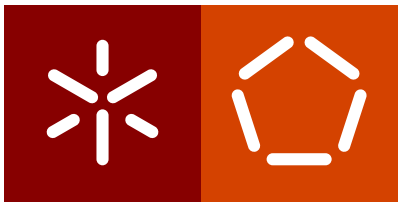
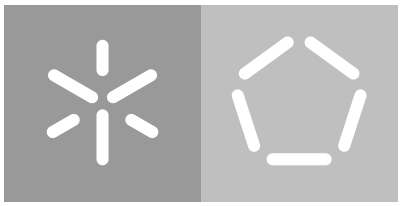


Universidade do Minho
Escola de Engenharia
Departamento de Informática

Raquel Dos Santos Silva

Estimating recombination frequency throughout the human genome using a phylogenetic-based method

October 2016



Universidade do Minho
Escola de Engenharia
Departamento de Informática

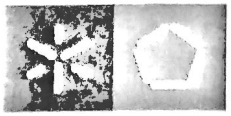
Raquel Dos Santos Silva

Estimating recombination frequency throughout the human genome using a phylogenetic-based method

Master dissertation
Master Degree in Bioinformatics

Dissertation supervised by
Pedro Alexandre Dias Soares

October 2016



Universidade do Minho

Declaração RepositóriUM: Dissertação de Mestrado

Nome: Raquel Dos Santos Silva

Nº Cartão Cidadão /BI: 14058714 Tel./Telem.: 912626391

Correio eletrónico: raqsilva8@hotmail.com

Curso Bioinformática Ano de conclusão da dissertação: 2016

Área de Especialização: Ciências e Tecnologias

Escola de Engenharia, Departamento/Centro: Departamento de Informática

TÍTULO DISSERTAÇÃO/TRABALHO DE PROJECTO:

Título em PT : Estimativa da taxa de recombinação no genoma humano utilizando um método filogenético

Título em EN : Estimating recombination frequency throughout the humen genome using a phylogenetic-based method

Orientadores : Pedro Alexandre Dias Soares

Declaro sob compromisso de honra que a dissertação/trabalho de projeto agora entregue corresponde à que foi aprovada pelo júri constituído pela Universidade do Minho.

Declaro que concedo à Universidade do Minho e aos seus agentes uma licença não-exclusiva para arquivar e tornar acessível, nomeadamente através do seu repositório institucional, nas condições abaixo indicadas, a minha dissertação/trabalho de projeto, em suporte digital.

Concordo que a minha dissertação/trabalho de projeto seja colocada no repositório da Universidade do Minho com o seguinte estatuto (assinale um):

1. ☐ Disponibilização imediata do trabalho para acesso universal;
2. ☒ Disponibilização do trabalho para acesso exclusivo na Universidade do Minho durante o período de

   ☐ 1 ano, ☐ 2 anos ou ☒ 3 anos, sendo que após o tempo assinalado autorizo o acesso universal.
3. ☐ Disponibilização do trabalho de acordo com o **Despacho RT-98/2010 c)** (embargo___ anos)

Braga/Guimarães, 27 / 10 / 2016

Assinatura: Raquel Dos Santos Silva

ACKNOWLEDGEMENTS

Em primeiro lugar, gostaria de agradecer aos meus pais, pelo apoio incondicional em todas as escolhas que fiz e que invariavelmente me levaram a este momento. Espero um dia poder retribuir de alguma forma tudo aquilo que me proporcionaram. Obrigada, também, Ricardo e Patrícia por todos os conselhos imprescindíveis e por facilitarem o acontecimento de muitas das oportunidades que me deram.

Não posso deixar de mencionar o Daniel e a Rafa, os melhores companheiros de casa que alguma vez podia desejar e que com o tempo se tornaram muito mais do que simples amigos. À Inês, à Daniela, ao Daniel, ao Vicente, à Vanessa; aos amigos da UTAD e do mestrado, um muito obrigada pelo companheirismo sem igual, pelas gargalhas e pelas palavras de apoio ao longo de todo este árduo processo. Ao Lucas e ao Nuno cuja ajuda e amizade foram preponderantes ao longo destes últimos dois anos. À Catarina, pelo apoio em todos os momentos, os bons, mas principalmente os menos bons, aqueles em que apenas as palavras de incentivo, força e carinho me ajudaram a continuar.

Finalmente, gostaria de terminar com um especial agradecimento ao prof Pedro, um orientador incansável e sem o qual nunca teria sido possível concluir este desafio com sucesso. Obrigada pelo apoio, disponibilidade, confiança, paciência e ajuda em todas as etapas da realização da minha tese.

ABSTRACT

Recombination rate is an essential parameter for most studies on human variation. Linkage disequilibrium (LD) measures the association between two variants in the same chromosome. When a new variant arises by mutation in a germinal line, that variant will be in complete linkage with the variants in the chromosomic background where it arises. Recombination through time (occurring during meiosis) will decrease the association decreasing the LD. Understanding how recombination occurs throughout the genome is the basis to interpret various association studies (search from causal variants for a given disease) and characterization of selective events. In this project the aim is to establish a novel methodology to estimate rate of recombination along a chromosome using a phylogenetic method. For this to be done, each chromosome will be divided into small overlapping windows of variation containing 20/30 variants. For each of these windows a phylogenetic network will be calculated using the reduced-median algorithm. Highly recombining regions will show a higher rate of cycles or reticulations in the network. A linkage map will be constructed for each chromosome using this novel methodology, compare the results with methods already available, locate region of low recombination of possible use for phylogenetic analysis and also explore some properties of the method for evaluation of selection.

Keywords: Linkage disequilibrium, single-nucleotide polymorphism, recombination, phylogenetics

RESUMO

A proporção de recombinação é um parâmetro essencial para os estudos baseados na variação encontrada nos humanos. O *linkage disequilibrium* (LD) mede a associação entre duas variantes no mesmo cromossoma. Quando uma nova variante aparece devido a uma mutação na linha germinativa, esta mesma variante irá estar em *linkage* completo com as outras variantes presentes no cromossoma onde esta apareceu. Com o passar do tempo, a recombinação genética (ocorre durante a meiose) irá diminuir a associação dos alelos, diminuindo o LD. A compreensão da recombinação ao longo do genoma humano é a base para a interpretação de vários estudos de associação (procura de variantes para uma doença especifica) e caracterização de eventos de seleção. O objetivo deste projeto é estabelecer uma nova metodologia para estimar a proporção de recombinação no decurso do cromossoma utilizando um método filogenético. Para isto ser realizado, cada cromossoma será dividido em janelas sobrepostas contendo 20/30 variantes. Para cada janela de sobreposição uma rede filogenética irá ser construída usando o *reduced-median algorithm*. Regiões com elevada recombinação irão mostrar um maior número de ciclos ou reticulações na rede. Um mapa de *linkage* será construído para cada cromossoma usando esta nova metodologia, comparando resultados com outros métodos já existentes, regiões de baixa recombinação irão ser localizadas para uma futura análise filogenética e explorar algumas propriedades desta metodologia de modo a avaliar a seleção.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF LISTINGS

LIST OF ABBREVIATIONS

1000genomes	1000 Genomes Project.
AF	Allele frequency.
cM	centiMorgan.
CNV	Copy-number variation.
DNA	Deoxyribonucleic acid.
DNA-seq	DNA sequencing.
ESP	Exome Sequencing Project.
EVS	Exome Variant Server.
ExAC	Exome Aggregation Consortium.
FTP	File transfer protocol.
GUI	Graphical user interface.
GWAS	Genome-wide association study.
HBB	Hemoglobin subunit beta gene.
HbS	Hemoglobin S.
LCT	Lactase gene.
LD	Linkage disequilibrium.
LP	Lactase persistence.
LPH	Lactase-phlorizin hydrolase.
MRCA	Most recent common ancestor.
mtDNA	Mitochondrial DNA.

NGS	Next-generation sequencing.
PAR	Pseudoautosomal region.
ped	Pedigree file format.
rdf	Roehl data format.
`SFS_CODE`	Selection on Finite Sites under COmplex Demographic Events.
SNP	Single nucleotide polymorphism.
tbi	Indexed tab-delimited file.
VCF	Variant Call Format.
VDE	VCFDataExporter.

1

INTRODUCTION

DNA (deoxyribonucleic acid) molecules are composed by four nucleotides (bases): Adenine (A), Thymine (T), Cytosine (C) and Guanine (G). In Eukaryotic organism, including animals, DNA is organized into chromosomes and most are diploid at the somatic level, meaning that with the exception of the germ line (the egg and sperm cells) all cells possess two copies of a given chromosome inherited respectively from the maternal and the paternal line of descent. The human genome contains 23 chromosome pairs, totaling 46 chromosomes from which 22 pairs are autosomes and a single pair, the X and Y sex chromosomes, which are different in females (two X chromosomes) and males (one X and one Y chromosome). Haploid cells in the germ line have only one chromosome from each pair and during the fecundation, a diploid egg is formed [4].

Although the genome of two unrelated individuals is almost 100% similar, a small set of differences (0.5%) is observable. These differences are responsible for many phenotypic differences between individuals but most are likely to be silent differences only observed at the DNA sequence level. This dissimilarity is explained due to sequence variations, which includes small insertions and deletions (indels) of one or more nucleotides, specific nucleotide substitution, which are the more common genetic variation, known as single-nucleotide polymorphism (SNP) [5], and copy-number variation (CNV), when abnormal copies of a chromosomal sections arise due to deletions and duplications (structural variants) [6]. New mutations that arise through time may increase or decrease the risk to have a certain disease, yet the majority of them are likely to have a minimal impact.

When there is a SNP present on a sequence, both alternative nucleotides are denominated alleles. The position with the variation is called a locus (and loci for various variable positions). Given this, the combination of several alleles within the loci of a region within a chromosome is called a haplotype. For instance, consider two SNPs from a region with six known SNPs (Figure 1); the former has alleles A and C; the latter has alleles T and G. The four possible haplotypes considering these two SNPs are AT, AG, CT and CG; however, only AT and CG are common [1].

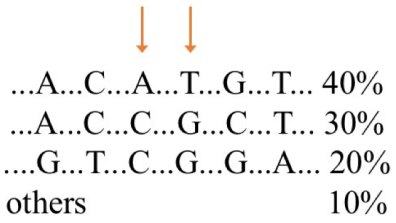


Figure 1.: Two common haplotypes from six SNPs. Retrieved from [1]

Considering that in diploid organisms (like humans) two copies of each chromosome exist in each individual it is not direct to extract haplotypic information. Seeing the case above one individual with the two most common alleles would appear as A/C in the third locus and G/T without a researcher knowing if the haplotypes were A-G and C-T or A-T and G-C. Although it is difficult to know individually haplotypes, on a grand scale of population units it is possible to statistically estimate what are the most probable haplotypes, in a process called phasing [7].

When present in the exome (in the protein-coding regions) SNPs may be classified as two types, nonsynonymous and synonymous. A change in the DNA alters the codon (set of three nucleotides that code for a given amino-acid). Nonsynonymous SNPs encode different amino acids due to the codon change, forming a different protein product. On the other hand, synonymous SNPs encode the same amino acid, and thus protein product, despite the allele difference (since different codons can code for the same amino-acid) [5].

1.1 Linkage disequilibrium

Recombination rate is an essential parameter for most studies on human variation. Linkage disequilibrium (LD) is the association between two or more alleles that are more likely to occur simultaneously at different loci on the same chromosome, meaning that some specific haplotypes are more common than expected from the frequency of the individual alleles. When a new variant arises by mutation in a germinal line, that variant will be in complete linkage with the variants in the chromosomic background (or the alleles in the haplotype) where it arises. This mutation is passed into descendants if the mutant sex cell takes part in fertilization (germinal line) [8], becoming part of the population gene pool. If no recombination occurs the variant would maintain the chromosomic background where

it initially emerged and the variant would be in complete LD with the variants in that background. But recombination through time (occurring during meiosis due to crossing-over events between homologous chromosomes) will decrease the association between alleles, decreasing the LD in this process. Nevertheless, recombination events occurring during meiosis in each generation in a population has a cumulative effect on patterns of LD, despite being relatively rare over small regions and negligible on a small temporal scale [9]. Furthermore, patterns of LD are powerful tools in leading to the identification of demographic events such as bottlenecks[1], admixture[2] and population growth [10] and to the identification of events of positive selection [11]. For instance, in case of admixture between two populations taking place, the descendant might have distinct segments within his genome inherited from different set of ancestors, possessing different genetic ancestries, which can be assessed through LD analysis [12]. Positive selection increases the frequency of a beneficial allele in a population. Reducing genetic variation on the region with the fixed advantageous variant on that population. This effect is known as the hitch-hiking effect, because it increases the frequency of an allele that is in LD with the allele under selection [13].

1.1.1 *LD measurements*

Several ways to measure LD were proposed. The two most commonly used are D' and r^2 (sometimes denoted $\triangle^2$), both related to the coefficient of LD (D) that is given by (1) [14]:

$$D = p_{AB} - p_A p_B \tag{1}$$

The frequency for allele A and for allele B are denoted by p_A and p_B, respectively. The haplotype frequency, considering allele A and B is denoted by p_{AB}.

The first measure was proposed by Lewontin [15] for the normalized measure of D (D'), where Dmax is the smaller value of $p_A(1 - p_B)$ and $p_B(1 - p_A)$, meaning the higher possible value of D given the allele frequencies:

$$D' = \frac{D}{D_{max}} \tag{2}$$

1 A temporary reduction in population size that causes the loss of genetic variation.
2 The mixture of two or more genetically distinct populations.

The second proposed measure is Pearson's squared correlation coefficient (r^2) [16] and it represents other way to quantify LD:

$$r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)} \tag{3}$$

Both measurements ((2) and (3)) are ranged between 1 (strong LD) and 0 (weak LD). If $|D'|$ equals to 1, two or three haplotypes are present; if it is significantly less than 1, all four haplotypes are present indicating recombination events. In case of r^2 being equal to 1, only two haplotypes are present [10]. While the interpretation for maximum LD (1) and no LD (0) is direct, any value in between is difficult to interpret in terms of how strong LD is, as the values of $|D'|$ and r^2 are highly dependent on allele frequencies and measurements are not directly compared.

1.1.2 *Patterns of LD*

Although LD patterns are difficult to predict, there are regions in the human genome with weak evidence of historical recombination (strong LD) composing a model known as haplotype blocks (haploblocks) [17]. In addition, this model suggests the hypothesis to the existence of sites where much of the recombination occurs (hotspots). Furthermore, some studies were made to examine this theory and concluding on a ubiquitous existence of hotspots within the human genome [18]. However, the human sex chromosomes in males do not recombine in the same manner as autosomes or female sex chromosomes, but they have two homologous regions in which the recombination occurs in a similar fashion as the remaining 22 chromosome pairs. These short regions are called pseudoautosomal regions (PAR1 and PAR2); the first is the longest and has a physical length of 2.6 Mbp (2.6 mega base pairs; $2.6\text{x}10^6$ bp) located at the short arms' tips of X and Y chromosomes; the latter, the PAR2, has a physical length of 0.32 Mbp and it is located at the long arms' tips [19]. PAR1 exhibits in males a much higher recombination rate than PAR2 or the genome average. PAR1 plays a key role in spermatogenesis and diseases [20, 21].

The traditional method to map recombination rates, and thus hotspots, is to compare the physical map to the genetic map. Physical distance is the number of base pairs that separates two loci which is obtained by genomic sequencing, while genetic distance is calculated through recombination frequencies and indirectly by linkage. On that note, the

measure for genetic distances is the centiMorgan (cM), defined by having a 1% chance of crossover between two genes on the chromosome. The average frequency of recombination in humans is 1cM per 1Mbp. Some regions near telomeres have generally much higher recombination frequency for both sexes, being higher in males [22], opposed to the low recombination (coldspots) found near the centromere for males, suggesting that recombination rate in general is higher near the tips of the chromosomes.

1.1.3 *LD observed in populations*

Shorter haploblocks, and thus generally less LD and more divergent patterns of LD[3] were observed on African populations that also display greater levels of diversity than non-African populations, probably suggesting an origin of modern humans in Africa (Figure 2) and a longer time for the disruption of LD patterns. Due to the Out of Africa movement migration of anatomically modern humans between 70 and 60 thousand years ago [23], that caused a bottleneck on the overall variation, the number and diversity of haplotypes is reduced when compared with Africa. This bottleneck consequently, increased severely LD on non-African populations [10, 24, 25].

