UNIVERSITY OF CRETE
IMBB-FORTH

MASTER THESIS

Tracing Disease Variants in Healthy Populations

Author:
Anna MATHIOUDAKI

Supervisor:
Dr. Pavlos PAVLIDIS

A thesis submitted in fulfillment of the requirements
for the degree of MSc in Molecular Biology-Biomedicine

in the

Biology Department, University of Crete, Greece

October, 2018

"Education is an admirable thing, but it is well to remember from time to time that nothing that is worth knowing can be taught "

Oscar Wilde

UNIVERSITY OF CRETE

Abstract

ICS-FORTH
Department of Biology

MSc in Molecular Biology-Biomedicine

Tracing Disease Variants in Healthy Populations

by Anna MATHIOUDAKI

The study of the genetic background of complex traits and diseases is of great interest due to its contribution to the understanding of the underlying related biological mechanisms. Here, we traced disease variants from the ClinVar database in 2,504 healthy individuals of the 1000 Genomes Projects. Interestingly, we identified 1.690 variants to be present. Our goal was to investigate their presence in healthy individuals; to trace their origin and course among populations and unravel their impact in the genotype. Thus, we implemented Population Genetics methodologies on haplotypes carrying the disease allele, clarifying the genetic variation within those haplotypes and detecting negative selection and demographic model for those haplotypes. We noticed that none of those are located in sex chromosomes. We could not identify any general pattern of population stratification since some patterns imply isolated populations and some mixed, with varying measurements of genotypic differentiation. Moreover, we identified 111 genomic regions in the neighborhood of pathogenic alleles indicating to be under positive selection. Though, there was no correlation between frequency within the healthy individuals and those selection events. Our study identifies specific population patterns of the variants' haplotypes and supports the aspect of misclassification of variants in ClinVar database.

...

Acknowledgements

At this point, I would like to thank all the contributors of this project. First of all, I would like to thank all the web-lab oriented groups that were full 1 year ago, while I was looking for a thesis.
Mostly I would like to express my gratitude my Supervisor, Advisor, Professor, Mentor, Motivational Speaker and above all Friend Dr. Pavlos Pavlidis for accepting me to his amazing team. For his patience while I was learning, for his multidisciplinary thought, for teaching me from scratch Bioinformatics, Statistics and Population Genetics.
I would also like also to thank the other two members of my advisory committee: Dr. Christoforos Nikolaou for introducing me to the field of Bioinformatics. I want also to thank Dr. Charalampos Spilianakis, that I also had the opportunity to join his team during one of my Rotations. He has been a great mentor to me.
I feel really grateful for having such amazing laboratory members and friends. Special thanks to Aggelos Koropoulis and my "office partners" Antonis Kioukis and Ioannis Koutsoukos for their help, for all their brainstorming, for just being there when everything looked black (with green letters) to make me laugh and then cry (because I was laughing so hard), Stefanos Papadantonakis for his ability to teach anything to dummies, my travel-buddy Maria Vasilarou, Joanna Garefalaki for the amazing collaboration and her passion for science and Alexandros Marantos for the "Intellectual Exchange".
Nothing would be accomplished without my friends who bravely endured me through this year and supported me in their very own way my flatmate and best-friend Dimitra, Kostas, John, Nikos, Tasos, Kostis, Panos, Lina.

This thesis is dedicated to my parents and brother, who made my studies possible all these years and support me in any way possible. ...

Contents

List of Figures

List of Tables

List of Abbreviations

AML	**A**cute **M**yeloid **L**eukemia
AD	**A**lzheimer **D**isease
SNV	**S**ingle **N**ucleotide **V**ariation
CNV	**C**opy **N**umber **V**ariation
LD	**L**inkage **D**isequillibrium
CDCV	**C**ommon **D**isease **C**ommon **V**ariant
CDRV	**C**ommon **D**isease **R**are **V**ariant
AF	**A**llele **F**requency
DHS	**D**NaseI **H**ypersensitive **S**ite
SFS	**S**ite **F**requency **S**pectrum
GWAS	**G**enome **W**ide **A**ssociation **S**tudies
MSMC	**M**ultiple **S**equentiallly **M**arkovian **C**palescent
TMRCA	**T**ime **M**ost **R**ecent **C**ommon **A**ncestor
HMM	**H**idden **M**arkov **M**odel
PCA	**P**rincipal **C**omponent **A**nalysis
1KG	**1000G**enomes
SAS	**S**outh **As**ian
EAS	**E**ast **As**ian
EUR	**Eur**opean
AFR	**Afr**ican
AMR	**Am**erican

Chapter 1

Background

1.1 Introduction

Recent progress in genetics and in computational biology has played an important role in the identification of genetic marks related to diseases. GWAS have significantly contributed to the identification of strong correlations between single nucleotide variations (SNVs) and genetic diseases. This is achieved through the inclusion of thousands of samples characterized by a specific disease. However, the way those contribute to the pathological phenotype has yet to be clarified.

The discovery of the genetic basis of many diseases can claim causality between genotype and phenotype, and thus to result in more accurate diagnosis. Still the genetic background for half of the known genetic disorders is not established yet (Chong et al., 2015). There are around 6.000 known genetic disorders and despite extensive related studies, they are poorly understood (Chong et al., 2015). For example, **Acute Myeloid Leukemia (AML)** comprises a set of hematological diseases, which display both genotypic and phenotypic heterogeneity. The 80% of the cases occurs in adult individuals. There is no reference biomarker of the disease but recently many genetic variations have been identified in patients with AML, mostly associated with treatment response and disease progress (Lagunas-Rangel et al., 2017). A further example is the progress of the **Alzheimer's disease (AD)** which is one of the most common neurodegenerative diseases. AD has been strongly linked with 'causal' genes (Bekris et al., 2010). For instance, with autosomal dominant familial AD (APP, PSEN1, and PSEN2) and one genetic risk factor (APOEϵ4 allele) (Mez et al., 2017).

1.2 Studying Genetic Variation

1.2.1 Terminology

Penetrance is the ratio of individuals in a population with a risk variant who also have the disease (Gibson, 2012).
Effective population size (Ne) is the size of an ideal population that would have the same level of genetic variance as the real population (Husemann et al., 2016).
Hitchhiking is described as the increased frequency of alleles linked to a beneficial allele, which has undergone natural selection (Kim and Maruki, 2011) **Heritability**is the phenotypic variance in a population, which can be explained by the genotypic differences among individuals (Griffiths et al., 2000).
Haplotype is referred as a group of alleles at neighboring loci of a chromosome that tend to be co-inherited more frequently than expected (Ardlie, Kruglyak, and Seielstad, 2002).
Reference genome refers to the genome based on which the sequence mapping was performed. It does not represent in all cases the major allele.

1.2.2 Genetic Variants

When a differentiation is observed between two genomes, this differentiation is referred as Variant. Genetic variants are categorized into three categories:

1. **Single Nucleotide Variants SNVs** (or Single Nucleotide Polymorphisms-SNPs)
   A substitution of a nucleic acid in a specific region. This substitution can be either transition, transversion or non-genic and can result to synonymous, non-synonymous, missense and nonsense variants.

2. **Indels**
   Those can be either insertions or deletions, which can range to hundreds of base-pairs.

3. **Structural Variants**
   That kind of variation mostly describes genetic alteration that occurs in a larger DNA sequence. This category includes Chromosomal rearrangements (Deletions, Insertions, Inversions and Duplications) and Copy Number Variations.

While studying genetic diversity, we have to always take into consideration the differences among sex chromosomes and autosomes. Those two "categories" differ in their effective population size, in the mutation and the demographic history, and the action of natural selection. X chromosome is characterized by reduced diversity, that can be explained by demographic events like bottlenecks and the smaller effective population size, because only a single copy is present in males. There are also areas of similar levels of diversity with autosomes, attributed to polygynous mating systems (Ellegren, 2009).

The genetic variants found to affect the gene expression are known as expression Quantitative Trait Loci (eQTLs). Most of those variants are located in regulatory, non-coding regions like enhancers. They have been proposed as great candidates in understanding the impact of genetic variation on complex traits (Nica and Dermitzakis, 2013).

1.2.3 Impact of Genetic Variation

Many studies have been conducted and theories established regarding the effect of variants in populations and their existence. In this section some of them are being analyzed.

Background Selection

Background selection is defined as the reduced nucleotide diversity in neutral loci linked to deleterious mutations, due to negative selection. Neutral variants can be preserved in moderately large populations through generations if they do not coexist with deleterious variants, and thus they are not eliminated from the population through selection (Charlesworth, Morgan, and Charlesworth, 1993). Background selection has been also shown to be a result of recombination events (Hudson and Kaplan, 1995).

Missing Heritability

Risk variants are usually rare in populations and thus their association with complex traits is hard to be identified. Rare variants with huge effect sizes seem to be

responsible for events of "missing heritability". **Missing heritability** is described as the difference in heritability of a trait measured in familial (family-based) studies and the heritability explained by SNVs (Manolio et al., 2009). Recent studies aiming to unravel the patterns of rare genetic variants (minor allele frequency of 0.5%) have unexpectedly identified a higher number of rare variants than those identified under previous studies of human population history. It has been shown that this can attributed to the recent burst of the human population size(Lohmueller, 2014).

Common Diseases and Variants

There have been proposed two models of explaining common disease prevalence, based on the genetic architecture; the **Common Disease-Common Variant (CDCV) model** and the **Common Disease-Rare Variant (CDRV) model**. The main difference of those two models relies on the number and the frequency of alleles in a given 'disease' locus (Schork et al., 2009). The CDCV model posits that genetic variants of high frequency and low penetrance in a population underlie common disease risk. The CDRV model posits that there exist many rare disease variants (Saint Pierre and Génin, 2014). Those two models can be further explained by two distinct evolutionary scenarios. In case of common variants as in the CDCV model, this can either mean that those alleles are not evolutionary deleterious, explaining later-onset diseases that do not affect reproduction efficiency, or they are maintained because of **balancing selection**, due to heterozygous advantage or antagonist pleiotropy. Also, alterations in the direction of selection, and variants characterized as neutral in the past could have been shifted to a pathogenic state (Saint Pierre and Génin, 2014). On the other hand, in the CDRV scenario, the rare variants that contribute to a common disease should be explained by higher mutation rates able to accumulate their putative loss due to selection and drift (Saint Pierre and Génin, 2014). The CDCV hypothesis is now contradicted in many recent studies due to the **missing heritability problem** (Saint Pierre and Génin, 2014).

