K. P., O’brien, S. J., & Dobrynin, P. (2020). GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data. *GigaScience*, *9*(3), 1–18. https://doi.org/10.1093/gigascience/giaa005

Novembre, J., & Ramachandran, S. (2011). Perspectives on human population structure at the cusp of the sequencing Era. *Annual Review of Genomics and Human Genetics*, *12*(August), 245–274. https://doi.org/10.1146/annurev-genom-090810-183123

Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P. W., Ávila-Arcos, M. C., … Warinner, C. (2021). Ancient DNA analysis. *Nature Reviews Methods Primers*, *1*(1). https://doi.org/10.1038/s43586-020-00011-0

Pääbo, S. (1985). Molecular cloning of Ancient Egyptian mummy DNA. *Nature*, *314*(6012), 644–645. https://doi.org/10.1038/314644a0

Pääbo, S. (2018). The Future of Ancient DNA. *Schrödinger at 75: The Future of Biology*. Dublin, Ireland.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., … Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*(3), 1065–1093. https://doi.org/10.1534/genetics.112.145037

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), 2074–2093. https://doi.org/10.1371/journal.pgen.0020190

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, *155*(June), 945–959.

Rambaut, A., & Grassly, N. C. (1997). Seq-gen: An application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Bioinformatics*, *13*(3), 235–238. https://doi.org/10.1093/bioinformatics/13.3.235

Rasmussen, M., Anzick, S. L., Waters, M. R., Skoglund, P., Degiorgio, M., Stafford, T. W., … Willerslev, E. (2014). The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature*, *506*(7487), 225–229. https://doi.org/10.1038/nature13025

Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., … Ruiz-Linares, A. (2012). Reconstructing Native American population history. *Nature*, *488*(7411), 370–374. https://doi.org/10.1038/nature11258

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian population history. *Nature*, *461*(7263), 489–494. https://doi.org/10.1038/nature08365

Richards, M. B., Sykes, B. C., & Hedges, R. E. M. (1995). Authenticating DNA Extracted from Ancient Skeletal Remains. *Journal of Archaeological Science*, *22*(2), 291–299. https://doi.org/10.1006/jasc.1995.0031

Rosenberg, N. A., & Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, *3*(5), 380–390. https://doi.org/10.1038/nrg795

Sanchez, G., Holliday, V. T., Gaines, E. P., Arroyo-Cabrales, J., Martínez-Tagüeña, N., Kowler, A., … Sanchez-Morales, I. (2014). Human (Clovis)-gomphothere (Cuvieronius sp.) association ∼13,390 calibrated yBP in Sonora, Mexico. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(30), 10972–10977. https://doi.org/10.1073/pnas.1404546111

Schlebusch, C. M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., … Jakobsson, M. (2017). Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science*, *358*(6363), 652–655. https://doi.org/10.1126/science.aao6266

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods*, (5), 16–18.

Skoglund, P., Mallick, S., Bortolini, M. C., Chennagiri, N., Hünemeier, T., Petzl-Erler, M. L., … Reich, D. (2015). Genetic evidence for two founding populations of the Americas. *Nature*, *525*(7567), 104–108. https://doi.org/10.1038/nature14895

Slatkin, M. (2016). Statistical methods for analyzing ancient DNA from hominins. *Current Opinion in Genetics and Development*, *41*, 72–76. https://doi.org/10.1016/j.gde.2016.08.004

Sousa, V., & Hey, J. (2013). Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Reviews Genetics*, *14*(6), 404–414. https://doi.org/10.1038/nrg3446

Tang, H., Peng, J., Wang, P., & Risch, N. J. (2005). Estimation of individual admixture: Analytical and study design considerations. *Genetic Epidemiology*, *28*(4), 289–301. https://doi.org/10.1002/gepi.20064

Taylor, R. E. (1985). The Beginnings of Radiocarbon Dating in American Antiquity: A Historical Perspective. *American Antiquity*, *50*(2), 309–325.

Wakeley, J. (2009). *Coalescent Theory: an introduction*. Greenwood Village: Roberts and company publisher.

Waters, M. R., Forman, S. L., Jennings, T. A., Nordt, L. C., Driese, S. G., Feinberg, J. M., … Wiederhold, J. E. (2011). The buttermilk creek complex and the origins of clovis at the Debra L. Friedkin site, Texas. *Science*, *331*(6024), 1599–1603. https://doi.org/10.1126/science.1201855

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, *15*, 323–354.

Zheng, Y., & Janke, A. (2018). Gene flow analysis method, the D-statistic, is robust in a wide parameter space. *BMC Bioinformatics*, *19*(1), 1–19. https://doi.org/10.1186/s12859-017-2002-4