As mentioned before, Africa is thought to be the ancestral homeland of modern humans. Genetic studies on African human populations provide important information about how genetic variations alter the phenotype and for fine-scale mapping of complex diseases [26, 27]. Moreover, complex diseases like hypertension, diabetes, obesity and prostate cancer are progressively present in African populations, presumably as a result to and urbanized Western lifestyle. The possibility of existing population-specific alleles that predispose to disease, is very alluring for mapping these particular variants [28].

3 Alleles that are in positive association in one population might be in negative association in another

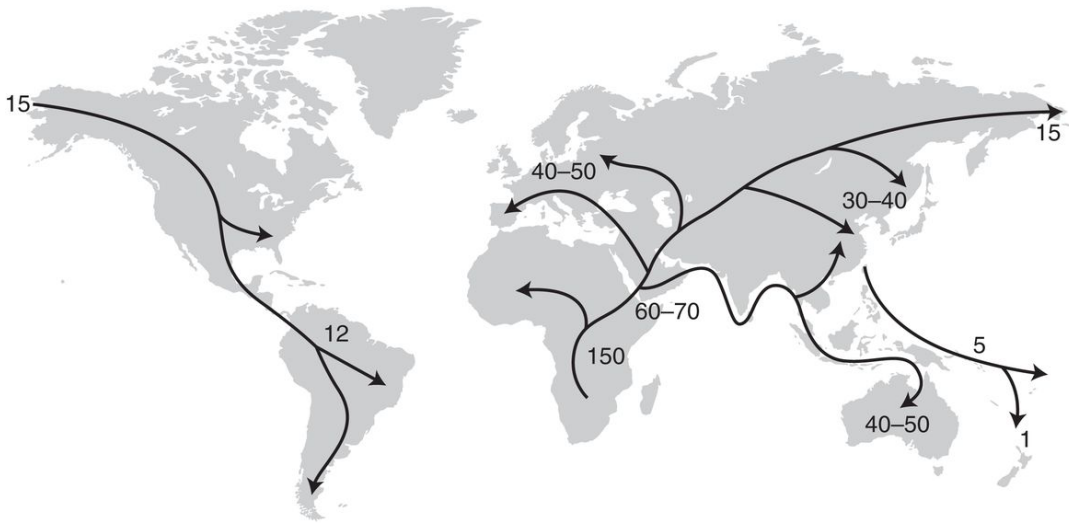


Figure 2.: Humans Out-of-Africa movement in thousand years. [2]

1.1.4 *Genome-wide association study (GWAS)*

Understanding how recombination occurs throughout the genome is the basis to interpret various association studies (search from causal genetic variants for a given disease) and characterization of selective events. Likewise, genome-wide association study (GWAS) is a research approach that seeks to [29]:

- Detect variants associated with complex traits in populations, offering suspect regions;
- Study patterns of LD between SNPs to map genomic loci that have an effect on illnesses or other complex traits;
- Detect genetically causality for common complex diseases such as heart disease, diabetes, auto-immune diseases and psychiatric disorder.

This type of studies will allow the identification of genetic markers associated with certain diseases, making a statistical estimation for the increased risk of developing the disorder. In some occasions, some variants are imputed, which means that genotypes are estimated for SNPs from the known LD patterns and haplotypes, using reference data. On that note, samples with shared haplotypes are used to estimate their frequencies among the genotyped SNPs. Association mapping studies look for LD between alleles that cause heritable diseases and other nearby alleles, even though only a few may affect the phenotype [30, 31].

1.1.5 *Positive selection*

Positive selection increases the frequency of a beneficial allele in a population. Selective sweep takes place when genetic variation on the region is reduced caused by positive selection. This is because the fixed advantageous variant on that population increased dramatically in frequency, and as its frequency increases at a fast rate, recombination was not able to break down the haplotypic background where the beneficial variant arose increasing the frequency of the complete haplotype, decreasing the overall haplotypic diversity [32]. This effect is known as the hitch-hiking effect, because it increases the frequency of an allele that is in LD with the allele under selection [33]. This process is the basis for selection tests, more exactly haplotype-based tests, that will measure the haplotypic diversity in one allele, in the allele under selection and the other allele. Such tests include the integrated haplotype score (iHS) [34] and the cross population extended haplotype homozygosity (XP-EHH) [35].

Addressing two regions to be explored later on that are an example of positive selection that took place recently due to the presence of a new favorable allele. The lactase gene (LCT) and the hemoglobin subunit beta gene (HBB) are responsible for the ability to digest milk (lactase persistence) and resistance to some forms of malaria, respectively.

Lactase persistence (LP) is mostly common on European populations and less common on Africans. It allows the synthesis of the lactase-phlorizin hydrolase (LPH) enzyme to digest the lactose present on milk and other dairy products after weaning, possibly owing to positive selection [36, 37]. Although the SNP that will be studied is from the MCM6 gene, it has a direct influence on the LCT gene that encodes LPH. For the European population the rs4988235 is strongly associated with LP, specifically the LCT-13910*T allele. The polymorphism is characterized by a C/T transition, where the individuals that have the T allele will carry the lactase persistence genetic trait [36, 37, 38].

The rs334 SNP located in the HBB gene, has an hemoglobin S (HbS) responsible for sickle-cell anaemia only on homozygote individuals (T;T). Despite having a strong deleterious effect in heterozygous individuals (A;T), its presence provides a strong protective effect against malaria infection. This mutation is characterized by an A/T transition and is highly present on African populations, due to being a region with high prevalence of Malaria [39, 40, 41, 42].

1.2 VARIANT DATABASES

To discover sequence variations, next-generation sequencing (NGS) technologies were a breakthrough on DNA sequencing (DNA-seq). While the first human genome draft was accomplished during a period of nearly a decade and a half (using traditional Sanger sequencing) [43] nowadays next-generation sequencing is allowing genomic and exomic data to be generated at an unprecedented level. As one example recently in a single Nature journal issues, hundreds of genomic sequences were published and analyzed [44, 45, 46].

Several databases focus on storing variation obtained from these genomes. In order to obtain a list of variants in a relatively user-friendly format the data needs to be processed using algorithms that will transform the data from raw data, that consists of millions of individual reads into, for example, a Variant Call Format (VCF) file, that will be widely used throughout this work.

An example of a pipeline from raw data to VCF files will be provided, focusing specifically on the Illumina[4] DNA-seq pipeline chosen by the phase 3 of the 1000 Genomes Project (1000genomes). This type of sequencing produces reads (raw data) that are more commonly stored in FASTQ files along with the quality scores. Quality diagnostics are performed on the raw data to determine if preprocessing is necessary. The next step is to align/map the reads to a reference genome resulting in a SAM (Sequence Alignment/Map) file format or BAM (binary version). The Burrows-Wheeler Aligner (BWA[5]) is used for short reads mapping to the large reference human genome. Moreover, a post alignment preprocessing to remove bias is done before the final step of variant calling with an estimate on allele frequency (AF) and storage in a VCF file [47, 48, 49, 50]. This will provide genotypes of an individual. However, the 1000genomes currently provides VCF files with haplotypic data (meaning that the data is in the so-called phase). This is obtained by a statistical method that uses complex algorithms to estimate the most probable combination of alleles given the population patterns. Such phasing is performed in the case of the 1000genomes with the software SHAPEIT2 [51, 52], that performs this task on the full chromosome level.

Furthermore, we will describe some databases of interest that, although all results focus on the 1000genomes, were used experimentally in the development of the work.

4 http://www.illumina.com/
5 http://bio-bwa.sourceforge.net/

1.2.1 1000 *Genomes Project*

The 1000 Genomes Project (1000genomes) purpose is to catalogue variations present in the human genome from different genomics regions across 26 populations categorized in 5 super-populations, shown on Table 1 (with ancestry from Europe, East Asia, South Asia, West Africa and Americas).The objective is to characterize, using high-throughput sequencing technologies, over 95% of variants with an AF of at least 1% and alleles present in coding regions with a frequency as low as 0.1% [53]. This dataset contains 78 million SNPs and was aligned using the GRCh37 reference genome assembly.

Throughout the 1000genomes, new variants were discovered and characterized, being added to the public database of short genetic variations (dbSNP[6]) or to the database of genomic structural variation (dbVar[7]).

Due to the advance of sequencing technologies it is now possible to sequence genomes with a lesser cost. This NGS data provides a resource on human genetic variation. Each dataset may have different assembly algorithms and different coverage [54, 55]. Genotyping these variants will reveal the alleles at particular sites or regions. All genotyped variants present in the 1000genomes data are phased, denoting the putative chromosome of the pair where that specific allele belongs. Most of the project contains low coverage data from populations suggesting that the individuals' DNA was sequenced approximately 4 times (4x). This dataset contains samples from 2504 individuals, 1233 males and 1271 females.

Furthermore, the project allows to document the genotype profiles of the 2504 individuals from the last phase, phase 3, of the 1000genomes for further study, being accessible through a public FTP (File Transfer Protocol) server in compressed files (GNU Zip extension, '.gz') [56]. Therefore, the fact that it is in a FTP server[8] means that it is possible to download, upload, delete, rename or change permissions of files however one cannot access the files' content remotely.

6 https://www.ncbi.nlm.nih.gov/projects/SNP/
7 https://www.ncbi.nlm.nih.gov/dbvar
8 ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/

Table 1.: Populations present on the 1000 Genomes Project, their Code and number of samples. Information retrieved from the 1000genomes web page

Population	Code	Samples
Chinese Dai in Xishuangbanna, China	CDX	93
Han Chinese in Bejing, China	CHB	103
Japanese in Tokyo, Japan	JPT	104
Kinh in Ho Chi Minh City, Vietnam	KHV	99
Southern Han Chinese, China	CHS	105
Total East Asian Ancestry	EAS	504
Bengali in Bangladesh	BEB	86
Gujarati Indian in Houston, TX	GIH	103
Indian Telugu in the UK	ITU	102
Punjabi in Lahore, Pakistan	PJL	96
Sri Lankan Tamil in the UK	STU	102
Total South Asian Ancestry	SAS	489
African Ancestry in Southwest US	ASW	61
African Caribbean in Barbados	ACB	96
Esan in Nigeria	ESN	99
Gambian in Western Division, The Gambia	GWD	113
Luhya in Webuye, Kenya	LWK	99
Mende in Sierra Leone	MSL	85
Yoruba in Ibadan, Nigeria	YRI	108
Total African Ancestry	AFR	661
British in England and Scotland	GBR	91
Finnish in Finland	FIN	99
Iberian populations in Spain	IBS	107
Toscani in Italia	TSI	107
Utah residents with Northern and Western European ancestry	CEU	99
Total European Ancestry	EUR	503
Colombian in Medellin, Colombia	CLM	94
Mexican Ancestry in Los Angeles, California	MXL	64
Peruvian in Lima, Peru	PEL	85
Puerto Rican in Puerto Rico	PUR	104
Total Americas Ancestry	AMR	347
Total		2504

1.2.2 *Exome Aggregation Consortium*

The Exome Aggregation Consortium (ExAC) is an affiliation of investigators with the objective to compile exome (protein-coding regions) sequencing data from different projects that range from disease-specific individuals to population-specific datasets. The dataset is composed by 60706 samples, some representing individuals with severe diseases, of unrelated individuals from 7 different populations (Table 2). This merging of variant data from different projects will provide a base for the discovery of disease-causing variants [57].

Table 2.: Populations on the ExAC data, their code and number of samples. Information retrieved from the ExAC web page

Population	Code	Samples
African/African American	AFR	5203
Latino	AMR	5789
East Asian	EAS	4327
Finnish	FIN	3307
Non-Finnish European	NFE	33370
South Asian	SAS	8256
Other	OTH	454
Total		60706

1.2.3 *Exome Sequencing Project*

The National Heart, Lung, and Blood Institute created the Exome Sequencing Project (ESP) with the intent of sequencing the exome of richly-phenotypes populations, achieving 6503 samples from two different ancestries (Table 3). Apart from a compressed dataset publicly accessible, this project has, a web interface known as the Exome Variant Server (EVS) that allows the visualization and extraction of filtered data to text files and Variant Call Format files. The ESP contains a mean coverage data of 81x [58].

Table 3.: Populations on the ESP data, their code and number of samples. Information retrieved from the EVS web page

Population	Code	Samples
African-Americans	AA	2203
European-Americans	EA	4300
Total		6503

1.2.4 *Simons Genome Diversity Project Dataset*

This database aims to include a list of genomes that were sequenced with a coverage of at least 30x using Illumina technology [45]. The database aims to display a wide range of anthropological and cultural diversity in terms of individuals.

This newly developed database offers an ideal dataset for the work described here, however, it only provides the dataset in an unfriendly format for download (10 terabytes of data) and requires a certificate. The download is through a software that allows fast and secure (via encryption) data movement. The certificate needs to be approved and is required for security reasons (privacy of the files), when accessing the FTP channel to download the data [59].

1.3 RELEVANT FILE FORMATS

There are several file formats used in genomics and their features and levels of details incorporated are direct consequences of their use in different software.

1.3.1 *Variant Call Format*

To store variants, plus annotations, a file format was proposed by the 1000genomes, the Variant Call Format (VCF) file. Its adaptability and flexibility led to an increasing endorsement of the VCF file. On a similar note, VCF files were established as standard files to store DNA polymorphism data, such as SNPs, insertions, deletions and structural variants, including rich annotations.