1.2.4 Genome Wide Association Studies

(Gibson, 2018) (Pearson and Manolio, 2008) Due to the recent development of the **Genome Wide Association Studies (GWAS)**, variations can be further investigated and maybe become strongly correlated with a disease. A GWAS is a genetic study of a genome-wide set of variants with the intention of identifying associations between variants and traits, even complex traits such as diseases (Gibson, 2018).
Briefly, those studies involve 2 groups - one that displays (cases) the studied trait and one that do not display (controls) the studied trait. Through the comparison of allele frequencies between those 2 groups, a higher frequency in affected group can indicate correlation of that specific variation with disease development (Pearson and Manolio, 2008). This design is often affected by population stratification due to admixture and migration events, since cases and controls might accidentally derive from distinct populations that among other traits can also be correlated with the disease(Hirschhorn et al., 2002. We have to note that most of the studies are biased towards protein-coding sequences, that consist only 1.5% of the genome. Although, 76% of GWAS investigated SNPs are located near or in DNase I hypersensitive sites (DHSs), marked as transcriptionally active regions (Schierding, Cutfield, and O'Sullivan, 2014).

1.2.5 1000 Genomes Project

One thousand Genomes Project , ran between 2008 and 2015, aimed to record genetic variants with frequency at least 1% in the populations studied. The participating individuals were 2,504, from 26 populations worldwide. Those populations can be further classified to 5 super-populations based on their geographical distribution: Africa (AFR), East Asia (EAS), South Asia (SAS) and Europe (EUR) (Consortium, 2015).

Firstly, 3 pilot studies were conducted in order to establish the design of the next phases of the project. The main project was subdivided into three phases. Phase 1 included 1,092 individuals and revealed 15 million single nucleotide variants, 1 million short insertions and deletions and 20.000 structural variants. Most of those were firstly described in that phase (Consortium, 2012). There was no publication about the second phase of the project. During the phase 3, the project was considered completed, and the number of the participating individuals reached 2,504 (Sudmant et al., 2015). During that phase 85 million SNVs were discovered, 3.6 million indels and 60,000 structural variants. The reference genome of the study was GRCh37 (hg19)(Sudmant et al., 2015). Sequencing data about variant calls (VCF files), alignment data in BAM or CRAM file format, raw sequence data but also exression RNAseq data, for both mRNA and miRNA (Lappalainen et al., 2013) are available. There are also Cell lines and DNA available from the 1000 Genomes samples used in the study.

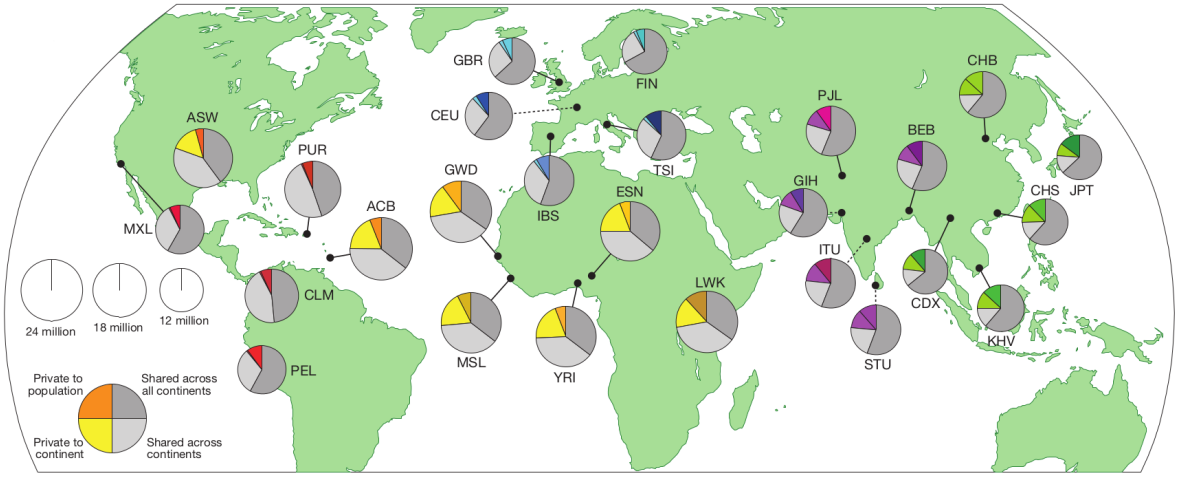

FIGURE 1.1: Occurrences of genetic variants the population studied in the 1000 Genomes Project. Each pie's section represents the number of polymorphisms within a population. Darker colours are indicating of variants only present in a specific population, lighter colours of variants shared in a continental territory. Light grey represents shared variants across continentals and dark grey shared across continents (Consortium, 2015).

1.2.6 ClinVar Database

ClinVar is a freely available database, provided by National Institutes of Health, which reports relationships between medically important human genetic variants and phenotypic traits, like diseases (Landrum et al., 2017, Landrum et al., 2013). It is in union of dbSNP and dbVar, that hold the information about human genome coordinates of the ClinVar Variants. The phenotypic traits are described as proposed by

MedGen. Data from ClinVar can be downloaded in multiple formats like html, XML, VCF and tab-delimited. ClinVar's variants, based on their clinical significance, can be assigned as Pathogenic, Likely Pathogenic, Benign and Likely Benign Variants. This classification is based on guidelines from the American College of Medical Genetics and Genomics and the association of Molecular Pathology (Richards et al., 2015). Each pathogenic criterion, based on the acquisition procedure followed, can be described as very strong (Null Variants), strong (amino acid change, De novo variants, functional studies, variant frequencies in control populations), moderate (e.g. variant in a critical domain) or supporting (segregation analysis). Each benign criterion is considered as stand-alone, strong or supporting (Richards et al., 2015).

Not much work has been performed regarding the importance of Population History in complex traits, and subsequently Mendelian diseases. In order to study the existence of variants in populations several approaches can be applied. Those involve population genetics techniques further discovered bellow. Briefly, the elucidation of LD patterns of variants, Demography, Population diversification and Natural Selection.

1.3 Understanding Linkage Disequillibrium

Linkage Disequillibrium is the non-random association of alleles. Let's assume two loci A and B. The LD measurement in this case is defined as the difference between the observed frequency of those 2 alleles occurring together $\mathbf{P_{AB}}$ and the expected frequency of those two alleles, if their segregation is considered random $\mathbf{P_A}$ $\mathbf{P_B}$ (Ardlie et al., 2001).

$$D = P_{AB} - P_A \cdot P_B \tag{1.1}$$

1.3.1 Patterns of LD across genome

LD in populations is expected to decrease across time and due to recombination distance. There is variability underlying LD and complex factors affecting it. This can be highlighted by the fact that not always adjacent markers are in LD (Ardlie et al., 2001). Moreover, really distant markers have been found to be under LD, either due to selection or due to non-adaptive stochastic processes (Reich et al., 2001).

1.3.2 Factors affecting LD

The major contributors in LD patterns are mutation and recombination. There are additional factors, which are listed bellow:
Genetic Drift. Allele frequency alterations, in every generation, due to the random combination of gametes. This phenomenon has effect on LD mainly in smaller populations.
Variable recombination and mutation rates across the genome and due to biochemical characteristics, like SNVs in CpG dinucleotides, respectively.
Admixture or Migration. LD can be created due to those events. Admixture is described as an event of introducing individuals from a distinct population into another. Gene flow is a consequence of migration.
Inbreeding. LD decomposition is delayed in self-mating populations.
Natural Selection. This phenomenon can effect disequillibrium in two ways. Either through a hitchhiking effect, that a haplotype flanks a favoured variant can be swept to high frequency or fixation or through epistatic selection, which though has yet to be discovered in humans (Ardlie, Kruglyak, and Seielstad, 2002).

1.3.3 LD meassures

In order to meassure LD, the expected and observed frequencies of a haplotype should be calculated and then the difference of those two values is considered as deviation D. Although, since D is dependent of allele frequencies, other measurements have been proposed based on D (Devlin and Risch, 1995). Here, we will only refer to Dprime (D') and r2.

- **D prime**
  The absolute value of D' is calculated by dividing D by its max possible value, given the allele frequencies at the two loci. When $D' = 1$, this is characterized as absolute LD. The extent of D' depends on sample size. Thus, it is difficult for samples to be compared. Consequently, values of D' near 1 should be indicating of recombination disruption but lower values should not be used for comparisons of the LD strength, among samples (Ardlie, Kruglyak, and Seielstad, 2002).
  $$D' = \frac{D}{D_{max}} \tag{1.2}$$
- **r^2**
  It is the correlation of alleles at the two loci and it is expressed by dividing D^2 by the product of the four allele frequencies, at the two loci. When $r^2 = 1$, it means that the alleles have not been disrupted by recombination and also that they have the same allele frequency (Ardlie, Kruglyak, and Seielstad, 2002).
  $$r^2 = \frac{D^2}{p_1 p_2 q_1 q_2} \tag{1.3}$$

1.4 Wright's F-statistics

Wright's F-statistics is a tool of describing genetic diversity, a measure of differentiation within and among populations (Wright, 1931). It includes both F_{ST} (see bellow) but also F_{IS}, which is a measure of deviation of the genotypic frequencies from Hardy-Weinberg proportions. When a locus is under Hardy-Weinberg proportions it means that the diploid genotype of that locus is equal to the expected from a random group of alleles (Holsinger and Weir, 2009). Here we will mostly focus on the Fixation Index (F_{ST}), the measure of population differentiation due to genetic structure. It is the most widely used descriptive statistic in Evolutionary and Population Genetics since it can provide insights for the demographic history of populations, describe a specific locus as 'under selection' and can provide a brief estimation of migration rates (Holsinger and Weir, 2009). It is deciphered in respect to allele frequencies among populations. Small F_{ST} can be translated to equal allele frequencies, while larger F_{ST} means different allele frequencies and consequently differentiated populations (Holsinger and Weir, 2009). The average F_{ST} among human populations is 15%, while in chimpanzee is 32%.
F_{ST} has a large number of formulas. A rather common definition is the following

$$F_{ST} = \frac{var(p)}{p(1-p)} \tag{1.4}$$

The variance of p comes from comparisons of sub-populations and $p(1-p)$ is the expected frequency of the heterozygotes.