Header

```
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	40	PASS	.	GT:DP	1/1:13	2/2:29
1	2	.	C	T,CT	.	PASS	H2;AA=T	GT	0|1	2/2
1	5	rs12	A	G	67	PASS	.	GT:DP	1|0:16	2/2:20
X	100	.	T	<DEL>	.	PASS	SVTYPE=DEL;END=299	GT:GQ:DP	1:12:.	0/0:20:36

Figure 3.: VCF file format example. Retrieved directly from [3]

At the top, the VCF file contains several meta-information lines, each starting with '##'. These lines describe the data stored, which includes the type of variable associated with that data (integer, string, float, etc) and customized relevant information about the dataset. Below meta-information lines the actual genomic data will be displayed, starting with '#', featuring 8 mandatory columns with specific fields, TAB-delimited. Particularly, this information includes the chromosome column (CHROM), the start position of the variant according to the reference sequence (POS), unique identifiers of the variant (ID), the reference allele (REF), the list of alternate alleles (ALT), quality score (QUAL), site filtering information (FILTER) and a list of additional annotation (INFO). Note that if samples are present in the file, sample columns will be added (sample ID) and a column regarding the information contained in the sample column (FORMAT). Genotypes separated by '|' indicate the alleles are phased, if they are separated by '/' they are unphased alleles [3].

1.3.2 *Input File Formats*

The pedigree (ped) and map file formats are complementing type of files that are the most commonly used type of format in GWASs and they can be implemented using the software PLINK [60]. The ped file contains pedigree information and genotype calls and the map file contains variant information to complement ped file regarding the genomic position of the variants.

Some software tools use small variations in the file format in relation to PLINK. Such an example is the Haploview software that uses the ped file in combination with a file with extension info (similar to map) that provides information on the location of the SNPs along the chromosome. Haploview calculates LD patterns within a region [61] which it will be

essential as a comparative tool in this work.

The Roehl data format (rdf) files are used by the Network software using a reduced median algorithm [62]. For the purpose of performing phylogenetics on the 1000genomes data the implementation of a software that builds networks makes more sense than into one that reconstruct straightforward trees, considering that the data can be shuffled by the so-called recombination between chromosomes.

The nexus (nex) file format is one of the most widely used formats in genetics. At its most simple form it includes the DNA data and the name of the individual samples but it can also include discriminated subsets.

The fasta file is considered the most basic DNA and protein sequence file format. It basically contains only information of the sample and the sequence (in nucleotides or amino-acids) for a given region in text format.

2

STATE OF THE ART

2.1 RELEVANT SOFTWARE

2.1.1 *VCFDataExporter*

The VCFDataExporter[1] (VDE) is a tool mostly developed over last year using Python functions simultaneously with PyVCF that allowed the extraction of genomic data from the 1000genomes VCF files and transformation of that data into various formats useful in existing genetic analysis software, mentioned before. The tool is now on his second release after further developments on the context of this project (v0.1.1) and provides the extraction of genomic data from the 1000genomes dataset, ExAC dataset, ESP dataset and a user uploaded VCF file. From the user specified information imputed it is possible to extract the data and transform it into map, rdf, fasta, nex, ped, info, VCF and excel files, and generate some basic statistic regarding the data. After the requested files are created it will provide links for download or pop-up.

The conversion tool developed allows a user-friendly manipulation for researchers not used to deal with the large datasets established by NGS. On that note, a GUI on the form of a web app was developed using a Python web framework, known as Django. For a faster extraction of the region and his information a tab-delimited files indexer[2] (Tabix) is being used to subset the VCF gzipped file with the completing indexed tab-delimited file (tbi). The subsetted VCF file will allow a faster retrieval of the genomic data and subsequent storage on different files.

2.1.2 *Network Software*

This software builds phylogenetic networks and trees, using several algorithms over a graphical user interface (GUI). It constructs a network following a multiple alignment of

1 https://github.com/raqsilva/VcfDataExporter
2 http://sourceforge.net/projects/samtools/files/tabix/

sequences, in which each sequence is represented by a node and the edges, that connect one node to the other, represents the difference between each other (variants). This difference is seen if they differ exactly in one mutation. The network is constructed using the reduced median algorithm [62]. This algorithm consists on median networks which are composed by nodes connected by edges if there is a mutation between them. This type of method only accepts binary sequences. The median sequence results from the median binary values of three sequences and is used to connect the nodes culminating on an unrooted and undirected phylogenetic network (Figure 4). The median node could also be an existing sequence if it corresponds to the median of the three other sequences. Median networks are mainly employed in the visualization of possible evolutionary pathways. On that note, the network will probably have all maximum parsimony trees; in other words, it will possibly have trees with the minimum number of nucleotide mutations [63].

The Network software will generate an output file (out file) for each rdf file used to run the program and build the phylogenetic network. Out files contain information about the data in rdf files and of the phylogenetic network. In this work, a modified version of the software provided by the developer was used. This version allowed batch work on several files to be performed while the downloadable version only allowed a single file to be run.

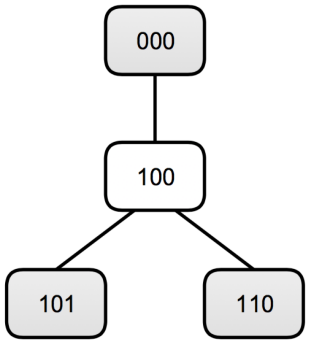


Figure 4.: Three haplotypes (grey) and the median node (white)

2.1.3 *Haploview*

Allows the visualization, from a GUI, of haplotype patterns within a region, calculates LD statistics, population haplotype frequencies and haplotype association tests. The several pairwise measures of LD calculated will be used to create a graphical representation of the region. This representation uses the haploblock model, mentioned on the introduction, to partition the region into segments of strong LD [61].

2.1.4 *PLINK*

Is a command line program that performs genome-wide association analyses from samples' genotypes. It is one of the main used programs in human genomics as it is used to filter data and merge datasets to be used in other software. Some of the main characteristics are [64, 60]:

- Data management: recoding, reordering, merging and extracting subsets of data;
- Summary statistics: missing rates and allele frequencies;
- Basic association analysis;
- Population stratification;
- Genotypic association models;
- Extracting SNPs of interest;
- Stratification analysis;
- Association analysis, accounting for clusters.

2.2 RELEVANT LIBRARIES

2.2.1 *PyVCF*

In order to parse and extract specific data from compressed VCF files, the programming language Python and an API called PyVCF[3] (v0.6.7) are the main mechanisms [65]. There is no need to decompress the compressed files before parsing, because the package comes with that attribute. Due to this feature the compressed files are preferred over the decompressed VCF files, since they are faster to transmit over the internet due to their reduced size and require less space on the disk for storage, contrary to the decompressed VCF files that have huge sizes (meaning it will be very difficult to handle them). Each VCF file has several records depending on how many positions are represented. It will attempt to parse the content of each record based on the data types specified in the meta-information lines, specifically the ##INFO and ##FORMAT lines. On that note, it is possible to extract genomic data from compressed VCF files and transform that data into various formats useful in existing genetic analysis software (map, rdf, fasta, nex, ped, info).

3 https://github.com/jamescasbon/PyVCF

2.2.2 *NetworkX*

In pursuance of building graphs and analyze their structure this Python library was used. In addition, NetworkX[4] has several characteristics including the possibility to generate directed or undirected graphs, apply algorithms and do measurements, draw networks and add information to the network, edges or nodes. Furthermore, it allows to store networks in different formats and improve the access speed to the network data. This characteristic is very valuable to process data from large networks, like our phylogenetic networks, in a considerable shorter time.

4 https://networkx.github.io/index.html

3

OBJECTIVES

Herein, this project objective is to extract information from the VCF files corresponding to the 1000genomes and analyze LD along a chromosome. This will be done through windows of polymorphic diversity along a chromosome. This work will partially build upon VDE, that was further developed also during this project. Due to their abundance, their importance in genomic variation and their simplicity of treatment, most of the data to be extracted and transformed in this work will be primarily SNPs. The objective is to transform each window into a network whose cycles will be an indication of the possible occurrences of recombination events. Such approach provides two advantages over traditional measures. The first is that cycles might be a direct estimation of recombination events and consequently of recombination rate. The second is that the measure is not dependent of frequencies of the alleles or haplotypes and the sample size.

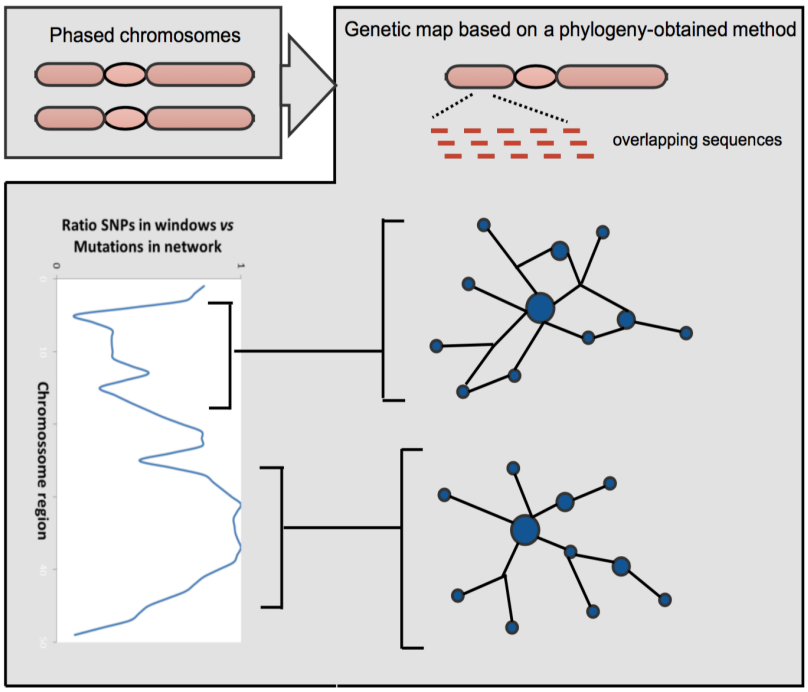


Figure 5.: Steps to build a recombination frequency map using a phylogenetic based method

From the output of the network analysis, the main interest is in obtaining three measures:

a) Number of mutations in the network. If there is no recombination (or multiple occurrences of a given mutations which should be substantially rarer in autosomic DNA) in the segment the number of SNPs in the segment will match the number of mutations in the network. With recombination multiple mutations will appear in the network so ratio between number of mutations in the network and number of SNPs will provide a good relative measure for recombination frequency.

b) Each recombination event can generate a cycle (or reticulation) in the network. Counting the number of possible cycles in each generated network is a direct estimate for the recombination rate in the region.

c) Frequency increments (expansions) generate starlike figures in a phylogenetic context. In a genomic context an expansion within a given region of the genome might represent a signal of positive selection. Phylogenetic starlikeness will be explored to see if it is a useful measure for selection evaluation.

Using measures in a) and b) a linkage map will be built. These linkage maps will be compared with maps based on traditional measures of linkage disequilibrium. VDE will be

directly used to create the complete chromosome files to be used in PLINK and Haploview. Additionally, in order to validate the approach, simulations will be used. A short evaluation of the potential of the use of recombination detection will also be included.
Approaches for studying selection and the detection of selective sweeps will be employed, namely recombination maps comparing the two alleles (the beneficial allele and the other), based in well-documented cases of positive selection, but also across general chromosome maps based on the starlikeness measure.
As a final analysis, if relatively long regions with no recombination are detected, networks/trees of those longer stretches will be built, first using Network to make a final test for no recombination along the whole region.

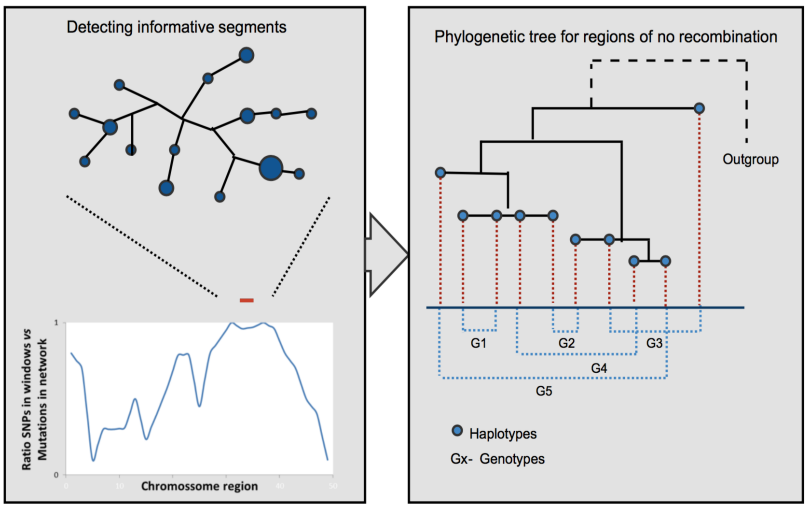


Figure 6.: Steps to build a phylogenetic tree of long regions with no recombination

4

METHODS

4.0.1 *Data filtering*

Through the analysis of existing VCF and rdf files, the comprehension and study of the structure of these files format was required. Parsing is done from a previously obtained VCF file, which contains the necessary data that will be used to create the rdf file with information to elaborate networks. First, it was considered multiallelic sites, however one decided that it was more accurate and straightforward to analyze only biallelic loci corresponding to binary data. To ensure that rdf files will only have polymorphic sites, SNPs are filtered by AF. In instances where analyses were based on groups with a continent-specific ancestry, the AF filtering for each population was performed based on the specific AF for the population and considering only samples belonging to that ancestry. Allele frequencies that are equal to 0 or 1 the SNP will not be included as it means that that locus is not polymorphic in that given population. Notwithstanding, this will allow the presence of at least the same number of mutations in the phylogenetic network as the number of SNPs in the rdf file.

Analyses were performed based on overlapping windows of 20 SNPs overlapping at 10 SNPs or 30 SNPs overlapping at 15 SNPs. Overlapping windows are acquired by considering the first 20/30 SNPs, previously filtered, and then passing the last 10/15 SNPs to the next window adding 10/15 new SNPs, appending these. As a result, for each window 20/30 SNPs are present for each chromosome. To divide the information for each chromosome each sample will be divided into two entries, corresponding to the two phased haplotypes, originating, for instance, HG00102a and HG00102b regarding the sample HG00102. However, concerning the case of the X chromosome in males there will be only one entry for each sample on sites that do not belong to PAR and two entries (a and b) on PAR. These regions were previously retrieved from the VCF file for future filtering by position, dividing chromosome X in three regions. The first region that corresponds to PAR1, the middle region and the last corresponding to PAR2.

Windows were successfully obtained extracting data along the VCF files with 20 or 30 SNPs overlapping by 10 or 15 SNPs, respectively, into various files. Output files are in rdf format for further use as input in the Network software and additionally a text informative file with the mutation number, SNP IDs and its genomic position was formed.

Each rdf file generated along a chromosome will be used to calculate a phylogenetic network, corresponding to the diversity in that window, and then produce an output file (out file) using the reduced median algorithm available in the Network software. The most effective way to use the algorithm for thousands of fragments was to contact the developers of the Network software, which were able to reprogram the software adding an additional feature capable of performing the algorithm on multiple files instead of requiring each file to be selected individually. This characteristic allowed to generate all the thousands of out files from all rdf files of windows along a from the chromosome.

Specific data from two interesting SNPs that have been firmly established as two examples of positive selection in the human genome will be analyzed. The SNPs chosen are responsible for having a direct influence on the expression of the LCT gene (rs4988235) and HBB gene (rs334) that respectively allows the digestion of lactose and offers protection against Malaria. Moreover, a region surrounding the genes/SNPs of about 4 million base pairs is chosen from the VCF file of chromosome 11 to include the HBB gene and from the VCF file of chromosome 2 to include the SNP (located in MCM6 gene) that has influence on the LCT gene. Following this, these regions are outputted in a new VCF file for each gene. In the 1000genomes data the transition of rs4988235 and rs334 is characterized by G/A (instead of C/T like mentioned before) and T/A, respectively. For each gene, samples from EUR or AFR were the only ones considered, and the haplotypic data was separated into different folders based on the geography. Following this, the data was further separated based on the presence of the SNP or not in the haplotype, where one folder will contain an rdf file with the haplotypic information of the chromosomes that have the reference allele, and the other folder will contain the haplotypic information in the chromosome that have the alternate positively selected allele.

4.0.2 *Data processing*