1.5 Detecting Natural Selection

1.5.1 Selective Sweeps

Maynard Smith and Haigh in their seminal work in 1974 described **Selective Sweep** as the phenomenon where a beneficial mutation spreads in a population and, subsequently, the frequency of adjacent (weakly selected or neutral) variants will increase as well. Such events can be detected by investigating the pattern of single nucleotide polymorphisms nearby the loci of interest, since a beneficial mutation can result in reduced variation (Smith and Haigh, 1974). There are macro- and micro-evolutionary approaches of detecting such events, with the second one referring to within species events. The variants discovered under those methods are believed to be important for local adaptation. Micro-evolution methods are methods of Population Genetics that are employed to understand human demographic history and evolution (Vitti, Grossman, and Sabeti, 2013). Those methodologies can be further distinguished into either Frequency based, that a variant can be identified due to its higher prevalence in the population, LD-based, that selective sweeps can be identified cause they bring a genetic region that contains both the "causal" allele and adjacent variants in higher prevalence, and Population differentiation based, were selection seems to act on an allele in one population but not in the other (Vitti, Grossman, and Sabeti, 2013).

1.6 Coalescence Theory

Wright-Fischer model considers that all the individuals of a generation release gametes for the next generation and the new individuals are randomly formed by those alleles. The model assumes that the population is constant through time, each individual gets replaced in each generation and that only genetic drift affects allele frequency (Bacaër, 2011). The Coalescence Theory relies on the Wright-Fischer model of forward in time allele frequency change but it examines a sample and the genealogical history of it backward in time. It is a mathematical way of estimating probabilities of genealogies, in populations and describes how the shape of a genealogy of sequences is affected by population genetics events (Kingman, 1982). It is suitable for estimating Population Size and Time Most Recent Common Ancestor (TMRCA) as well as the evolutionary forces acting on populations.

1.7 Background

1.7.1 Principal Component Analysis

Dimensionality Reduction Algorithms
Real world data are usually characterized by high dimensionality. Especially, genetic data where each polymorphic site is a dimension, are characterized often by millions of dimensions. To be able to visualize individuals using their genetic information, it is necessary to reduce the number of dimensions. However, dimension reduction is necessarily accompanied by loss of information. Principal Component Analysis (PCA) allows for dimensionality reduction with the minimal amount of information loss (Van Der Maaten, Postma, and Herik, 2009). **Principal Component Analysis (PCA)** is a linear dimensionality technique. In PCA the data are projected into a lower dimensional linear subspace, that describes as much of the variance as possible. Briefly, the procedure of yielding the directions (Principal Components)

that maximize the variance involves computing of data covariances (standardization), deducing of the eigenvectors by multiplying original data with eigenvalue, re-orientation and transformation. PCA is identical to the metric Multiple Dimensional Scaling (MDS), that uses Euclidean distance (Van Der Maaten, Postma, and Herik, 2009).

1.7.2 Hidden Markov Models

A Hidden Markov Model (HMM) is a statistical model suitable for solving systems that are characterized by the Markovian property (given the current state, the next state does not depend on the past) with unobserved (hidden) states (Baum and Petrie, 1966). It can be used for the characterization of the progression of the observed events, which depend on subjective points. Those points are not visible and they are called *hidden states*, while the observed events are known as *observations*. The transitions and the probabilities in a HMM are stated by arrows (Yoon, 2009). The two main stochastic processes of an HMM are:

- An invisible process of hidden states
- A visible process of observable symbols

The current state affects the hidden states of the Markov Chain and the probability distribution of the observed symbol (Yoon, 2009).

Chapter 2

Methods

2.1 Data used

There are several public resources reporting genetic variation and its impact in the phenotype like HapMap, dbSNp. Here we used ClinVar database, which is a public database, freely available that links SNPs to diseases annotating the severity of the association. We downloaded genetic variation data available from ClinVar from `ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/`. From those, we searched those assigned as Pathogenic, based on the guidelines from the American College of Medical Genetics and Genomics and the association of Molecular Pathology (Richards et al., 2015). We used phase 3 genotypic information of those individuals, downloaded from the 1000 Genomes Repository. The variant coordinates were mapped on GRCh37 reference genome.

2.2 Scripts and Pipelines

The scripts and pipelines used during this analysis can be found either in the GitHub repository (`https://github.com/annamath/MSc2018_PopGenOfDiseaseVAriants`) and in the Appendix provided A. During this thesis, a variety of softwares was used. Those are further analyzed. The table bellow represents represents a list of those and their reason used.

TABLE 2.1: Table of Softwares used

Softwares	Purpose
VCFtools	F_{ST} and VCF file processing
tabix	VCF file processing
ms	Population Simulations
PLINK	LD Calculation
SweeD	Detection of Selection Sweeps
MSMC	Demographic Inference

2.3 LD analysis

2.3.1 PLINK

PLINK is an open source genome association analysis package, that performs large - scale analyses. PLINK focuses on analysing geneotype/phenotype data (for GWAS) and not for preprocessing steps like CNV calls from raw data. Here we used PLINK to calculate LD for a list of variants, from a VCF file.

2.4 F_{ST}

2.4.1 VCFtools

VCFtools is a tool set designed for workjing with VCF files, like those used in our analyses. We used VCF tools for 2 purposes: **F_{ST} estimation** is performed based on Weir and Cockerham's 1984 work. The required input files must contain lists of individuals, which correspond to the populations compared. The calculations by default are performed at per-site basis. **Subsetting and slicing of VCF files**, in a specified genomic region (using tabix), for specific individuals.

2.5 Simulating Populations

2.5.1 ms Software

We used ms program Hudson, 2002 in order to generate datasets under a variety of neutral models. We chose ms software for our simulations since it is able to simulate neutral demographic models with migration given demographic parameters and the mutation and the recombination rate. In this work we simulated migration models. The purpose of those simulations was to test the performance of the MSMC software (more details bellow), in scenarios of migrating populations.

2.6 Assessing the Haplotype Population History

2.6.1 MSMC

We used Multiple Sequentially Markovian Coalescent (MSMC) in order to perform the Inference of Population's History (Schiffels and Durbin, 2014). MSMC uses an HMM model that exploits the density of the heterozygous sites and manages to infer branch lengths and coalescent times. The hidden states of the model are: the 1^{st} coalescent event, the total length of the tree's singleton branches and the identities of the two participating sequences. It does not provide point estimation of the tMRCA but it provides cross-coalescence rate, which measures relative gene-flow. Thus, when the gene flow is equal to zero, that is the point of common ancestry. MSMC allows the estimation of effective population size, population separation history and TMRCA.

For each variant, we used sequences of 250,000 bases that contained the variant of interest in the middle-point, for each individual separately, using time segment patterning of 1*2+25*1+1*2+1*3.

It is important to note that among the different methods in bibliography about TMRCA calculation, we chose MSMC based on the consistent findings presented in Zhou and Teo, 2016. It provides a really valid estimation of TMRCA, necessary for population size inference.

2.7 Diversification of Haploid Sequences

We performed PCA (method detailed above) in order to test whether we could determine the genetic ancestry of the haplotypes which contained the pathogenic variant. The sequences were converted from VCF format to ms, using an inhouse perl script. Each haploid sequence was projected separately. In the plots shown here we have projected the first two components, that explained the most variance.

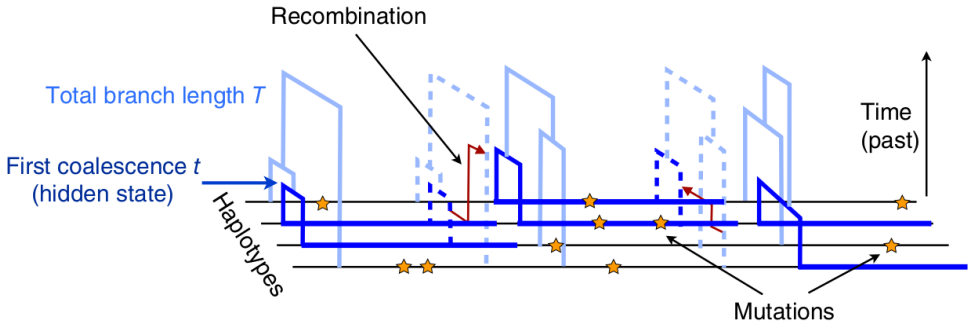

FIGURE 2.1: MSMC representation of the model.
Recombination can change local genealogies along the sequences. Few mutations on branches are indication of recent coallescent event. The hidden states of the model are the time of the first coallescent and the identity of the participating sequences to that event.

2.8 Detection of Selective Sweeps

We localized the action of positive selection by detecting selective sweeps in order to explain pathogenic variants' existence in healthy individuals, especially the most frequent ones. We hypothesized that a beneficial allele that has reached fixation, due to natural selection might be linked in pathogenic variants (hitch-hiking effect). This is a plausible mechanism that could explain their existence in populations.

2.8.1 SweeD tool

For Selective Sweeps detection we chose Sweep Detector (SweeD), a likelihood-based tool for detecting sweeps in whole genomes, through analyzing Site Frequency Spectra of Single Nucleotide Variant frequencies in a given sample Pavlidis et al., 2013.
Our goal was the investigation of selective sweeps events in the haplotypes that carried the pathogenic variants. Thus, we ran the SweeD algorithm in a range of 1 million base pairs, adjacent to the variation of interest. Within that locus, we chose to test for that kind of events every 1000 base pairs. The likelihood threshold above of which a variation was considered as under Natural Selection was calculated by creating null distributions from random positions of the whole chromosomes. To calculate the threshold we applied the following methodology. We ran SweeD for whole chromosomes, setting the grid parameter to 10,000. After calculating the 10,000 likelihoods (one for each grid-point), we created 1.000 random datasets of 1.000 randomly sampled grid-points. From each dataset we sampled the maximum value. Those values consisted the points of the null distribution. The threshold was set at the 95% max of the distribution, above of which we denoted a loci as positive for selection.