Two files (out file and text informative file) are essential to extract required information on phylogenetic networks. To ensure these steps Python scripts were written, including the development of efficient methods to store, retrieve and process large amounts of data. The out file contains the necessary data to build a graph with NetworkX and count existing

mutations. This data consists in nodes and edges with the respective mutation number. Moreover, it was possible to count all cycles (reticulations) obtained from the connection of 4 nodes (1 being equal; first and last node) by 4 edges that have the same 2 mutations; and obtain the SNP IDs that are contained in those edges (Function B.1). The text file provides the SNP ID and the genomic position of each SNP, which allows calculating the region size by subtracting the genomic position of the last SNP with the first. This gathered information is used to build a plot for the visualization of the measurements of recombination frequency throughout the chromosome with matplotlib[1]. Recombination rate in the region is measured by the ratio between cycles and region size; while a second linkage measure was established by the ratio between mutations and region size. All the data was stored in a tab-delimited text file.

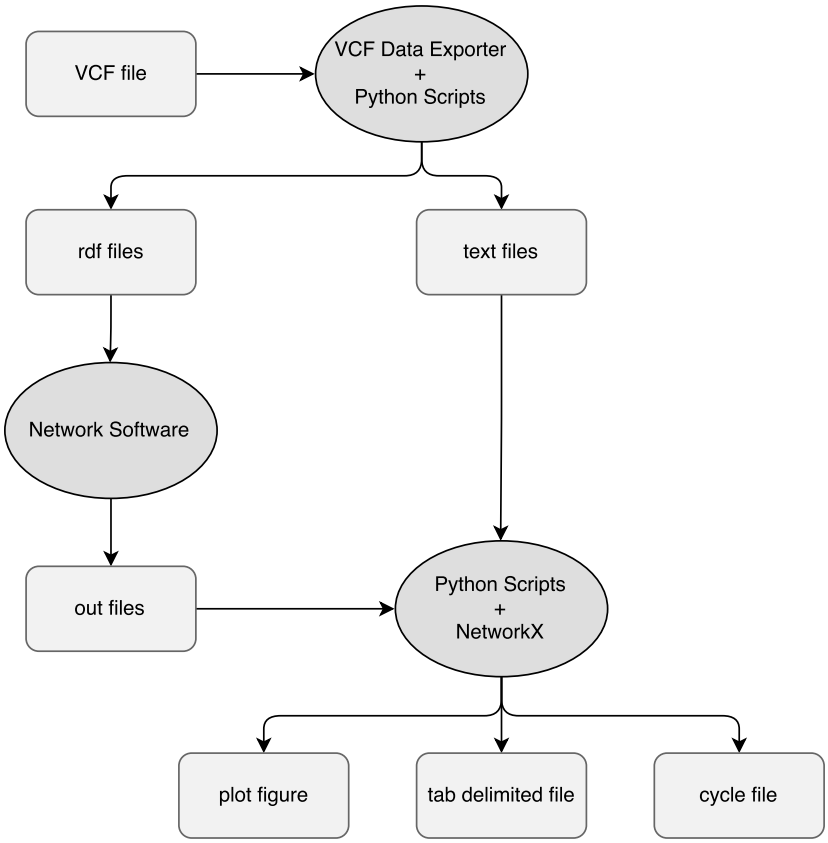


Figure 7.: Pipeline of the process to generate required data. Light gray are the output files and dark gray are software or scripts implemented during the workflow.

1 http://matplotlib.org/

Cycles in populations

The output from the reticulations accounted for each population are important to explore the use of recombination as a clock, it will be explored assuming an African origin for Modern Humans and the Out-of-Africa movement. In this regard, all cycles from each of the five main population groups (Africa, Europe, South Asia, East Asia and America), along the different windows of chromosome 22, were calculated and they were compared with the cycles from the other 4 populations. For this to be possible, scripts were made elaborating an algorithm to course through the entire network from each window finding all reticulations (cycles) in the least time consuming way and considering the computing resources available. The algorithm (Function B.1) will course only through the nodes that are connected, owing to being those that aren't leaf nodes and probably resolving into cycles. It will select a node (n1) coursing through all neighbors and saving the SNP ID (e1) linking it with the neighbor (n2) and edge, then checking all neighbors from n2 (namely node n3), not equal to n1, and saving the respective SNP ID (e2) of the edge. Then it will search for the neighbors of n3 and compare e1 with the SNP ID of the edge that connects n3 to the neighbor (n4). If the SNP IDs are equal it will continue the course if not, the algorithm will try other neighbors. This n4 node will have to be connected to n1 and have the same SNP ID as e2 to complete a cycle. All information related regarding edges and neighbors is stored. The SNP IDs that define each cycle are in a list ([e1, e3]) which will be stored in a text file. The forthcoming analysis between comparing matches in the cycles of each specific population against another simply corresponds to a search for a cycle in one population or its inverse ([e3, e1]) in the file of the other population.

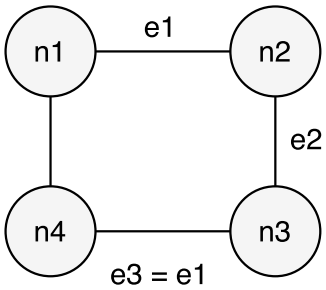


Figure 8.: Cycle example.

Starlikeness

Using our phylogenetic network, we built a phylogenetic tree based on the most probable evolutionary path with the minimum number of evolutionary events established from the maximum parsimony method, eliminating edges that are part of the reticulations (cycles) [63, 66]. The graph will have to be weighted for the calculation of a Dijkstra path from the

central node (root) to all other nodes present in the network. The central node is the node with higher degree, which is the node with more edges connected to it. Dijkstra algorithm will find the minimum cost path (optimal path) between all nodes in the graph [67]. In this case we will find the optimal path from the central node to all other nodes. The optimal path is found because all edges have a weight attribute that was implemented specially for the algorithm. The weight of each edge depends on the frequency of haplotypes on each node that compose the edge. Moreover, nodes without samples (like median vectors) are given an edge weight of 2 and all other nodes have an edge weight of 1 (Function B.4). Furthermore, Dijkstra paths are computed and phylogenetic trees are built based on those paths.

The tree root is considered the most recent common ancestor (MRCA) and the other existing nodes are related to this central node by mutations that occurred on specific sites connected by edges. Additionally, this tree demonstrates the mutational history. The coalescence time is the time that has elapsed since the presence of the MRCA of several copies of a locus [68, 69]. Given each node with n haplotypes in a tree with i edges and that edge with m mutations, we estimated the variance (ρ^2) of the coalescence time (ρ) that can be calculated by [70]:

$$\rho = \frac{\Sigma\, n_i \times m_i}{n_{total}} \tag{4}$$

$$\sigma^2 = \frac{\Sigma\, n_i^2 \times m_i}{n_{total}^2} \tag{5}$$

Where the theoretical value, which when equal to 1 corresponds to a perfect star phylogeny, is given by [70]:

$$\sigma_t^2 = \frac{\rho}{n_{total}} \tag{6}$$

And the final calculation for the starlikeness of the network is:

$$starlikeness = \frac{\sigma_t^2}{\sigma^2} \tag{7}$$

After calculating the variance, the theoretical variance and the starlikeness for each overlapping windows, a plot with those values was built to verify the starlikeness throughout

the chromosome.

4.0.3 *Data simulation*

Simulations were performed using `SFS_CODE` (Selection on Finite Sites under COmplex Demographic Events), which is a software that generates population genetic simulations. It has the capacity to generate populations of individuals and follow them generation by generation evolving by natural selection. Evolutionary algorithms are applied so that a given number of individuals hypothetically evolve under different levels of variation, selection and inheritance, including important recombination mechanisms like crossing-over and gene conversion (unidirectional transfer of genetic material) that can be modulated by the user. Demographic effects and migration can be included. Furthermore, DNA sequences are generated for each individual in the simulations and aligned. Due to the simulation of finite-sites mutations this alignment may contain several sites that could have being a target of several mutations (homoplasy) which could cause false possible recombination events [71, 72, 73]. Nevertheless this is also a strong possibility in real data which can allow the simulations to mimic false positives that we could find in real data.

All simulations using one chain of iterations were only performed in a population containing 1000 diploid individuals for a single locus of 50000 bp with a mutation rate of 0.01 per site. Outputs are given in VCF files. To analyze how the novel measure performed, several simulations were done on different regions with high/low recombination, like larger regions with a stable rate of recombination, regions with 2 different rates, small regions with high recombination (hotspots) to search for peaks and different peak sizes depending on the recombination rate. `SFS_CODE` takes a recombination map file which employs crossing-over and gene conversion. This file contains the number of sequence intervals (regions with different recombination probability), the base pair position and its cumulative recombination probability until this site. Each simulation was performed three times with the same recombination map file.

4.0.4 *LD heatmap*

Haploview has a limitation of 100kbp (100 kilo base pairs; $100x10^3$ bp) for the graphical representation of the region. This limit can be overcome with higher computing resources and a specific command of haploview, but in this case the resources were far from sufficient. As a result, the visualization of haploblocks from the measures of LD was not possible us-

ing this software. A pipeline was developed to be able to visualize these LD patterns on larger regions. To subset large regions, like a substantial part of a chromosome, the use of bcftools[2] was needed in behalf of its fast subsetting and filtering command (*view*). This tool allows to subset the VCF file into smaller regions with a filtering option of keeping only the SNPs that have an AF higher than a threshold value. The subsetting will require the tbi file of the VCF file to retrieve the information faster. Then, vcftools[3] will get as input the VCF file originated from bcftools and output the Pearson's squared correlation coefficient (r^2) in a tab-delimited file containing the chromosome, the SNPs positions of the two SNPs compared and the corresponding r^2. With the outputted LD file, a Python script was written to be able to read the file and extract the useful data. Additionally, the data is composed into a matrix and then with the resources of plotly[4], a heatmap was built. The heatmap contains a color scheme, which the blue color represents lower values of r^2 and red represents higher values of r^2. Seemingly, haploblocks are represented as triangular red regions.

2 https://samtools.github.io/bcftools/
3 http://vcftools.sourceforge.net/
4 https://plot.ly/

5

RESULTS AND DISCUSSION

All the plots shown are the mean of 14 values (windows) for the measures of Cycles/Region and Mutations/Region. From the original tab-delimited files all the values are retrieved with an overlap of 7 values. This was done so that a tendency across a given portion of the chromosome was obtained and so that individual single high values do not generate a disproportionate peak that could actually be caused by sequencing faults (as the 1000genomes data is low coverage data).

5.1 LINKAGE MAPS FROM THE STUDY ON CHROMOSOME 22

Chromosome 22 is the most recombinant chromosome on humans, having a recombination frequency of 2.46cM per Mbp against the genome average of 1cM per Mbp [74]. This is expected as it is the shortest chromosome, so if homologous chromosomes connect during meiosis at least in one point and crossing over occurs, the rate of recombination will be higher considering its size. Chromosome 22 due to its high recombination rate and the smallest size (for computational reasons) is a good model to test the new approach.

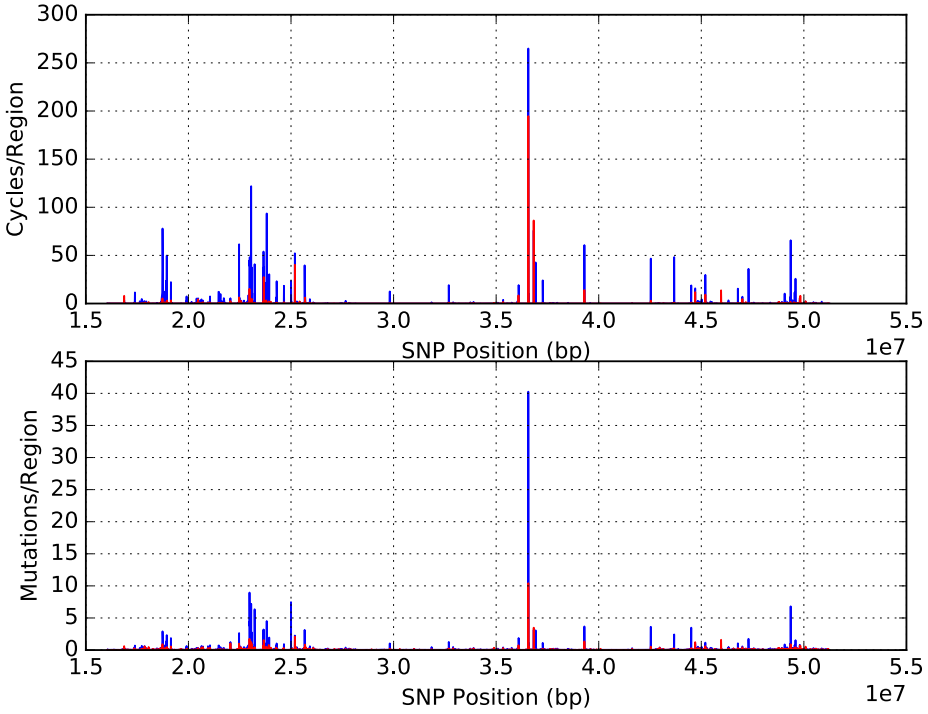


Figure 9.: Linkage map for chromosome 22 with 30 SNPs (blue) and 20 SNPs (red), overlapping by 15 and 10, respectively.

Figure 9 shows that independently of the measure employed (cycles or number of mutations divided by size) or the fact that measures are based on 20 or 30 SNPs the maps are very similar. Also it becomes really explicit that along the chromosome it is possible to identify regions with low levels of reticulation in the region, likely meaning very low recombination, as well as regions displaying peaks that could represent recombination hotspots. Throughout the thesis the maps of these measures will be generally referred as "linkage maps" although the first map could better represent a recombination rate map and the second, based on the excess of mutations in the phylogenetic reconstruction is more difficult to define.

It becomes really explicit that along the chromosome it is possible to identify regions with low levels of reticulation in the region, likely meaning very low recombination, as well as regions displaying peaks that could represent recombination hotspots. At least four or five regions with putatively high recombination rate are visible, corresponding to various peaks on the same general area (namely around the positions, in bp, 2.2×10^7 to 2.6×10^7 and just after 3.5×10^7).

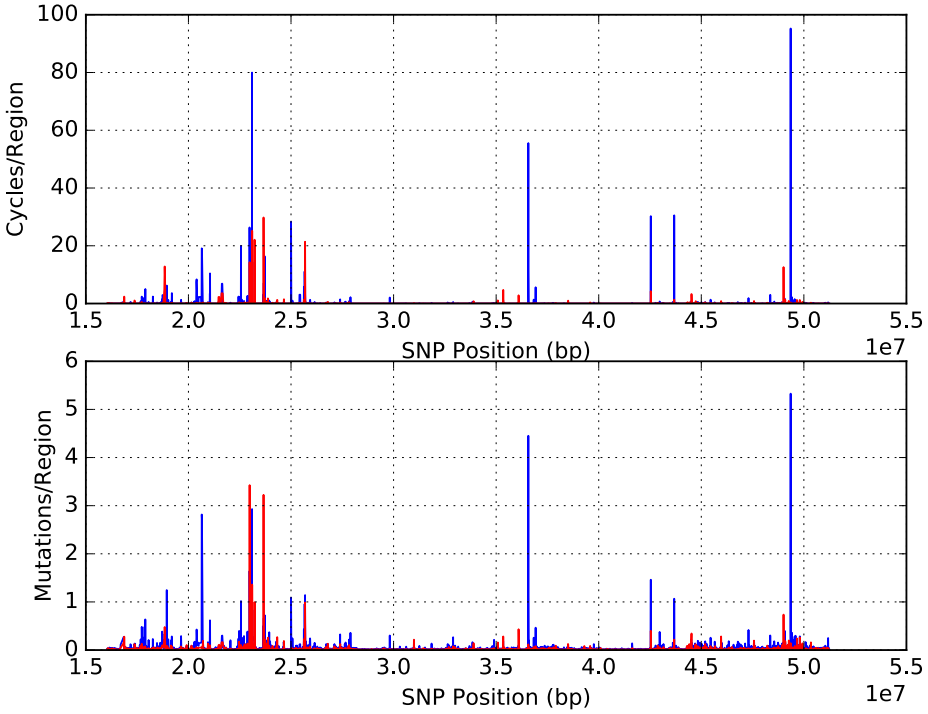


Figure 10.: Linkage map for chromosome 22 with 20 SNPs, overlapping by 10. African individuals (blue) and European individuals (red).

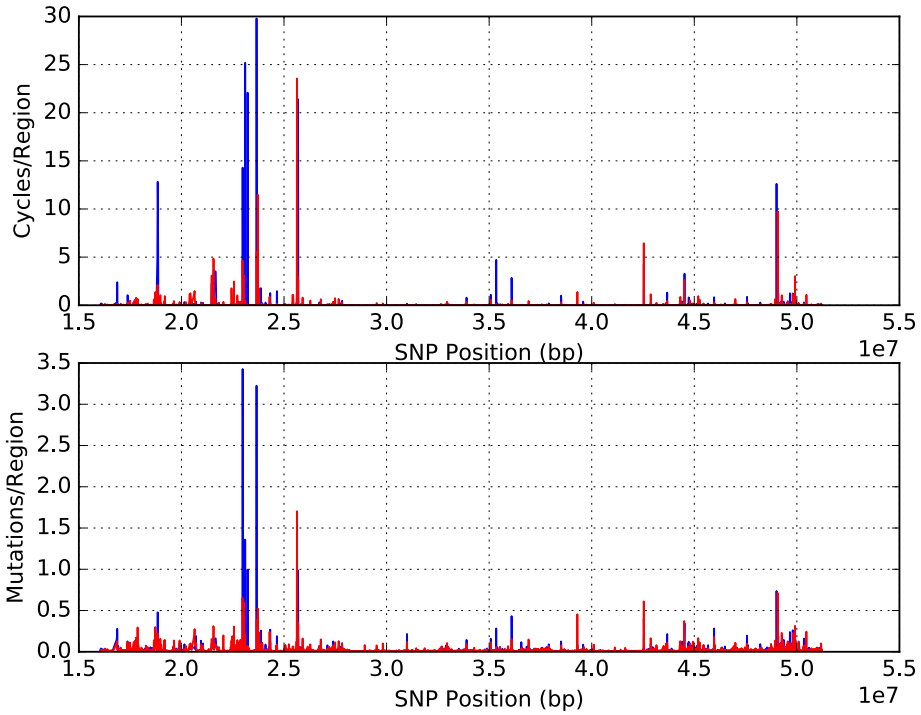


Figure 11.: Linkage map for chromosome 22 with 20 SNPs, overlapping by 10. European individuals (blue) and South Asian individuals (red).

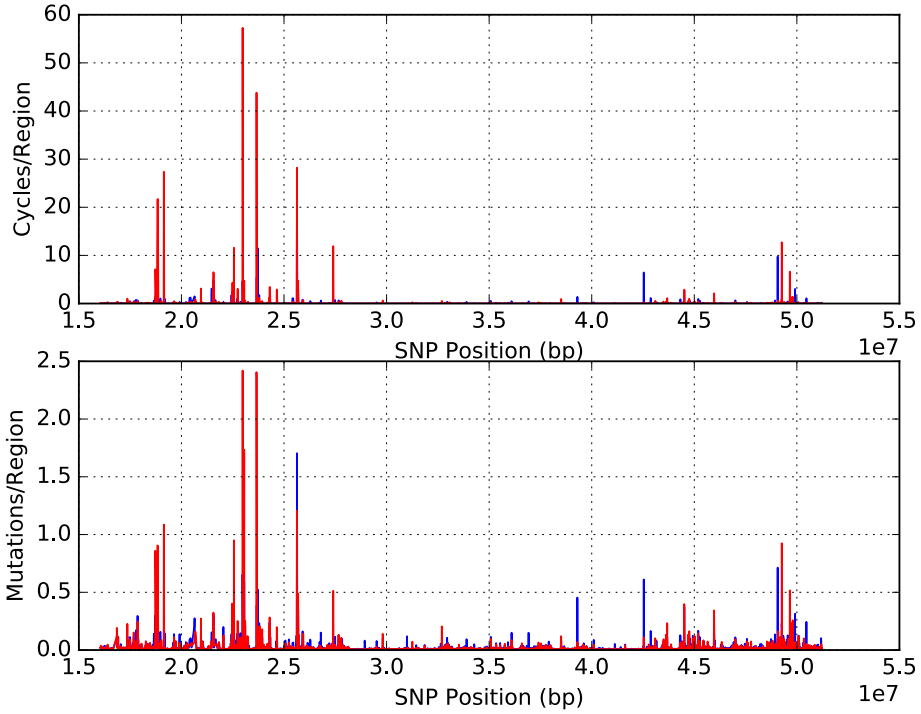


Figure 12.: Linkage map for chromosome 22 with 20 SNPs, overlapping by 10. South Asian individuals (blue) and East Asian individuals (red).

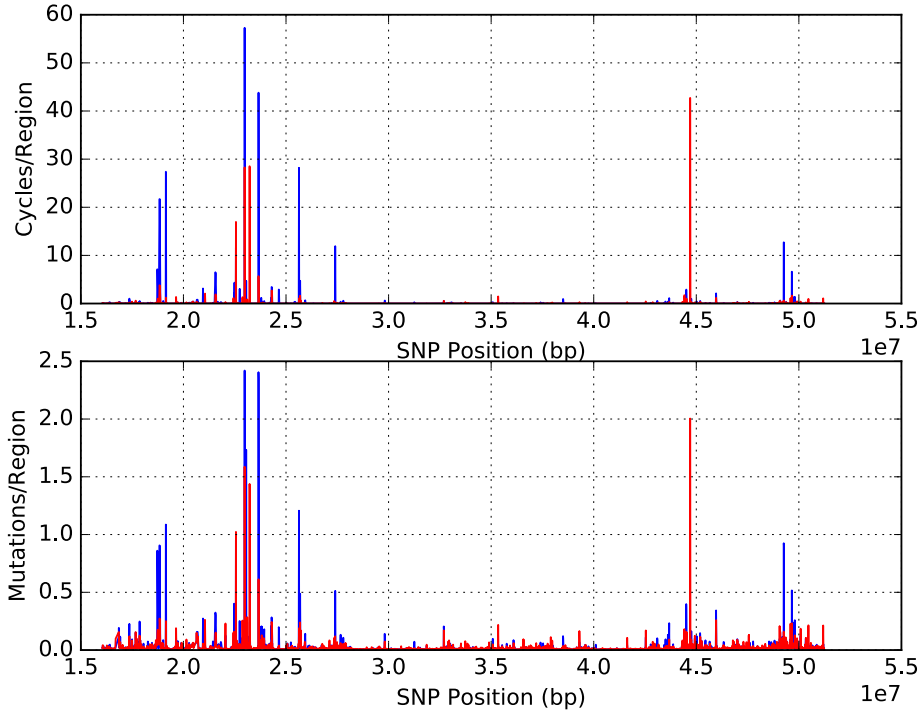


Figure 13.: Linkage map for chromosome 22 with 20 SNPs, overlapping by 10. East Asian individuals (blue) and American individuals (red).

Another relevant analysis is the comparison of maps between networks established on a single population. (Figures 10, 11, 12, 13). The objective is to observe if recombination maps varied drastically depending on the dataset, especially considering that one of the advantages of this methodology is that it should not be so dependent on allele frequencies. The maps did not vary significantly between population. For example, all graphics show a region between 20Mbp and 26Mbp (base region) that has a higher recombination frequency than other regions on the plot independently of the population analyzed (Figures 10, 11, 12, 13).

A comparison between African (AFR) and European (EUR) populations shows that, as expected, observed recombination is larger in African populations (Figure 10). Although it is evident that the recombination rate on individuals with African ancestry on chromosome 22 is higher than European individuals, being especially clear on the region 22Mbp and 26Mbp. The higher peak of recombination (hotspot) is almost 100 Cycles/Region for AFR and 30 Cycles/Region for EUR populations. This observation is in line with the fact that modern humans had an origin in Africa while outside Africa a bottleneck occurred when starting the populating of the globe that would partially increase the linkage.

Apart from a few larger peaks, the recombination frequency is very similar between South Asian (SAS) populations and European populations (Figure 11). The difference between SAS and East Asian (EAS) populations is narrow, which is expected considering that even though South Asia was settled before East Asia, it was likely a very rapid migration [75, 76] leaving little time for differentiation (Figure 12). The Americas were likely colonized by modern humans from a Northeast Asian source [76]. It is clear that apart from a few peaks, America has a lower base recombination rate than East Asia (Figure 13). With the course of time and space between the possible movement of humans from Africa to America, the observed recombination is thought to decrease which is caused by successive bottlenecks. Nevertheless, maps are not extremely different between populations.

5.2 LINKAGE MAPS COMPARED TO HEATMAPS

In order to compare this new methodology with methodologies that have been established for a long time in genetics, like traditional linkage disequilibrium measures, a comparison of different regions will be established. It is impossible to compare the full chromosome for r^2 since software tools like Haploview would just allow 100kbp, that is a very small size in a chromosomal context. The pipeline developed here allowed to build LD maps for 20 times larger regions than haploview. The blue color represents lower values of r^2 and red represents higher values of r^2. Following we will compare different stretches of chromosomic regions using the linkage map developed here and the heatmaps. For a simplistic descriptive statistic of the region, an average of the Cycles/Region and linkage measure for that portion will be presented. The examples aim to provide instances of agreement and disagreement between the two methodologies.

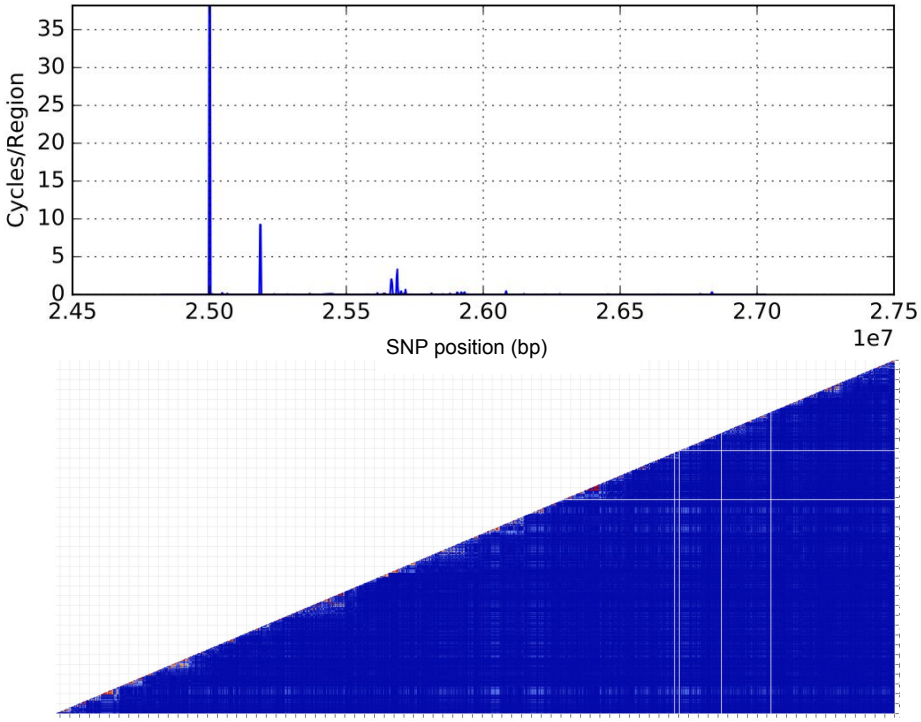


Figure 14.: Linkage map and heatmap for chromosome 22 on a region between $2.45x10^7$ bp and $2.75x10^7$ bp.

The mean Cycles/Region for the linkage map (Figure 14) is 0.17. Comparatively, this is a region with relatively high recombination rate considering this measure. Analogously, the heatmap does not have any significant haploblocks, hypothetically due to the high recombination rate.

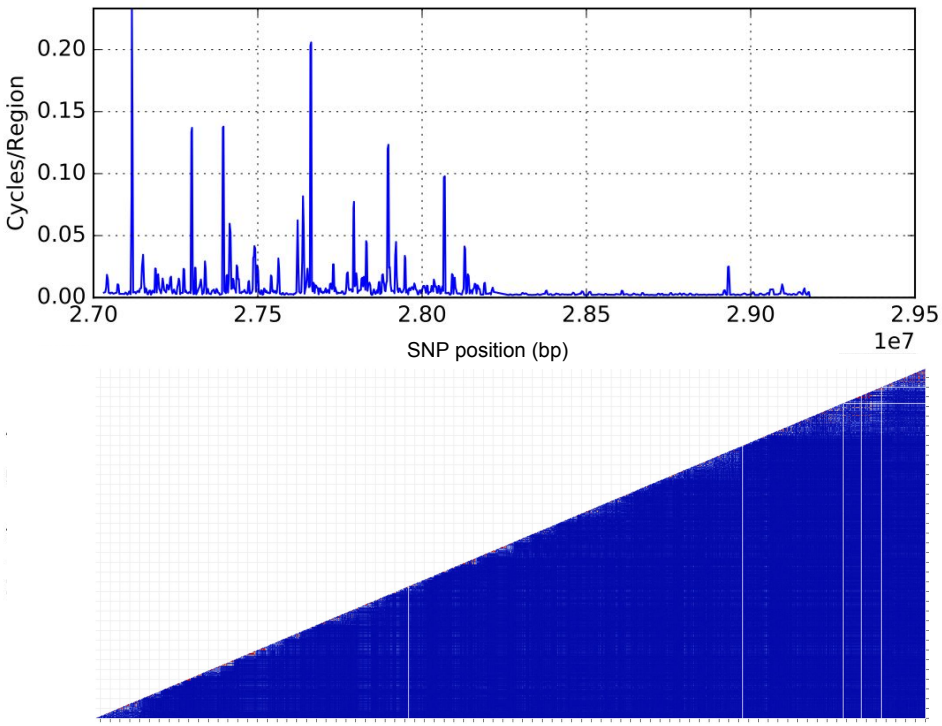


Figure 15.: Linkage map and heatmap for chromosome 22 on a region between 2.70x10^7 bp to 2.95x10^7 bp.

On figure 15 the mean recombination frequency based on cycles is approximately 10 times lower than the region displayed in figure 14, but there are not relevant haploblocks present. It could be explained by a not very accurate calculation of LD (using r^2) or the linkage map not being correct in that region (possibly due to a small number of SNP on the overlapping windows). This is the region that shows lower correlation between the results we obtained with r^2 and the new methodology. It is not straightforward to explain and deserves a detailed examination in the future in terms of frequency of the different alleles along the region, since a first analysis does not reveal anything particular.