Chapter 3

Results

3.1 Pathogenic Variants present in Healthy Populations

From the 392,929 variants present in the ClinVar Database, 57,731 of those were characterized as Pathogenic due to the guidelines of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (Richards et al., 2015). From those, only 1,690 were found present in the 2,504 healthy individuals of the 1000 Genomes project. Specifically, 651 of those were found in Europeans, 526 in African samples, 456 in Americans, 420 in South Asians and 411 in East Asians. In addition, 924 out of 1,690 were singletons (meaning present in only one genome), 227 doubletons, 110 tripletons while some of the variants were found in really high frequency. For instance, rs4784677 was present in all individuals (either in homozygous or heterozygous state), rs6025 in 2,503, rs820878 in 2,501 etc. Moreover, 107 of the mapped variant were characterized as pathogenic for Acute Myeloid Leukemia with Maturation. Five of those AML variants (rs10800598, rs6063971, rs10800597, rs3735819, rs2815822) belong to the 10 most common with frequencies that range from 2,371 individuals to 2,476.

We mapped those variants on chromosomes, using the hg19 reference genome (Figure 3.1). Interestingly, we noticed the absence of strongly correlated variation with pathogenicity in both sex chromosomes. This is something that could implicate severity of symptoms when it comes sex related pathogenesis. Variants were also absent in most of the cases from chromosome centromeres. Moreover, approximately 10% (180/1690) of the variants mapped seem to be strongly correlated with Accute myeloid leukemia with maturation.

TABLE 3.1: Number of pathogenic variants present in each population

Population	Counts
EUR	651
AFR	526
AMR	456
SAS	420
EAS	411

3.2 Linkage Disequilibrium of Pathogenic Variants

We used PLINK software for the pairwise estimation of LD among variances variants. For that we used both r^2 and D′ measurements. No differences were noticed among those two approaches (Figure 3.2).
In most of the cases, pairwise LD is close to zero with some extreme positive values

(e.g. in chromosomes 1, 3, 11, 16). This is an unexpected result since those variants have been implicated to disease.

3.3 Estimating Variant Origin using PCA

We performed PCA analysis to estimate origin of variation and also to test the possibility of haplotype grouping due to variant existence. We left out singleton variants. Bellow we are commenting on some of the variants that displayed interesting behavior. The PCA plots are shown in Figure 3.3. For all the variants tested, that kind of sequence distinction was only observed in ***rs9660525 and rs12406470***. Those two variants were identified in all populations, and formed "groups" of pathogenic-non pathogenic sequences. Moreover, ***rs12406470*** is not present in many AFR individuals. ***rs1008642 and rs1137617*** are characterized as mixed, since they are found in all populations, displaying no pattern of distinction. ***rs429358*** is also present in multiple populations, shaping sequence "clusters" but this does not seem to be related to variant existence. ***rs185790394 and rs11570112*** can be classified as single origin, that have recently arise variants since they seem to be isolated within a single population.
We chose to continue to the next analysis with the variants mentioned above.

3.4 F_{ST} measurements of Pathogenic Haplotypes

The results discussed in this sections are displayed in Figure 3.4. We performed estimations of F_{ST} calculations between the individuals who had a variant and the null individuals of each superpopulation, in haplotypes of 250,000 bases long. We worked on distinct populations in order to avoid false-positive results due to among populations variation.

For the variants studied we observed in some cases really low F_{ST} values, close to zero, indicative of no differentiation among the cases and null individuals ***(rs1008642, rs1137617)***. This could be a consequence of increased recombination events in that locus.
In most of the cases ***rs12406470, rs9660525, rs429358, rs185790394*** our results indicated moderate diversification among our two groups studied, with the most variation occurring in the middle of the haplotype (i.e. in the coordinate of the pathogenic variant studied) and then average F_{ST} values, around 0,5 to 0,5, which do not indicate neither population diversification nor population resemblance.

In the case of ***rs3120649***, we noticed difference between AFR and EAS populations. In the case of AFR populations the "pathogenic" haplotypes showed average diversification, while in the EAS the values were close to zero. In the case of rs16904774 we observe increased F_{ST} values.

3.5 Demographic History of Haplotypes

For the demographic inference, as mentioned above we used MSMC tool.

The list of individuals and populations we performed those experiments on are listed in AppendixB. Regarding the original data, ***rs34526199*** we performed MSMC analysis for AMR, EUR, SAS samples. We observed no differences among the distinct populations. The history we can assume looks like a bottleneck and a recent population expansion (Figure 3.5).

In the cases of ***rs429358 and rs11570112*** we can assume a stable demographic past and a recent population growth as well (Figure 3.6, Figure 3.7).

Recent increase in effective population size could either mean migration and thus gene flow between population or it could be attributed to the recent explosion in human population size. Thus, that kind of increase is not directly related to the demographic history of a specific haplotype. An MSMC analsysis on migrating simulated data was performed, using ms software. Interestingly, the outcome of that analysis reinforces the idea of migration events, since a similar population "explosion" is observed (Figure 3.8 and A).

3.6 Explaining Pathogenic traits with selective sweeps

In order to explain the existence of Pathogenic variants in healthy populations, we investigated the existence of selective sweep phenomena in adjacent areas of the Pathogenic Variants, except for the singletons. From the 700 haplotypes we checked, **111** were found located in haplotypes positive for selection events. We hypothesized that those under selective events would be present in higher frequencies. Though, we found no significant difference among the frequencies in populations among the group of variants' adjacent loci under positive selection and those not under positive selection.

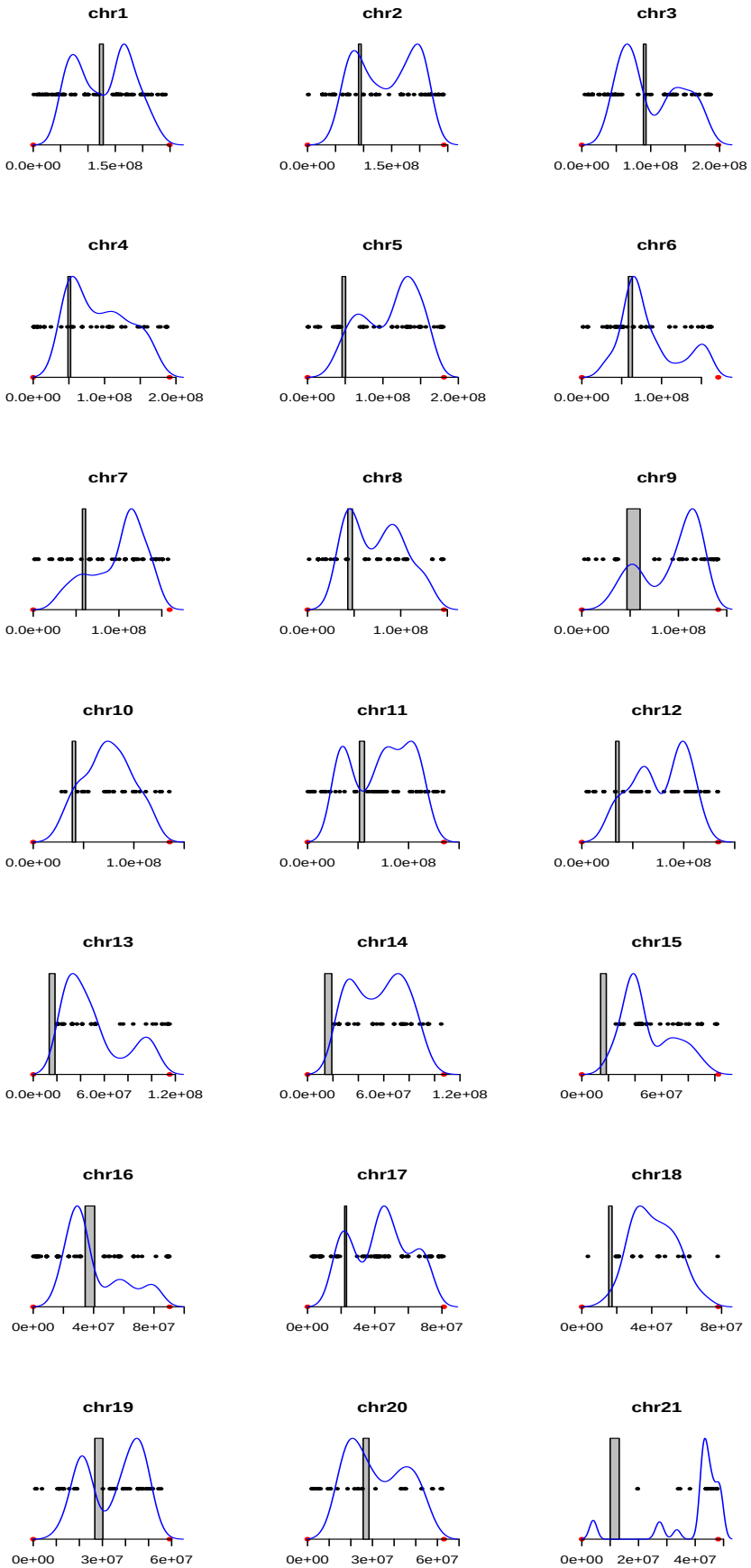

FIGURE 3.1: Distribution of 1.690 Pathogenic Variants found in healthy populations. The gray line represents the centromeres' positioning. The black dots are the indicate the variant occurences across chromosomes. No variants were mapped on Sex-Chromosomes and on Chromosome 22.