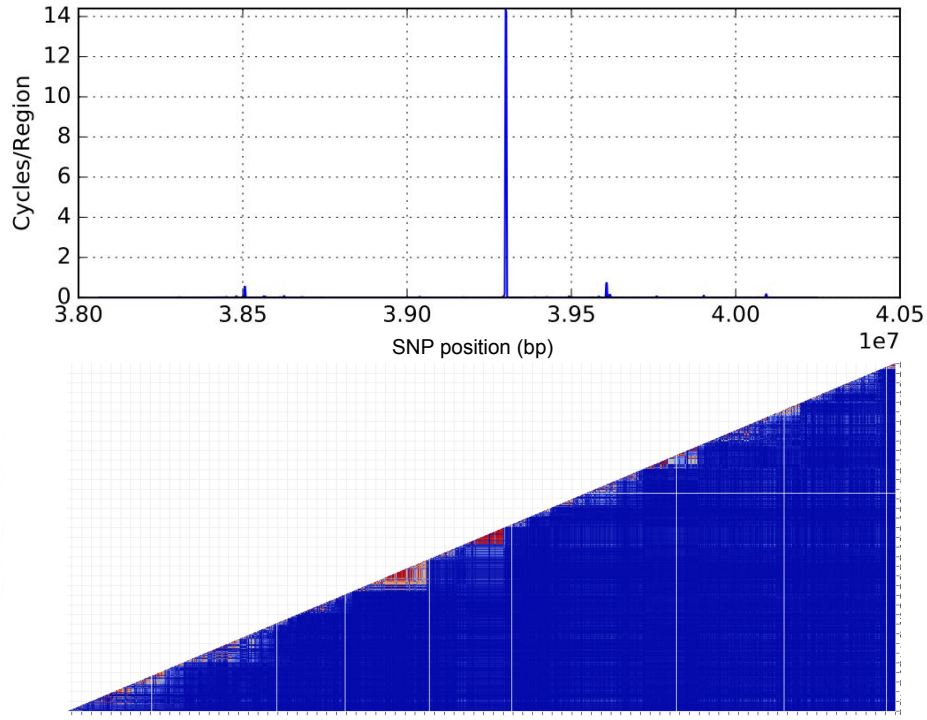


Figure 16.: Linkage map and heatmap for chromosome 22 on a region between $3.80x10^7$ bp to $4.05x10^7$ bp.

The region displayed in plot 16 has an average recombination frequency of 0.046, yet this value is high due to a peak of approximately 14 Cycles/Region that is increasing the overall recombination rate. When observing the heatmap it is possible to see that some haploblocks are found, specially two stronger ones between 3.88 to 3.93 confirming the low recombination rate in the region. However, right after these blocks r^2 is low, corresponding to the peak in the new methodology.

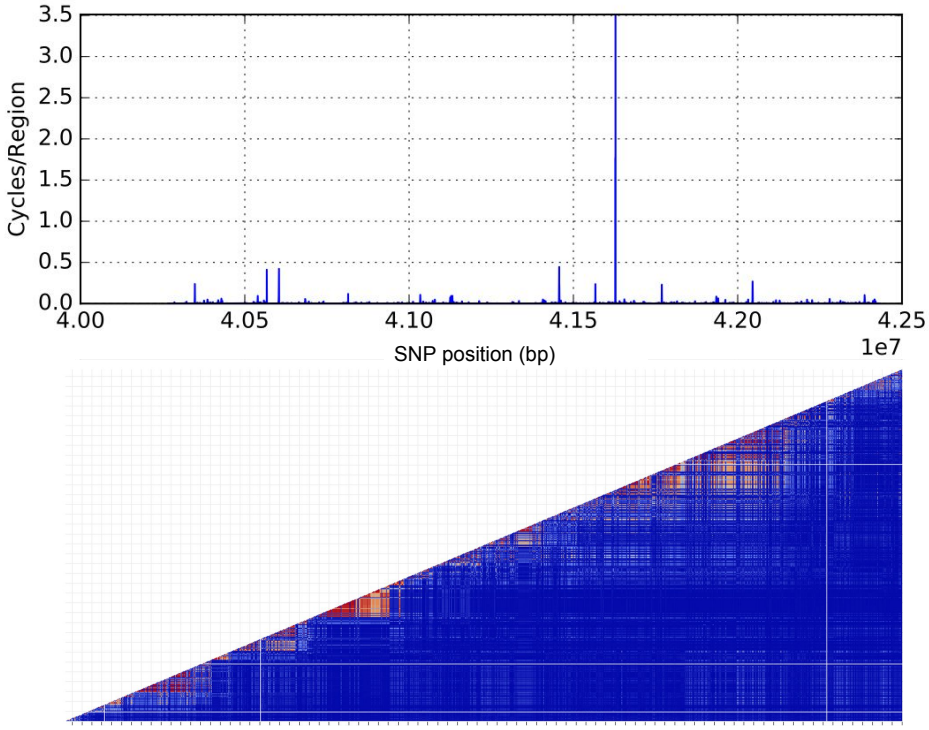


Figure 17.: Linkage map and heatmap for chromosome 22 on a region between $4x10^7$ bp to $4.25x10^7$ bp.

The final example corresponds to a region whose heatmap (Figure 17) suggests the presence of many haploblocks and specially a large triangular region of high LD. The mean value of Cycles/Region for this linkage map is 0.0038, significantly lower than the first plots correlating well with the heatmap.

Overall, the regions displaying low recombination in the Cycles/Region measure display a higher rate of haploblocks.

5.3 SIMULATION RESULTS

A validation of the method was attempted using simulations. All plots (linkage maps based on Cycles/Region and mutations/regions) of simulations are on the original format, without displaying any averages. Each color represents a different run of the simulation; all plots have 3 runs. The first run has the blue color, the second is red and the third is green. Following, different conditions will be used in order to establish how measures like the Cycles/Region measure are directly dependent of the recombination rate. A higher value of Cycles/Region can translate on a greater recombination frequency on that region.

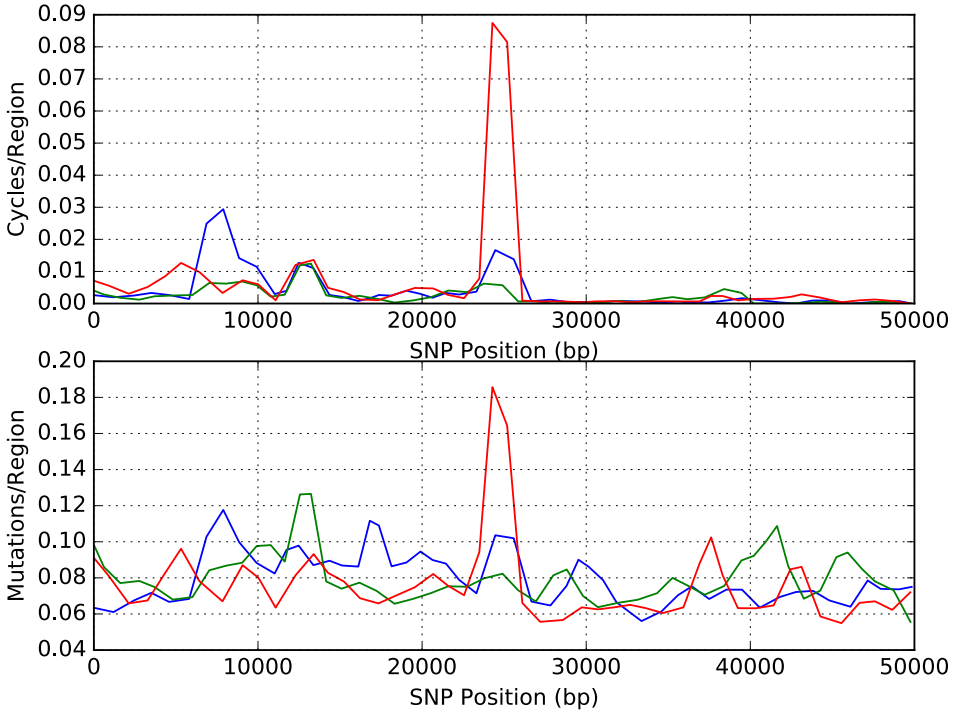


Figure 18.: Simulation for 2 hotspots.

In this simulation a VCF file was created with two hotspots, the first one as a region of 5kbp, from base 5k to 10k, with a low recombination probability of 0.10, the second one with a region of 10bp (from base 25000 to 25010) with a recombination probability of 0.65 and a base static recombination probability of 0.05. The results (Figure 18) are not clear considering each run individually, but overall the three plots considered together have two peaks with different sizes, mostly when considering the Cycles/Region measure; the first has a smaller height but wider due to a larger region of recombination and the second is exactly the opposite (higher but less wider).

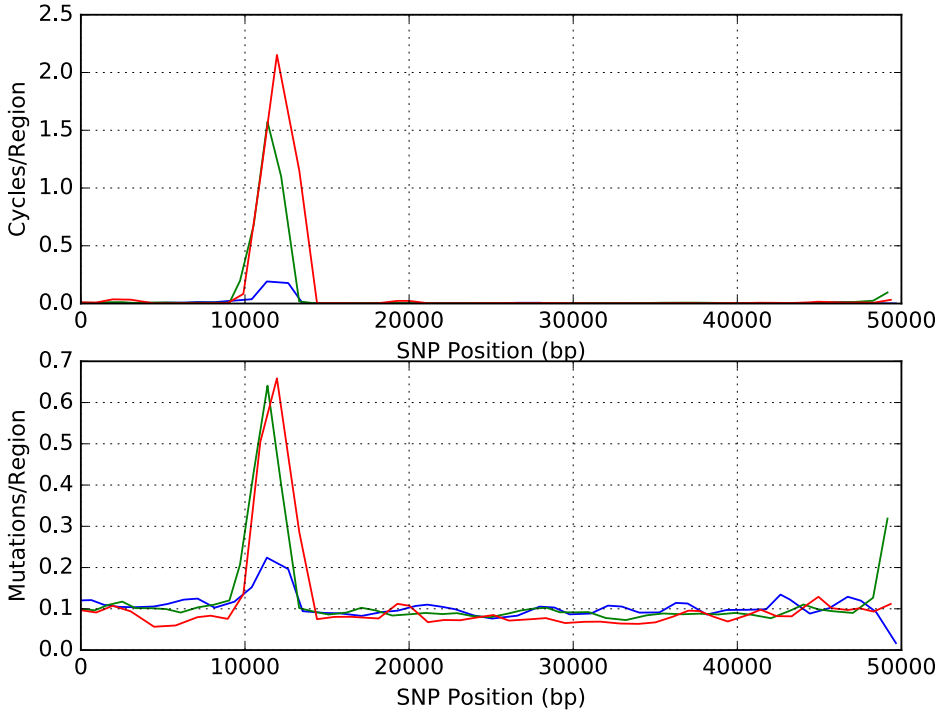


Figure 19.: Simulation for 1 small platform.

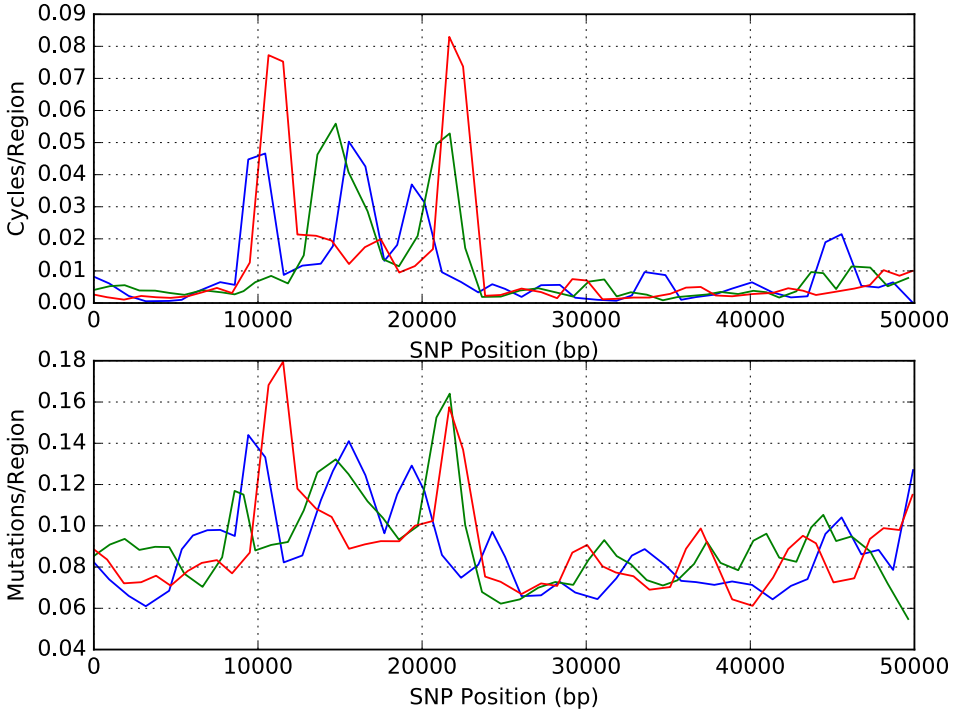


Figure 20.: Simulation for a platform with double the length of figure 19.

In order to see how recombination and our measurement behaves on larger regions with a constant recombination probability, two simulations were performed for these cases. Figure 19 shows the simulation for a region of 2500bp (from base 10000 to 12500), with a recombination probability of 0.6 (recombination hotspot) and a base recombination probability of 0.1 for the remaining wider region. The results clearly show the hotspot with the absence of any peak within the wider region, that has a constant but much lower recombination probability.
The second plot 20 with the same base recombination probability of 0.1 and 0.6 for the higher recombinant region, but with a wider recombinant area of 12500bp (from base 10000 to 22500). Again, when taken together, the three runs clearly delineate the region with higher recombination probability and show an absence of peaks outside that region.

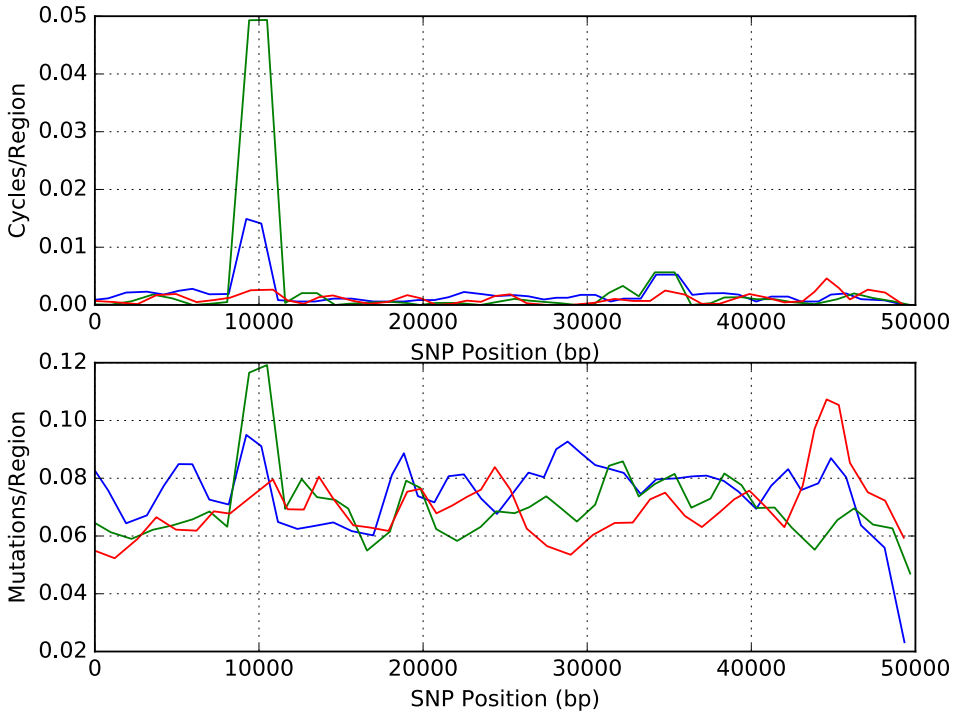


Figure 21.: Simulation for 2 hotspots.

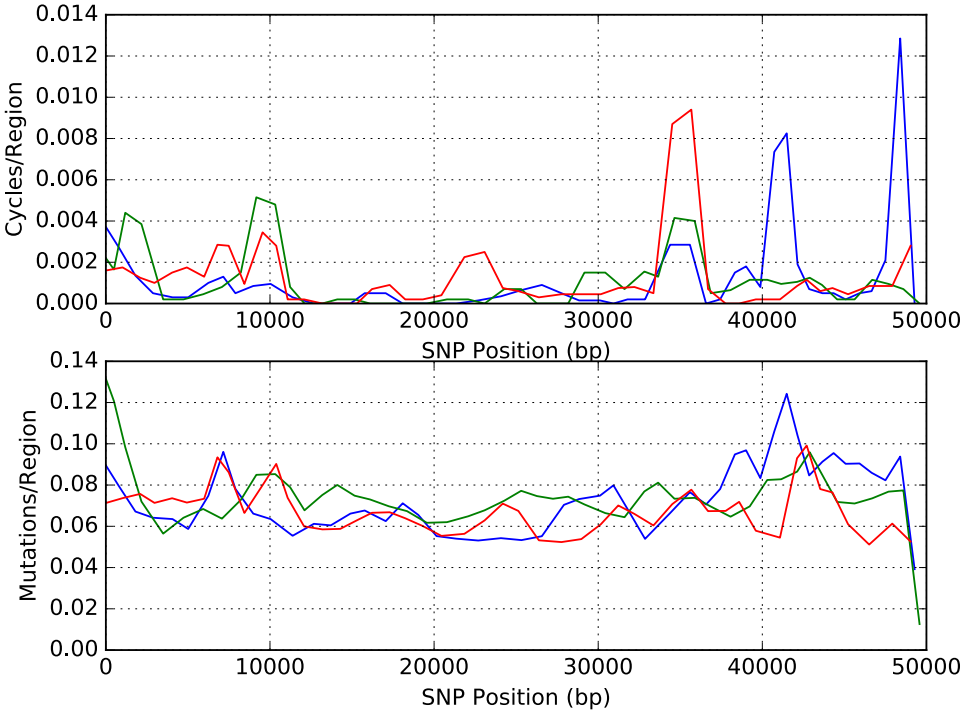


Figure 22.: Simulation for 2 hotspots.

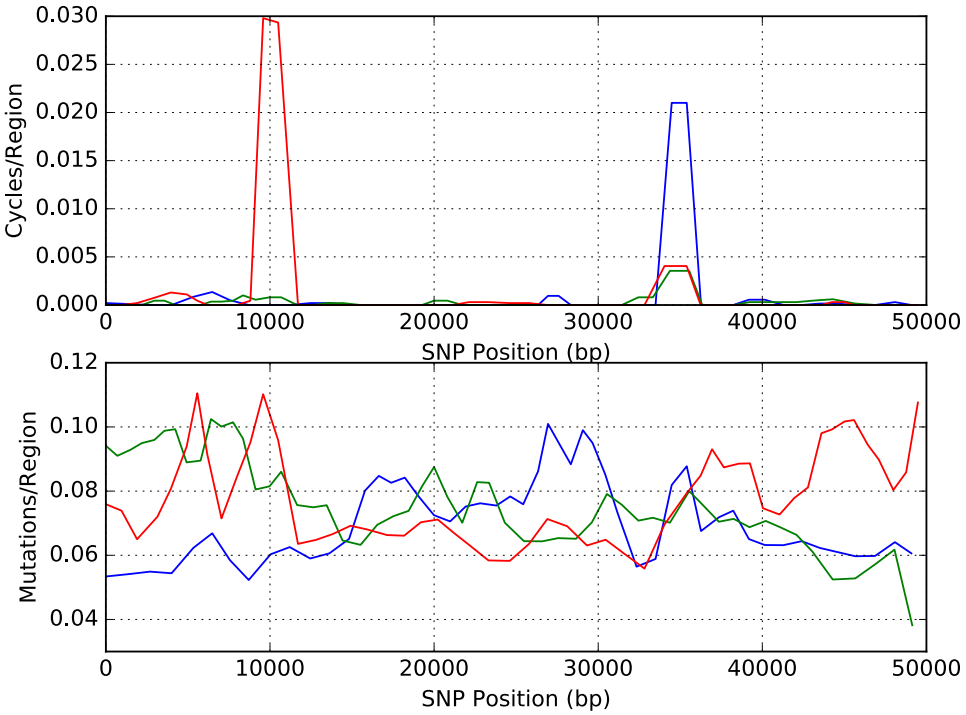


Figure 23.: Simulation for 2 hotspots.

Finally, simulations displayed in figures 21 to 23 aimed at looking of the effect of considering very specific hotspots with a very small size. Simulation 21 has 2 hotspots with the same length of 10bp, the first has a recombination probability of 0.3 located at 10000bp and the second a probability of 0.55 located at 35000bp. Simulation 22 has 2 hotspots with the same length, the first has a recombination probability of 0.2 and the second a probability of 0.55. The base recombination rate for the remaining regions in both simulations is 0.05. Simulation 23 has a lower base recombination rate of 0.01 and 2 hotspots with a recombination probability of 0.49. While peaks are discernible in very few simulations, these simulations with a very small region for the hotspot are not very informative.

Hotspots with larger regions are much more easily detected on the linkage map. For this reason, hotspots with a length of 2500bp are always represented by a peak or several peaks, instead of hotspots with a 10bp region that are more dependent of the base recombination rate. Nevertheless, recombination hotspots are likely to be represented by regions of the genome with considerable sizes. Therefore, more simulations with larger regions of recombination needed to be performed. However, the first simulations were promising but need to be improved. It is likely that resolution of most of the simulations was not enough, which is reflected on the low number of points represented in each plot. Further simulations need to be reproduced considering larger regions with more variation and individuals.

5.4 Linkage maps from the study on the X chromosome

Although we centered the analysis on chromosome 22, chromosome X was also analyzed due to two important features in the context of this work. The first one is that we can analyze the chromosome independently in males and females. This is important since while X-chromosome in females is phased based on phasing algorithms like SHAPEIT2 (as in the case of chromosome 22 and the other autosomes), in the case of the X-chromosome in males, one can obtain directly the haplotype. This allows to show how reliable the phasing estimate is, since the work highly depends on that. The second feature is that X-chromosomes have Pseudoautosomal regions (PARs) that links with the Y-chromosome. Mostly one, PAR1, is probably the overall region of the genome with higher recombination rate. The detection of this PAR region would also be a validation for the method.

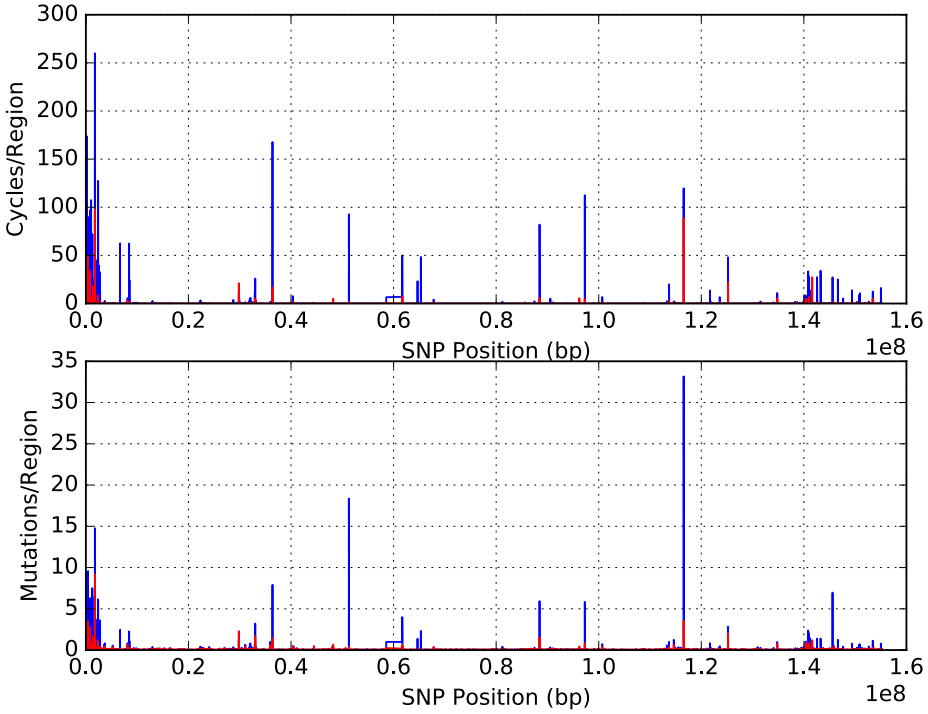


Figure 24.: Linkage map for the X chromosome of male individuals with 30 SNPs (blue) and 20 SNPs (red), overlapping by 15 and 10, respectively.

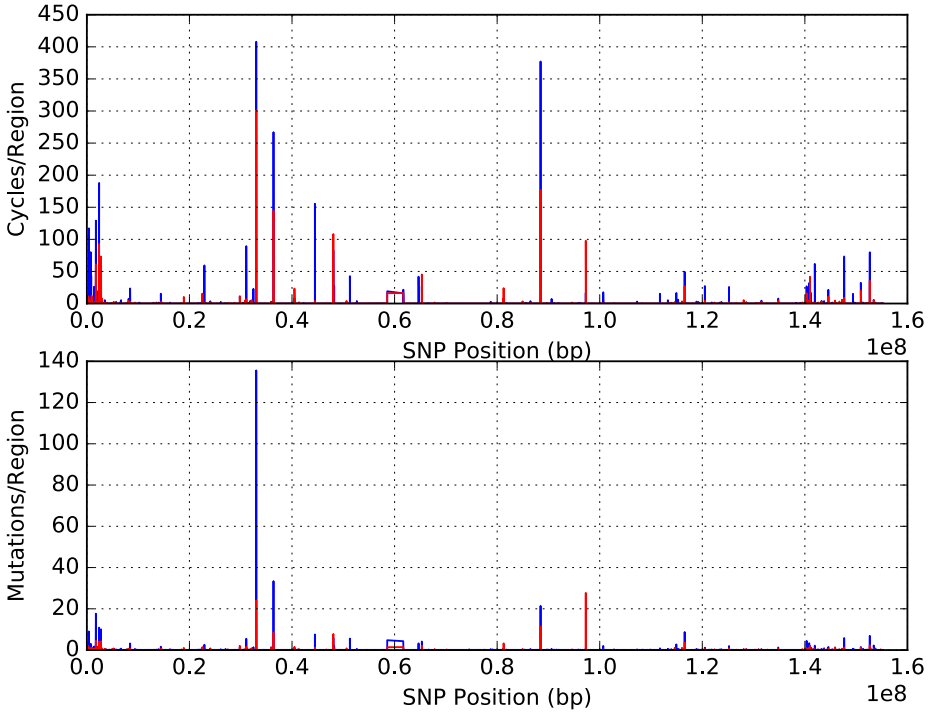


Figure 25.: Linkage map for the X chromosome of female individuals with 30 SNPs (blue) and 20 SNPs (red), overlapping by 15 and 10, respectively.

Recombination frequency at the beginning of the X chromosome for both males and females, as seen on figure 24 and 25, is very high overall. This shows a clear representation of PAR1. The tip of the long arm (PAR2) has a smaller recombination frequency than PAR1 and it does not display a recombination rate substantially higher than the remaining chromosome, which is not unexpected from previous studies.

In terms of comparing the maps in both males and females there are some odd peaks but most of the features are similar in both analysis, including a region displaying probable high recombination starting at 1.4×10^8 bp but also around 0.3×10^8 bp and 0.6×10^8 bp. The major differences are peaks just before 1.2×10^8 bp in male data and at about 0.9×10^8 bp in females. Overall the maps are very similar.

5.5 RETICULATIONS ON DIFFERENT POPULATIONS

Recombination rate is an important force in shaping the genome. It is expected that recombination occurs at a relative steady rate along time. In this section the aim is to explore the possibility of using recombination as a molecular clock, meaning check how the amount of

probable recombinants can help in understanding the divergence between populations.
All cycles for each population along chromosome 22 were counted, including duplicates. Duplicates happen as a result of cycles forming a cube, repeating each cycle two times. It was only considered cycles with 4 nodes, not passing through other nodes. The total number of cycles in each population is displayed in Table 4.

Table 4.: Results for the number of cycles in each super-population

Cycles	Ancestry