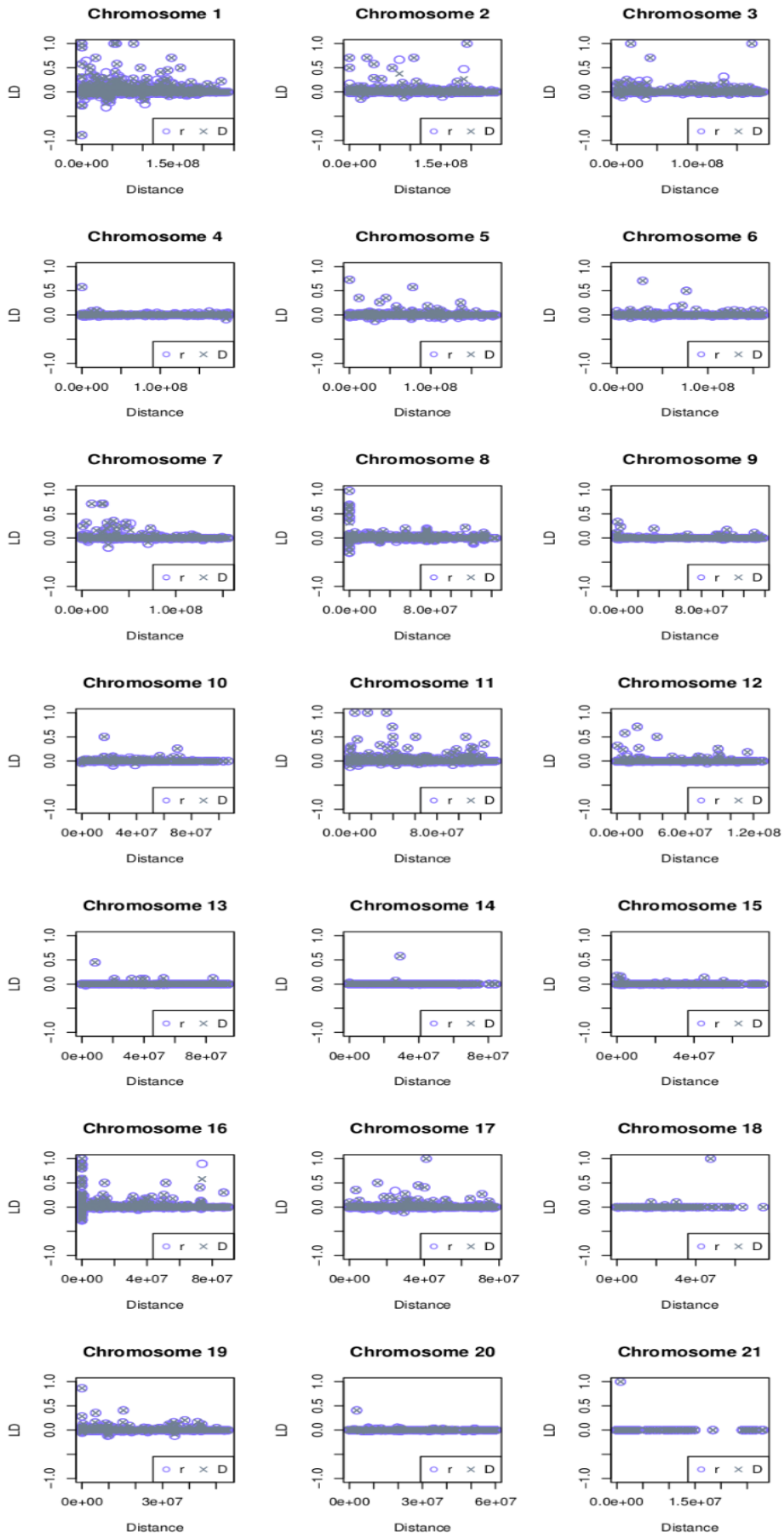

FIGURE 3.2: Pairwise Linkage Disequillibrium estimates concerning variants' distance, using PLINK software. In most of the cases, LD is close to 0. There are cases where LD is increased. LD is calculated using both r^2 and D'. The results produced in both cases are consistent.

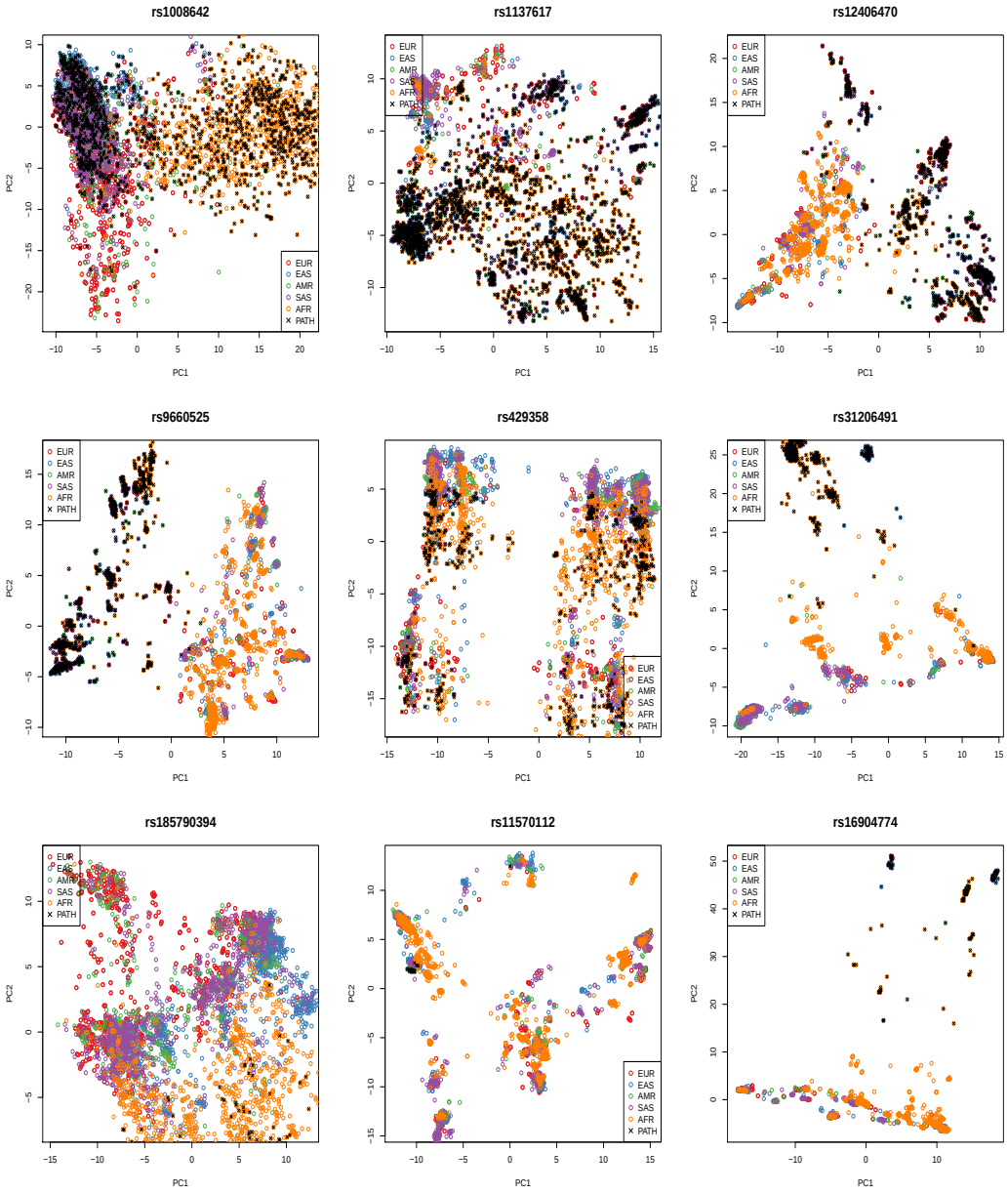

FIGURE 3.3: PCA for Haploid Pathogenic Sequences, among the 5 super-populations studied. Different diversification is observed among pathogenic variants. The axes are representative of the two first principal components.

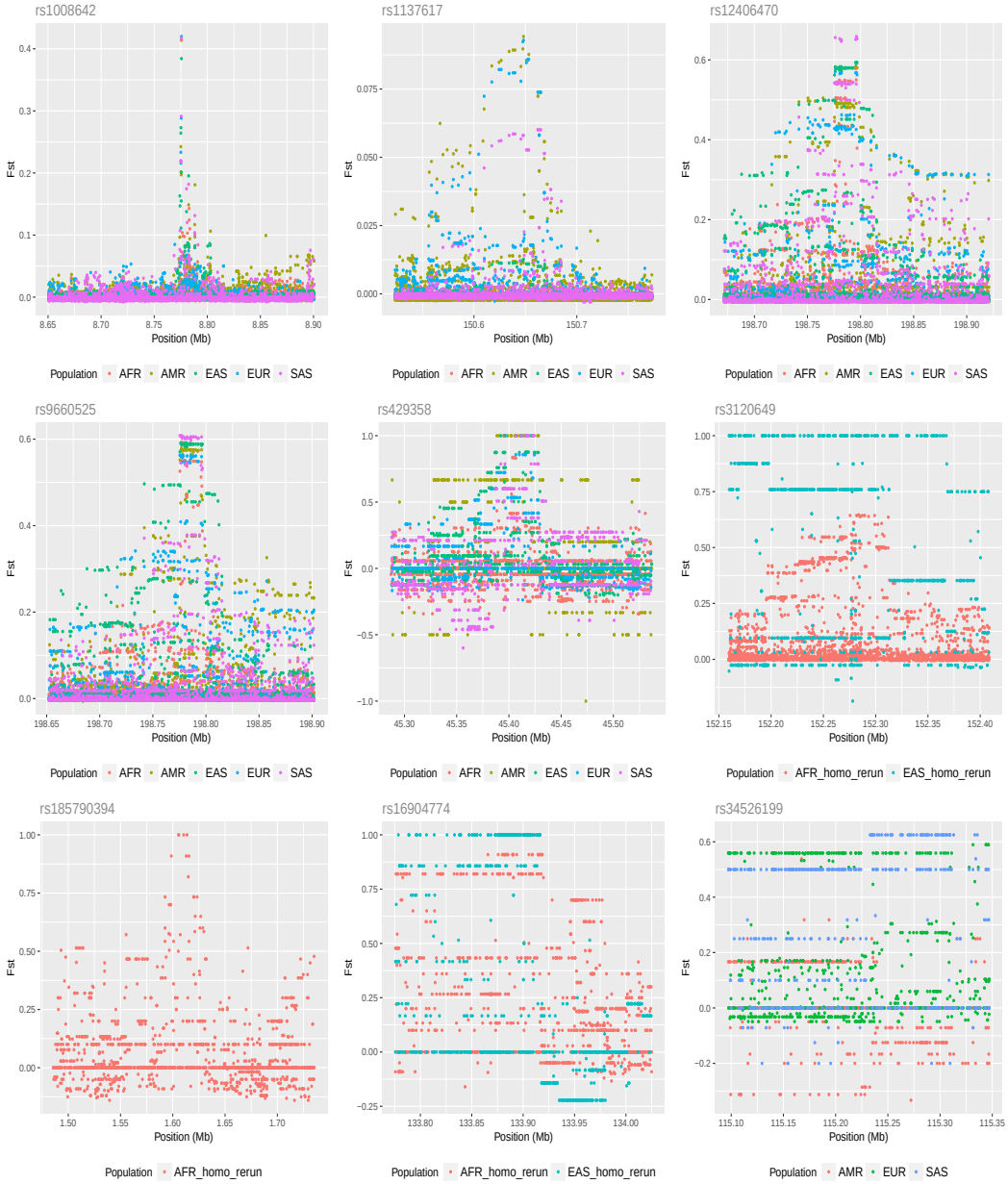

FIGURE 3.4: F_{ST} measurements in haplotypes which contain pathogenic variants. x axis represents the positions in base-pairs, while the y axis the F_{ST} measurements. No common pattern of "population" structure is observed.

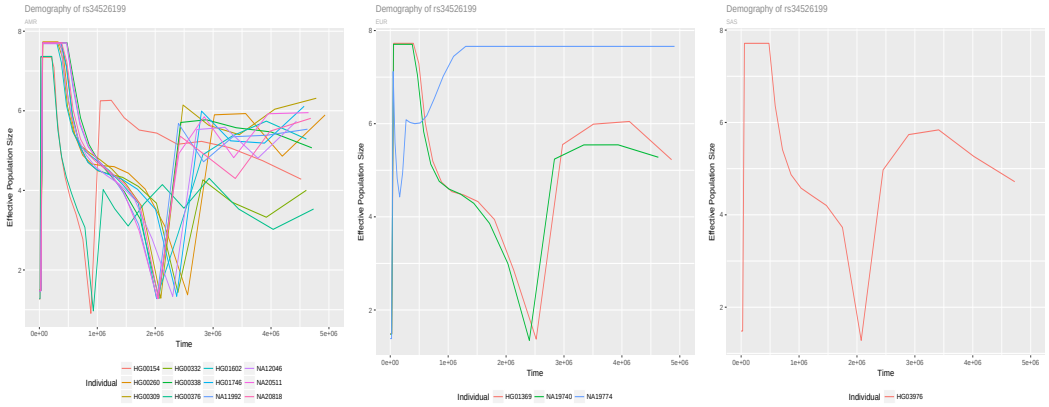

FIGURE 3.5: MSMC output for haplotypes of 250.000 bases which contain rs34526199 for AMR, EUR and EAS populations No difference in the effective population size among populations is observed

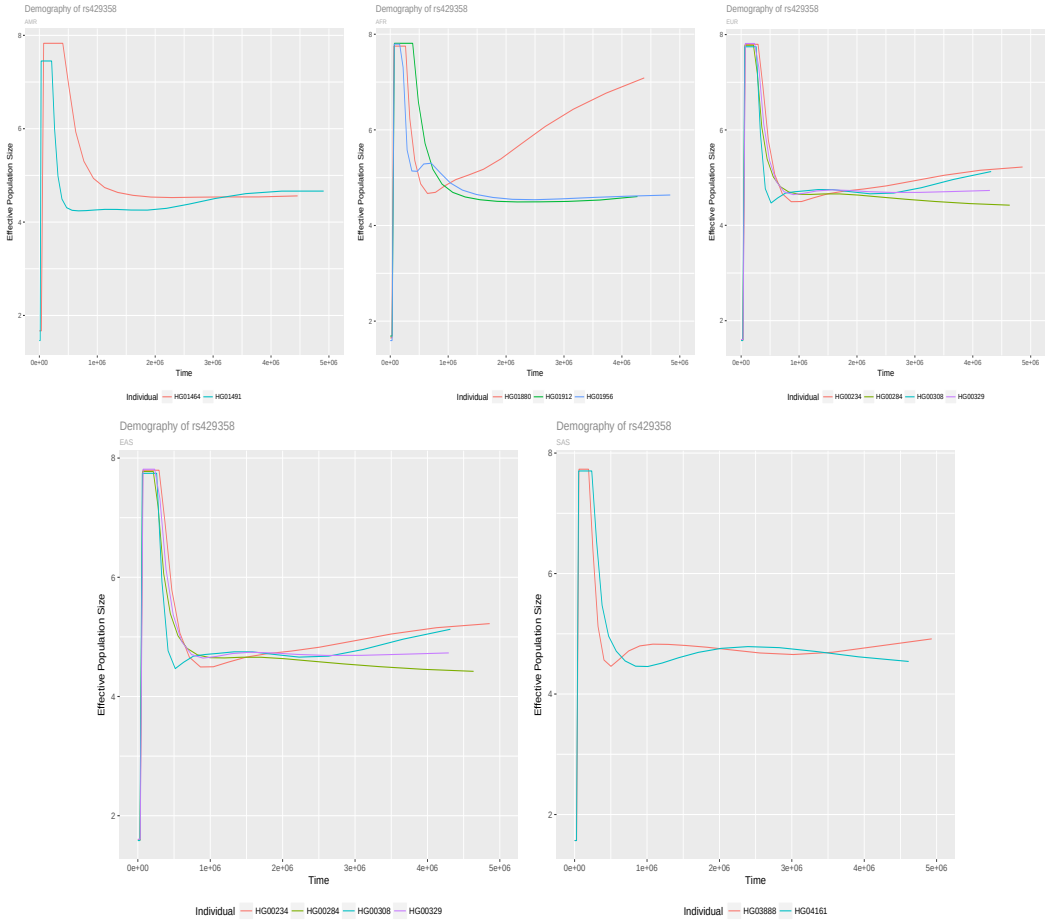

FIGURE 3.6: MSMC output for haplotypes of 250.000 bases which contain rs429358 for AMR, AFR, EUR, EAS and SAS populations. No difference in the effective population size among populations is observed.

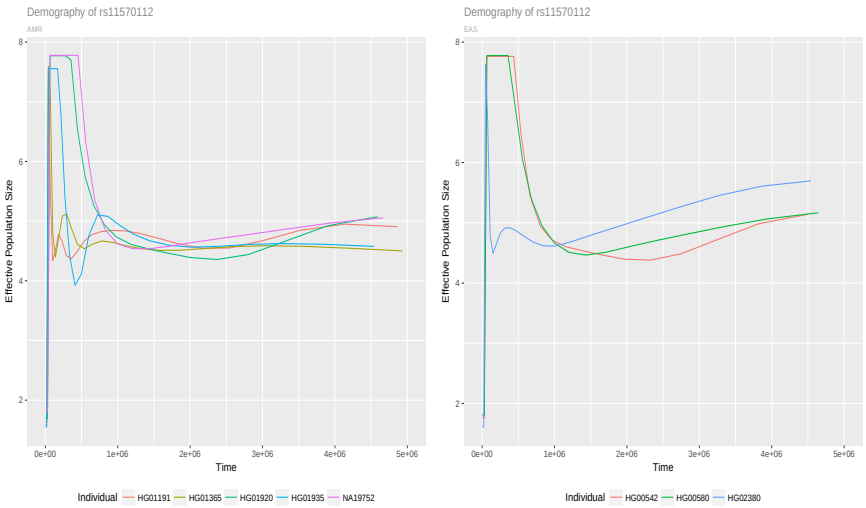

FIGURE 3.7: MSMC output for haplotypes of 250.000 bases which contain rs11570112 for AMR and EAS populations. No difference in the effective population size among populations is observed.

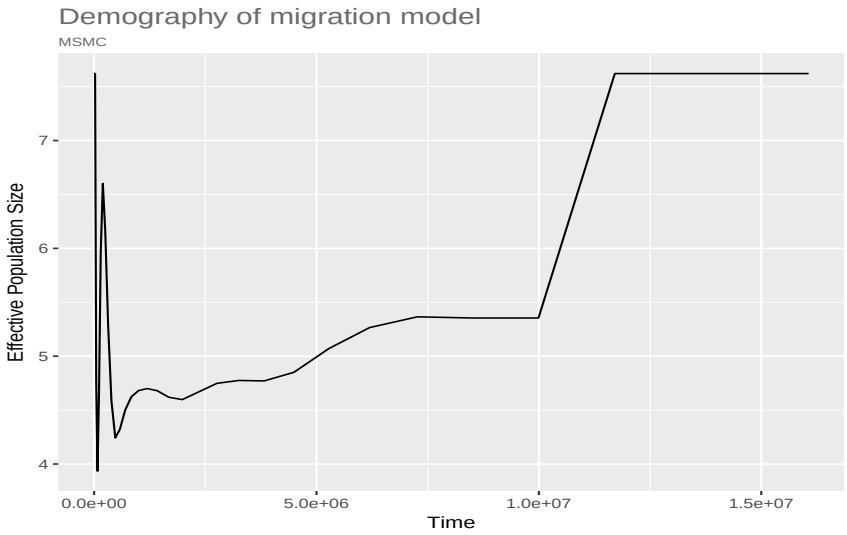

FIGURE 3.8: MSMC Demographic Estimation on simulated migration data, using ms software

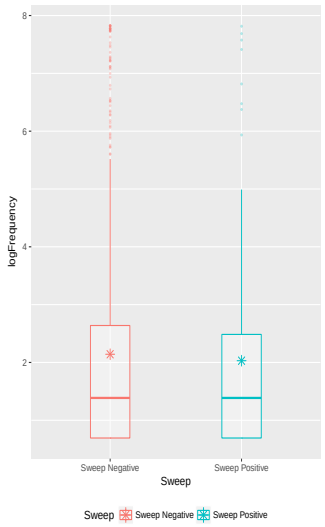

FIGURE 3.9: The variants in haplotypes positive for selection do not exhibit higher population frequencies than those which are negative.

Chapter 4

Discussion-Future Goals

Here we present the methodology we followed in order to explain the existence of variants associated with diseases in healthy individuals and we examined the properties of their genomic neighborhoods in terms of population genetics related statistics. For this, we employed Population Genetics methodologies, considering in some of the cases (e.g. Demographic inference analysis) the haplotypes as individuals in order to trace the course of a specific haplotype and not of the whole individual. Different genome loci might have different origin. For example, there are extreme cases of viruses integration in the human genome (Keane, Wong, and Adams, 2012).

In all the estimations of the demographic inference, we observed a recent population growth, followed by rapid population decline.

Our first observation can be attributed in the recent explosion of Human Population, a phenomenon also responsible for the excess of rare variants observed in humans(Keinan and Clark, 2012). The population decline suggests two plausible explanations. Either the variants became recently pathogenic thus the effective population size of their flanking neighborhood was shrunk recently or the observed past expansion can be better explained by inter-population migration events. Ancestral migration events increase the coalescent time, thus they effectively resemble an increase in the population size.

The ms simulations we performed on data with migration are concordant to our last observation. Indeed the analysis of ms data, without selection events, suggests that a similar increase followed by a population decline of the inferred ancestral population size can be observed when gene flow has occurred between two different populations, supporting that the studied variants were introduced in a population because of migration events.

Even if the variants tested are characterized as disease variants, meaning deleterious for the population, we did no observe any shared genetic variability pattern in the containing haplotypes. Thus, we cannot attribute any common possible effect on the genotype. In the case of *rs16904774*, the containing sequences are distant from the rest of individuals and this finding is also supported by the increased F_{ST} values, which implies population differentiation. In the cases of mixed populations, the F_{ST} estimations around that loci were low, indicating that little variation can be attributed to the existing variant. One of the variants studied is the *rs429358*, a risk allele for Alzheimer's disease development (Mez et al., 2017). This variant is well represented among different populations, suggesting older origin, which is also supported by the estimations of effective population sizes (Figure 3.6) and by the population diversification (Figure 3.4), where *rs429358* haplotypes display average variation among cases and controls.

We observed that pathogenic variants are absent from the centromeres and from adjacent areas, a result consistent to the reduced genetic variation in centromeres due to background selection, first described by Charlesworth, Morgan, and Charlesworth,

1993. Another possible explanation is that the quality of NGS data in the proximity of the centromeres is poor, thus the annotation of genetic variants nearby the centromere should not be trusted. The absence of disease variants from sex chromosomes can be attributed to the reduced effective population size of both X and Y chromosomes (Ellegren, 2009). That is also an indication of the severity of sex chromosomes-linked diseases (Arnold, 2017, Rogers and Shapiro, 1986) ; when a disease mark occurs in a sex chromosome, it is more likely to develop a disease and thus not be considered as a healthy individual.
When working with genetic markers that have effect in pathogenesis, it is expected those variants should not be co-inherited, thus to be under non-positive LD. Most of the pairwise LD measurements were around zero, with some cases of strong LD even in distant variants but also negative LD in neighboring variants (Reich et al., 2001, Ardlie et al., 2001). The uncertainty while predicting LD and the variability highlights the complexity of the factors affecting LD.
We pointed out a weakness of the PLINK software we used for LD calculations, that identified two variants to be under negative LD, *rs9660525* and *rs12406470*. Those two variants are located in neighboring genetic locations in the first chromosome and they are both characterized as 'responsible' for Acute Myeloid Leukemia with maturation. Though, they are not under negative LD.
It is important to note that 1000 genomes is not considered as the most suitable choice for the study of rare low-allele frequency variants. This is the case for some of the variants analyzed in this work. For that reason, it might be more appropriate to work with data available from gnomAD or TOPMed databases. Although, in those two cases the phenotypic state of the participating individuals is not clear and the required genotypic information for our analysis is not available. It would be also recommended to work with projects like 10.000 Genomes project, recently released by Craig Venter's research team (Telenti et al. (2016)).
The investigation of selective sweeps events, and thus of selection, in loci where a pathogenic mark is present is rather helpful in the understanding of their presence in healthy individuals. A phenomenon like that would explain the preservation of such a trait in the population, in high frequencies. Thus, we are planning to further examine selective sweep patterns by running Omega Plus, an LD based approach to estimate selective sweeps (Alachiotis, Stamatakis, and Pavlidis, 2012).
The value of detecting genetic variation is small, if it is not accompanied by the study of its possible effects on a biological pathways. We aim to investigate the impact of an observed variation in gene expression, by focusing on variants discovered to be under selective sweep (Stranger et al., 2012). For that we will investigate gene expression differences among individuals "positive" and "negative" for specific disease variants, by utilizing expression data available for the 1000 Genomes individuals (Lappalainen et al., 2013).

Appendix A

Commands and Pipelines

A.1 PLINK

All the analyses are applied with the commands below. In the first case the LD calculation was based on r while on the second case on Dprime.

```
plink --vcf <vcf-file> --r inter-chr --ld-snp-list
plink --vcf <vcf-file> --r dprime inter-chr --ld-snp-list
```

A.2 VCFtools

All the F_{ST} analyses were applied with the commands bellow.

```
vcftools --vcf <vcf-file> --weir-fst-pop <samples1.txt> --weir-fst-pop <samples2.txt>
```

The slicing of VCF files for a specific chromosomal range and for specific individuals was performed using the command bellow.

```
tabix -h <vcf_file> chr:start-end |  vcfsubset -c <samples> | bgzip -c > output.vcf.gz
```

A.3 SweeD

For the SweeD analysis, on whole chromosomes the following commands were used:
Creation of osf files, for whole chromosomes:

```
SweeD -input <1KG_vcf_file> -osf <file_name> -name <name_of_run> -folded
```

SweeD runs for whole chromosomes:

```
SweeD -grid 20000 -name <name_of_run> -input <osf_file> - folded
```

For specific genetic regions:
GridFiles, which contain information about chromosome and position, were created by GridFileCreator.py script.

```
Sweed -grid 1000 -gridFile -name <rs_ID> -input <whole_chromosome_vcf> -threads 4
```

`-folded` argument considers the SFS folded, meaning that the ancestral and the derived states of the alleles cannot be distinguished.

A.4 ms

We used ms program in order to generate data with population parameters:

A.4.1 A Brief Guide to ms Software

- `-r p sites`
  Sites are normally considered as the haplotype length plus 1 but since it would not be computationally effective we chose values between 1000 3000.
  ρ is described by the equation bellow:

$$\rho = 4 \cdot N_e \cdot r \cdot l \tag{A.1}$$

  l denotes for the length of the simulated genotype in bases, and $\mathbf{N_e}$ for the effective population size. The effective population size of humans is 10^4

- `-t`
  The **t** parameter is the mutation parameter θ:

$$\theta = 4 \cdot N_0 \cdot \mu \tag{A.2}$$

  Where N_0 is the diploid population size and where μ is the neutral mutation rate for the entire locus.

- `-e`
  This parameter is used in order to specify past demographic changes.
  `-em t i j x`
  Where the migration matrix M_{ij} is consisted of elements $4N_0m_{ij}$.
  `i, j = 1, ...` n_{pop} and m_{ij} is the portion of subpopulation `i` which came from migrants of `j` subpopulation, each generation.
  `-ej t i j`
  This argument is used in order to move all lineages from subpopulation **i** to subpopulation **j** at time **t**. Growth rates of populations are stable. This scenario corresponds to population splitting.

- `-I npop` n_1 n_2 ...
  This argument produces samples under island models and it is followed by the number of subpopulations (npop) and a list of integers (n_1 n_2 ...) which indicate the number of chromosomes exchanged among populations.

A.4.2 Scenarios Simulated

Migration Effect Model
For the migration model we ran the following command:

```
ms 4 10 -t 250 -r 100 1000 -I 2 4 0 -em 0 2 1 2 -ej 2 1 2
```

A.5 MSMC

A.5.1 A Brief Guide to MSMC

```
msmc2 -t <threads> -p <time segmentation patterning> -I <haplotypes> -o <output prefix> --sk
```

`-o` Output prefix.
`-I` Starting from the index of zero, it denotes the haplotypes to be analyzed. When working with two haplotypes `-I 0,1` etc.
`-p` Defines time segmentation patterning. The default is `-p 1*2+25*1+1*2+1*3`
`-m` This is the scaled mutation rate. In case no mutation rate is given, the tool uses watterson estimator in order to determine θ (Watterson (1975)).

`--fixedRecombination` When working for only one individual (thus, two haplotypes) it is recomended to skip this flag.
`--skipAmbiguous` This flag is proposed to be used while assessing gene flow. That way, sites with unclear phasing are removed from the analysis.

A.5.2 Generation of MSMC input files

During standardizing our MSMC methodology, we used ms output data.

The generatemultihetsep.py and the ms2multihetsep.py scripts are provided by the authors of the MSMC software. Those script are used for the generation of multihetsep files, which are the input files for MSMC.

For the simulated datasets, the following commands were used:

```
ms2multihetsep.py <ms output file> <chrom> <length>  > output.multihetsep.txt
```

A.5.3 Testing Time Segmentation

Different time segmentation **-p** patterning parameters were tested, in datasets of 2 samples.

```
msmc2 -t 4 -I 0,1 -p 1*2+25*1+1*2+1*3 -o <output> <input>
msmc2 -t 4 -I 0,1 -p 1*2+15*1+1*2 -o <output> <input>
msmc2 -t 4 -I 0,1 -p 10*1+15*2 -o <output> <input>
```

The command used for the `msmc2 -t 8 -I 0,1 -p 1*2+25*1+1*2+1*3 -o <output> <input>`
For the original datasets the following commands were used:

```
tabix -h <vcf_file> chr:start-end | vcfsubset -c <sample> | bgzip -c > output.vcf.gz
generate_multihetsep.py --mask <strict_mask_file> output.vcf.gz > output.multihetsep.txt
msmc2 -t 8 -I 0,1 -p 1*2+25*1+1*2+1*3 -o <output> <input>
```

A.5.4 Analyzing MSMC output

The output files of MSMC runs contain scaled times.
While plotting, we convert the scaled times to **generations** by dividing scaled times with mutation rate. The mutation rate is set to **mu = 1.25 e^{-8}**, since we are working with humans.

$$generations = \frac{output\ scaled\ times}{mutation\ rate} \tag{A.3}$$

The conversion of generations to years is performed by multiplying generations by per generation years. Here, we set generation years equal to 30.

$$years = \frac{(output\ scaled\ times) \cdot\ (generation\ years)}{mutation\ rate} \tag{A.4}$$

Appendix B

Individuals of MSMC runs

TABLE B.1: Table of individuals and corresponding populations on which we ran MSMC analysis.

Variant	AMR	AFR	EUR	EAS	SAS
rs34526199	HG00154, HG00332, HG01602,NA12046, HG00260, HG00338, HG01746, NA20511, HG00309,HG00376, NA11992, NA20818	-	HG01369, NA19740, NA19774	-	HG03976
rs429358	HG01464, HG01491	HG01880, HG01912, HG01956	HG00234, HG00284, HG00308, HG00329	HG00234, HG00284, HG00308, HG00329	HG03888, HG04161
rs11570112	HG01191, HG01365, HG01920, HG01935, NA19752	-	-	HG00542, HG00580, HG02380	-

Appendix C

Allele counts of SweeD positive variants

TABLE C.1: Allele counts of Variants assigned as under selective sweeps events, in individuals from the 1000Genomes Project

2	rs121434391, rs121908171, rs121908288, rs121908958, rs121912697, rs121912992, rs121918028, rs121918138, rs137852947, rs137852987, rs137852990, rs137891647, rs140950220, rs142609245, rs145541911, rs150766139, rs183261547, rs183589498, rs185392267, rs190834116, rs200214298, rs201114717, rs201230446, rs201284672, rs202073531, rs202138550, rs202145681, rs267606647, rs28940574, rs368263958, rs377619732, rs534438354, rs541299023, rs61753245, rs63750330, rs63750783, rs63750959, rs78290141, rs80338688, rs80338794
3	rs104894408, rs104895503, rs121912763, rs121912863, rs139517732, rs139620139, rs142724470, rs142800871, rs149712114, rs200837270, rs541873609, rs550921485
4	rs104886488, rs121434556, rs121909192, rs121912762, rs138326449, rs149989682, rs1800028, rs181690344, rs34993780, rs536522394, rs574552037, rs6475, rs80358198
6	rs13306515, rs142328166, rs28936395, rs5122
7	rs121913016, rs80338660
8	rs121908326, rs200908035, rs56307355
9	rs121918677, rs547709692, rs6445
10	rs121918426, rs28940885
11	rs121909075, rs35717904, rs73015965
12	rs144648002, rs35882952
13	rs115746363;rs564976220, rs121964926
14	rs201899866
16	rs59513011
19	rs74315310
20	rs61730328
21	rs121908627
23	rs28730837
24	rs187930476
37	rs769455
48	rs61761068
70	rs6711223
77	rs76863441
79	rs115079861
87	rs121964976
93	rs532781899
105	rs3212989
140	rs1800937
148	rs185790394
381	rs34116584
596	rs11539445
655	rs10509305
920	rs1064039
1669	rs1131695
1965	rs3755319
2201	rs6467
2498	rs2815822

Bibliography

Alachiotis, Nikolaos, Alexandros Stamatakis, and Pavlos Pavlidis (2012). "OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets". In: *Bioinformatics* 28.17, pp. 2274–2275.

Ardlie, Kristin et al. (2001). "Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion". In: *The American Journal of Human Genetics* 69.3, pp. 582–589.

Ardlie, Kristin G, Leonid Kruglyak, and Mark Seielstad (2002). "Patterns of linkage disequilibrium in the human genome". In: *Nature Reviews Genetics* 3.4, p. 299.

Arnold, Arthur P (2017). "Y chromosome's roles in sex differences in disease". In: *Proceedings of the National Academy of Sciences*, p. 201702161.

Bacaër, Nicolas (2011). "Wright and random genetic drift (1931)". In: *A Short History of Mathematical Population Dynamics*. Springer, pp. 105–109.

Baum, Leonard E and Ted Petrie (1966). "Statistical inference for probabilistic functions of finite state Markov chains". In: *The annals of mathematical statistics* 37.6, pp. 1554–1563.

Bekris, Lynn M et al. (2010). "Genetics of Alzheimer disease". In: *Journal of geriatric psychiatry and neurology* 23.4, pp. 213–227.

Charlesworth, Brian, MT Morgan, and Deborah Charlesworth (1993). "The effect of deleterious mutations on neutral molecular variation." In: *Genetics* 134.4, pp. 1289–1303.

Chong, Jessica X et al. (2015). "The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities". In: *The American Journal of Human Genetics* 97.2, pp. 199–215.

Consortium, 1000 Genomes Project et al. (2012). "An integrated map of genetic variation from 1,092 human genomes". In: *Nature* 491.7422, p. 56.

— (2015). "A global reference for human genetic variation". In: *Nature* 526.7571, p. 68.

Devlin, B and Neil Risch (1995). "A comparison of linkage disequilibrium measures for fine-scale mapping". In: *Genomics* 29.2, pp. 311–322.

Ellegren, Hans (2009). "The different levels of genetic diversity in sex chromosomes and autosomes". In: *Trends in Genetics* 25.6, pp. 278–284.

Gibson, Greg (2012). "Rare and common variants: twenty arguments". In: *Nature Reviews Genetics* 13.2, p. 135.

— (2018). "Population genetics and GWAS: A primer". In: *PLoS biology* 16.3, e2005485.

Griffiths, Anthony JF et al. (2000). "Quantifying heritability". In:

Hirschhorn, Joel N et al. (2002). "A comprehensive review of genetic association studies". In: *Genetics in Medicine* 4.2, p. 45.

Holsinger, Kent E and Bruce S Weir (2009). "Genetics in geographically structured populations: defining, estimating and interpreting F ST". In: *Nature Reviews Genetics* 10.9, p. 639.

Hudson, Richard R (2002). "Generating samples under a Wright–Fisher neutral model of genetic variation". In: *Bioinformatics* 18.2, pp. 337–338.

Hudson, Richard R and Norman L Kaplan (1995). "Deleterious background selection with recombination." In: *Genetics* 141.4, pp. 1605–1617.

Husemann, M et al. (2016). *Effective population size in ecology and evolution*.

Keane, Thomas M, Kim Wong, and David J Adams (2012). "RetroSeq: transposable element discovery from next-generation sequencing data". In: *Bioinformatics* 29.3, pp. 389–390.

Keinan, Alon and Andrew G Clark (2012). "Recent explosive human population growth has resulted in an excess of rare genetic variants". In: *science* 336.6082, pp. 740–743.

Kim, Yuseob and Takahiro Maruki (2011). "Hitchhiking effect of a beneficial mutation spreading in a subdivided population". In: *Genetics* 189.1, pp. 213–226.

Kingman, John Frank Charles (1982). "The coalescent". In: *Stochastic processes and their applications* 13.3, pp. 235–248.

Lagunas-Rangel, Francisco Alejandro et al. (2017). "Acute Myeloid Leukemia—Genetic Alterations and Their Clinical Prognosis". In: *International journal of hematology-oncology and stem cell research* 11.4, p. 328.

Landrum, Melissa J et al. (2013). "ClinVar: public archive of relationships among sequence variation and human phenotype". In: *Nucleic acids research* 42.D1, pp. D980–D985.

Landrum, Melissa J et al. (2017). "ClinVar: improving access to variant interpretations and supporting evidence". In: *Nucleic acids research* 46.D1, pp. D1062–D1067.

Lappalainen, Tuuli et al. (2013). "Transcriptome and genome sequencing uncovers functional variation in humans". In: *Nature* 501.7468, p. 506.

Lohmueller, Kirk E (2014). "The impact of population demography and selection on the genetic architecture of complex traits". In: *PLoS genetics* 10.5, e1004379.

Manolio, Teri A et al. (2009). "Finding the missing heritability of complex diseases". In: *Nature* 461.7265, p. 747.

Mez, Jesse et al. (2017). "Alzheimer's disease genetic risk variants beyond APOE ε4 predict mortality". In: *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 8, pp. 188–195.

Nica, Alexandra C and Emmanouil T Dermitzakis (2013). "Expression quantitative trait loci: present and future". In: *Phil. Trans. R. Soc. B* 368.1620, p. 20120362.

Pavlidis, Pavlos et al. (2013). "SweeD: likelihood-based detection of selective sweeps in thousands of genomes". In: *Molecular biology and evolution* 30.9, pp. 2224–2234.

Pearson, Thomas A and Teri A Manolio (2008). "How to interpret a genome-wide association study". In: *Jama* 299.11, pp. 1335–1344.

Reich, David E et al. (2001). "Linkage disequilibrium in the human genome". In: *Nature* 411.6834, p. 199.

Richards, Sue et al. (2015). "Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology". In: *Genetics in medicine* 17.5, p. 405.

Rogers, David B and Larry J Shapiro (1986). "X-Linked diseases and disorders of the sex chromosomes". In: *Genetic Disorders and the Fetus*. Springer, pp. 341–368.

Saint Pierre, Aude and Emmanuelle Génin (2014). "How important are rare variants in common disease?" In: *Briefings in functional genomics* 13.5, pp. 353–361.

Schierding, William Stewart, Wayne S Cutfield, and Justin Martin O'Sullivan (2014). "The missing story behind Genome Wide Association Studies: single nucleotide polymorphisms in gene deserts have a story to tell". In: *Frontiers in genetics* 5, p. 39.

Schiffels, Stephan and Richard Durbin (2014). "Inferring human population size and separation history from multiple genome sequences". In: *Nature genetics* 46.8, p. 919.

Schork, Nicholas J et al. (2009). "Common vs. rare allele hypotheses for complex diseases". In: *Current opinion in genetics & development* 19.3, pp. 212–219.

Smith, John Maynard and John Haigh (1974). "The hitch-hiking effect of a favourable gene". In: *Genetics Research* 23.1, pp. 23–35.

Stranger, Barbara E et al. (2012). "Patterns of cis regulatory variation in diverse human populations". In: *PLoS genetics* 8.4, e1002639.

Sudmant, Peter H et al. (2015). "An integrated map of structural variation in 2,504 human genomes". In: *Nature* 526.7571, p. 75.

Telenti, Amalio et al. (2016). "Deep sequencing of 10,000 human genomes". In: *Proceedings of the National Academy of Sciences* 113.42, pp. 11901–11906. ISSN: 0027-8424. DOI: 10.1073/pnas.1613365113. eprint: http://www.pnas.org/content/113/42/11901.full.pdf. URL: http://www.pnas.org/content/113/42/11901.

Van Der Maaten, Laurens, Eric Postma, and Jaap Van den Herik (2009). "Dimensionality reduction: a comparative". In: *J Mach Learn Res* 10, pp. 66–71.

Vitti, Joseph J, Sharon R Grossman, and Pardis C Sabeti (2013). "Detecting natural selection in genomic data". In: *Annual review of genetics* 47, pp. 97–120.

Watterson, GA (1975). "On the number of segregating sites in genetical models without recombination". In: *Theoretical population biology* 7.2, pp. 256–276.

Wright, Sewall (1931). "Evolution in Mendelian populations". In: *Genetics* 16.2, p. 97.

Yoon, Byung-Jun (2009). "Hidden Markov models and their applications in biological sequence analysis". In: *Current genomics* 10.6, pp. 402–415.

Zhou, Jin and Yik-Ying Teo (2016). "Estimating time to the most recent common ancestor (TMRCA): comparison and application of eight methods". In: *European Journal of Human Genetics* 24.8, p. 1195.