11030525	AFR
7564062	EUR
5129539	SAS
8741894	EAS
3955390	AMR

As expected, African populations, the continent where modern humans emerged, displayed a much higher number of reticulations, which is expected as the evolution of the human diversity is much deeper there than anywhere else worldwide. European and East Asian populations have higher values than expected considering South Asia. Excluding the duplicates, the number is substantially lower (Table 5) and the difference is minimum between Europe, South Asia and East Asia which makes sense considering that most probably all these regions were colonized with a small time difference.

Table 5.: Number of cycles present on a population that are on another population, without duplicates.

	AFR	EUR	SAS	EAS	AMR
AFR	162782				
EUR	27123	94633			
SAS	27658	35649	99271		
EAS	22750	28991	31828	91419	
AMR	37166	36995	33778	28209	96996

If we consider that the founder population that migrated from Africa could have already carried cycles in their genomes that did not occurred out-side Africa, we can exclude cycles that are present between Africa and Europe, South Asia and East Asia. Individuals with European or South Asian ancestry have similar values of cycles that are present in individuals

with African ancestry, meaning that 27k cycles from each population comes from African populations. East Asian individuals have less cycles belonging to an African ancestry (22k), but these individuals are already the result of a further migration that moved through South Asia, where it is possible to see that they share over 30k cycles. The American samples have a higher number of cycles than expected, this could be explained by a biased number of individuals in America that are African-American and European descendants; and an inferior number of samples.

Table 6.: Results for the number of cycles in each super-population

Cycles	Ancestry
162782	AFR
67510	EUR
71613	SAS
51420	EAS
30336	AMR

We can establish a hypothetical time-frame based on the number of cycles where the African count is direct (it is the ultimate source of original diversity), where common cycles with Africa were excluded from European and South Asian populations, where common cycles with Africa and South Asia were excluded from the number of cycles in East Asia and finally where the private cycles in the American population were the only ones considered (Table 6).

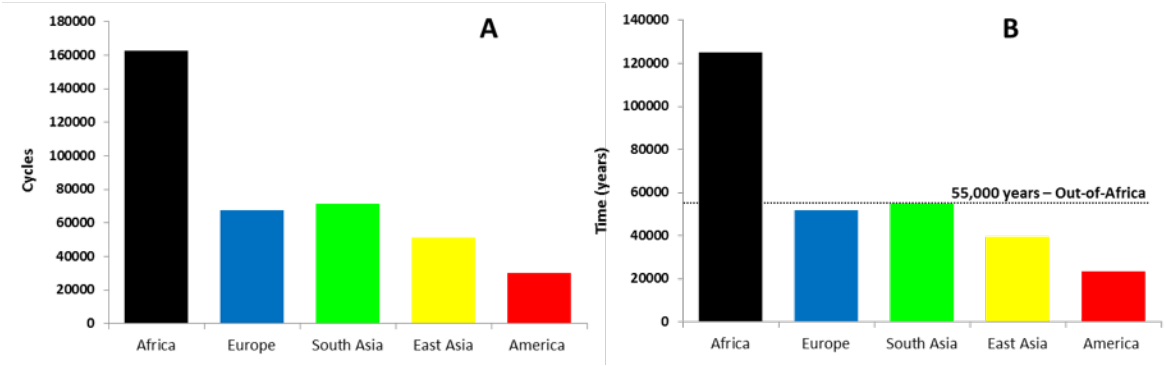


Figure 26.: time-depth.

It is possible to establish depth based on the amount of cycles. Following the establishment of a model of human migrations, a depth in terms of cycles was obtained (Figure 26-A). If we assume a specific point of colonization, it is viable to transform this measure

based on cycles into time-depth, considering the sample size for each continent is not so divergent. A hypothetical time of settlement of 55 thousand years was used for the colonization of South Asia following the Out-of-Africa migration [77]. Using this point of calibration, we can establish a settlement of Europe at about 50 thousand years, a settlement of East Asia at about 40 thousand years and a settlement of the Americas at about 20 thousand years (Figure 26-B). It is noticeable the similarity between these values and the times of colonization appointed in figure 2 of the introduction.
Although the analysis is somewhat experimental and crude, it displays the possibility of using recombination events and a recombination rate as a molecular clock to date events.

5.6 LINKAGE MAPS FROM THE STUDY OF POSITIVE SELECTION

Considering that the measure might be good in detecting level of recombination in the genome, the next exploration is to test it against cases of natural selection, mostly positive selection. If one positive allele arises and increases fast in frequency in the population, time will be short for recombination to occur and most of the haplotypic background around the allele will increase in frequency with it (hitchhiking effect). The loss of diversity caused by this increment of an haplotypic background is called selective sweep. To test its performance two examples from the bibliography of positive selection were used, one related with lactose persistence in Europe and another related with resistance to Malaria in Africa.

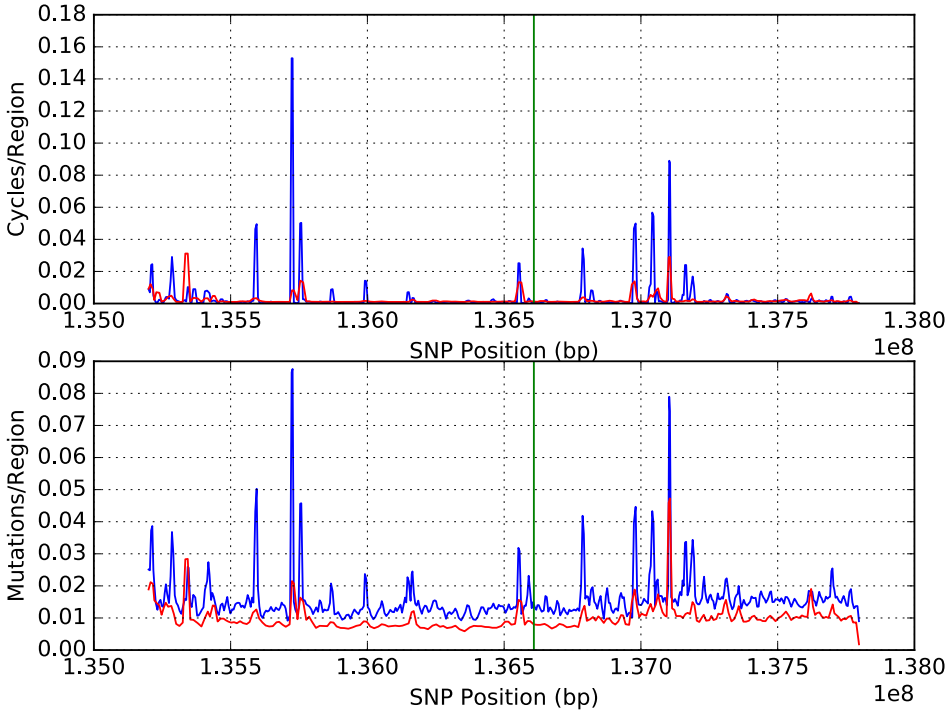


Figure 27.: Linkage map for AFR (blue) and EUR (red) with the SNP that has a direct influence on the LCT gene (green).

Figure 27 shows the linkage map for chromosome 2 on the region that contains rs4988235 (position 136608646) for the African and European ancestry. On African populations a higher recombination is seen on the adjacent region to the SNP as opposed to European populations. Considering that on European populations the beneficial allele is found on more individuals, the adjacent region will have lower recombination due to a selective sweep. If one really aims to detect the effect of selection, it is required that the two alleles (the beneficial one and the ancestor) are analyzed independently.

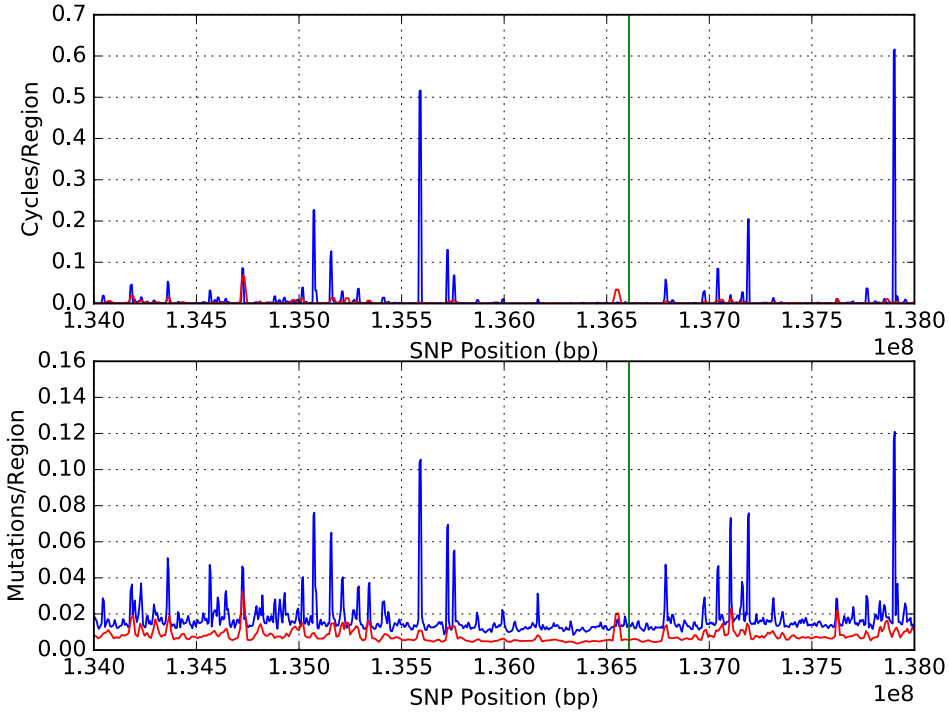


Figure 28.: Linkage map for AFR (blue) and EUR (red) with the SNP that has a direct influence on the LCT gene. Only individuals with the ancestral allele (G) (green).

Figure 28 represents a linkage map for individuals from African and European populations with the ancestral allele (G). On the European ancestry an increase on recombination rate is seen near the rs4988235. However, as a result of a selective sweep on the region neighboring the beneficial allele (A), a low recombination rate is seen on figure 29. This reflects a recent origin of the allele A, following a rapid increment in frequency in the population providing little opportunity for recombination. Again the new measures are based on estimation of recombination and not frequency of the haplotypes which can be an advantage in cases where by chance an early recombination broke part of the recombination background.

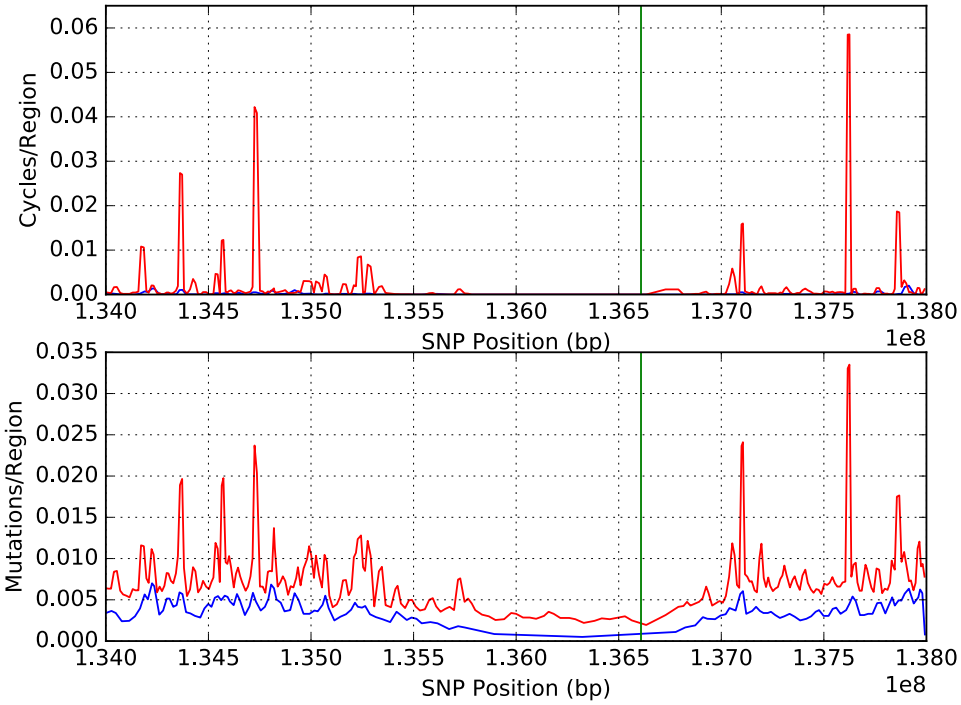


Figure 29.: Linkage map for AFR (blue) and EUR (red) with the SNP that has a direct influence on the LCT gene. Only individuals with the alternate allele (A; beneficial) (green).

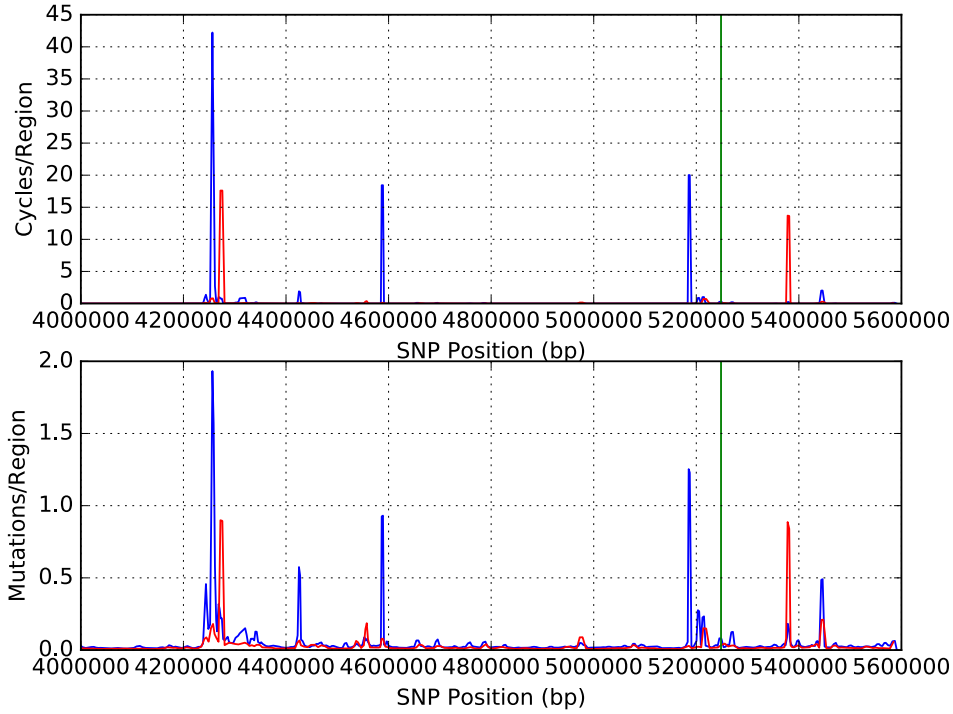


Figure 30.: Linkage map for AFR (blue) and EUR (red) with the SNP that has a direct influence on the HBB gene (green).

In terms of the HBB gene, figure 30 shows a small peak of recombination for individuals with African ancestry on the vicinity of rs334 (position 5248232).

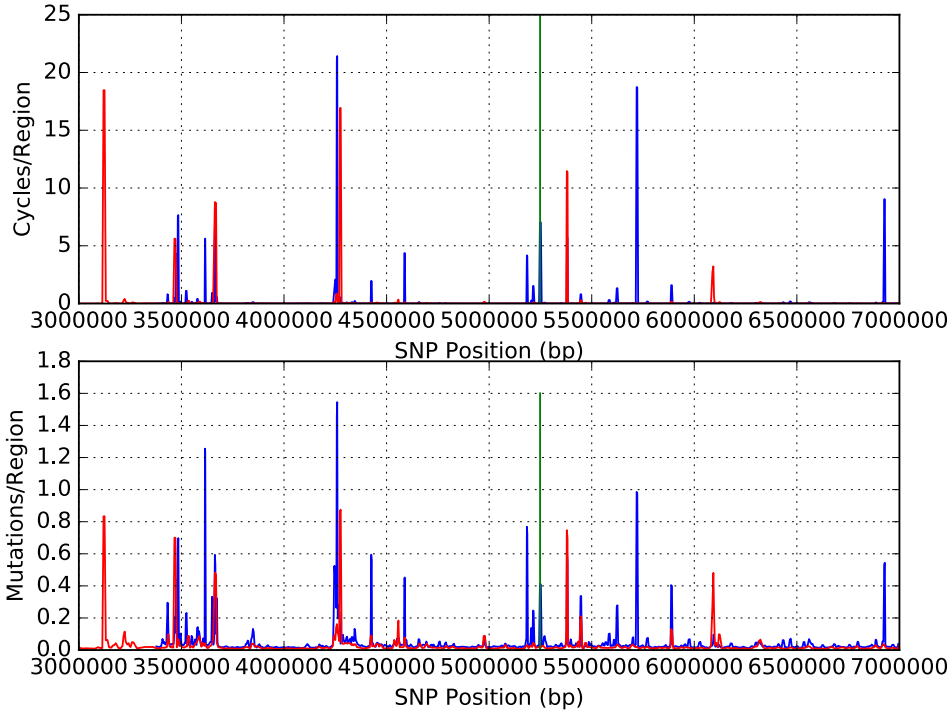


Figure 31.: Linkage map for AFR (blue) and EUR (red) with the SNP that has a direct influence on the HBB gene. Only individuals with the ancestral allele (T) (green).

Figure 31 represents a linkage map for individuals from African and European populations with the ancestral allele (T). Not surprisingly the recombination is high in the region around the SNP of interest, including peaks observed in the general map before.

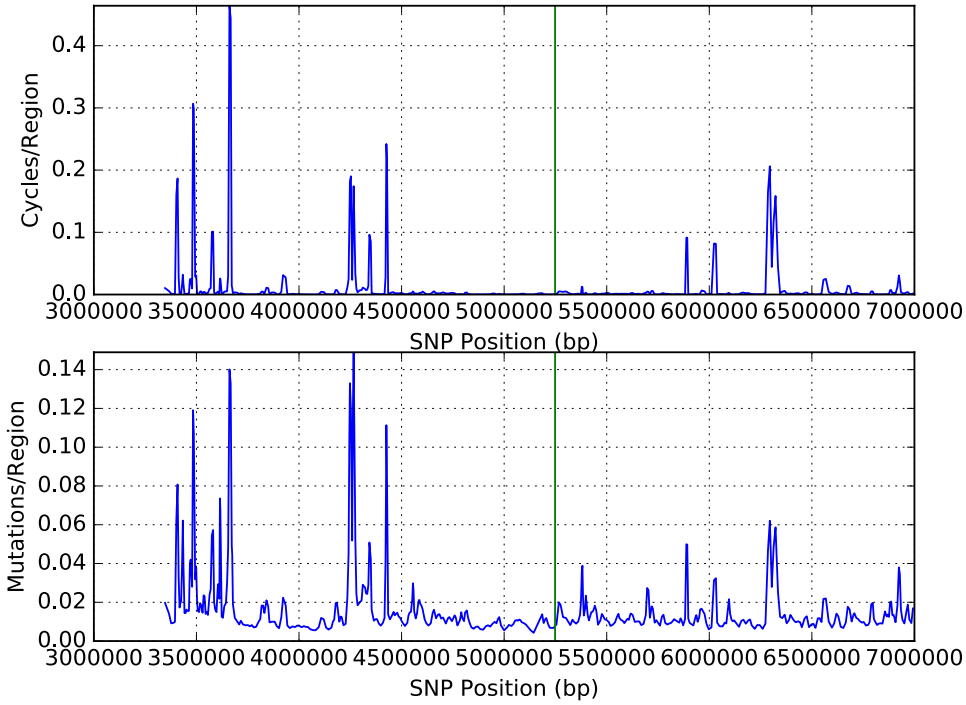


Figure 32.: Linkage map for AFR (blue) with the SNP that has a direct influence on the HBB gene. Only individuals with the alternate allele (A) (green).

The alternative allele of the rs334 (the beneficial one in Africa) does not appear on European populations (0% frequency), having 100% of the ancestral allele. Outside a Malaria endemic region, the SNP does not provide any beneficial advantage and it is deleterious. On Africa, the genetic map near rs334 is very low comparing to other regions, due to a selective sweep, and even the strong peaks around the allele that were observed around in the general map failed to appear here. (Figure 32)

To compare the last plots on the occurrence of selective sweep to an SNP that was not related before with positive selection,a new random SNP (rs2075984) from chromosome 22 was selected (Figure 33). The SNP was chosen based on the AF of the African and European populations that desirably needed to be as close as possible to the AF of the SNP that has a direct influence on the LCT gene (rs4988235). For European populations the AF is approximately 50% for both SNPs and for African populations the ancestral AF is 3% for rs4988235 and 4% for rs2075984.

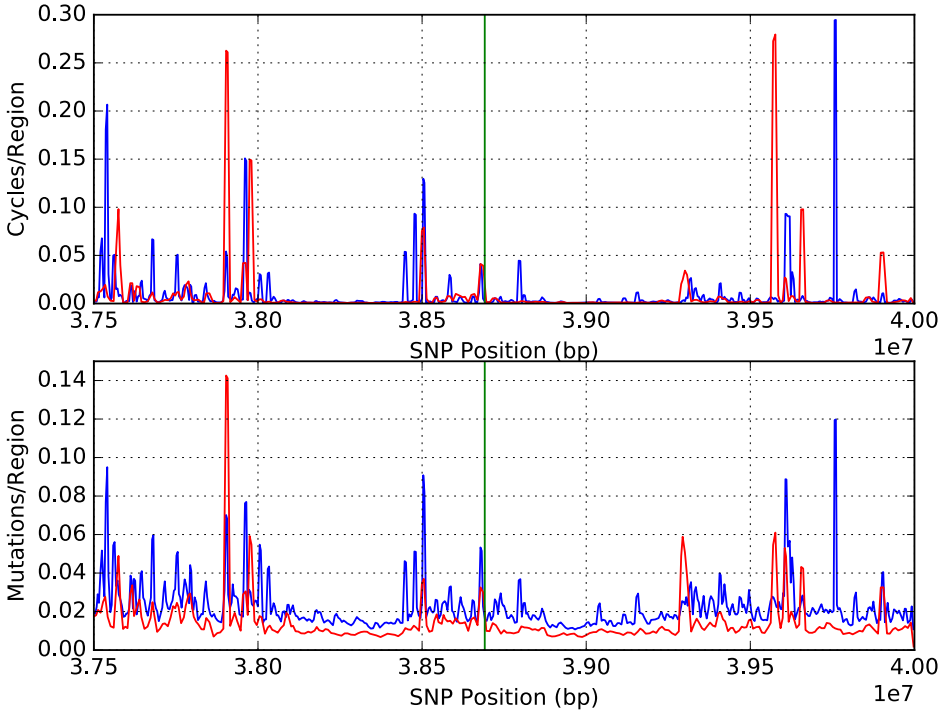


Figure 33.: Linkage map for AFR (blue) and EUR (red) with rs2075984 (green).

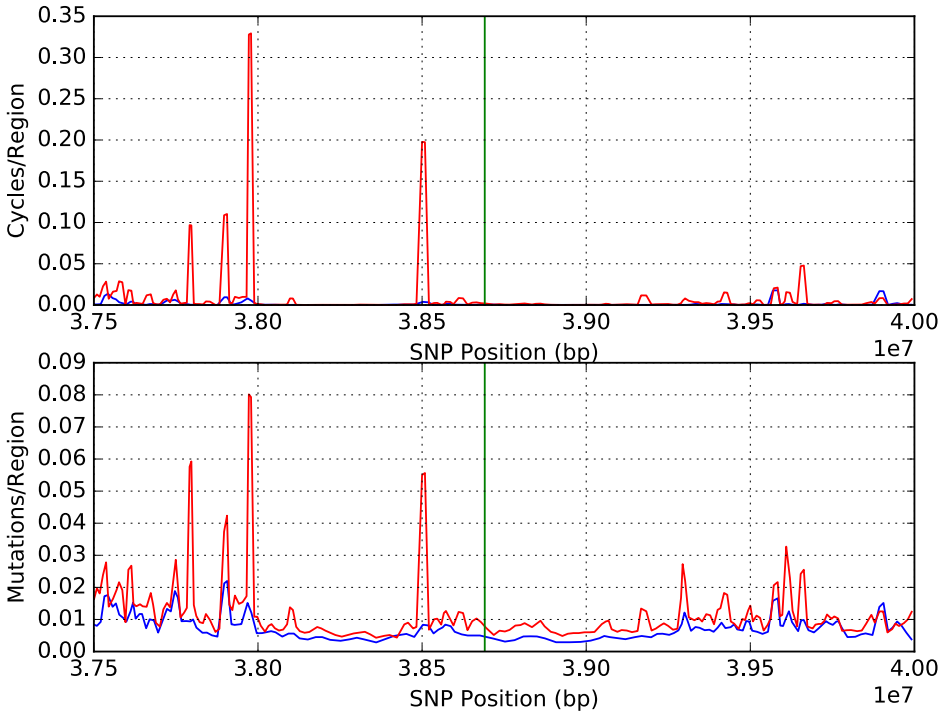


Figure 34.: Linkage map for AFR (blue) and EUR (red). Only individuals with the alternate allele (green), for the rs2075984 SNP.

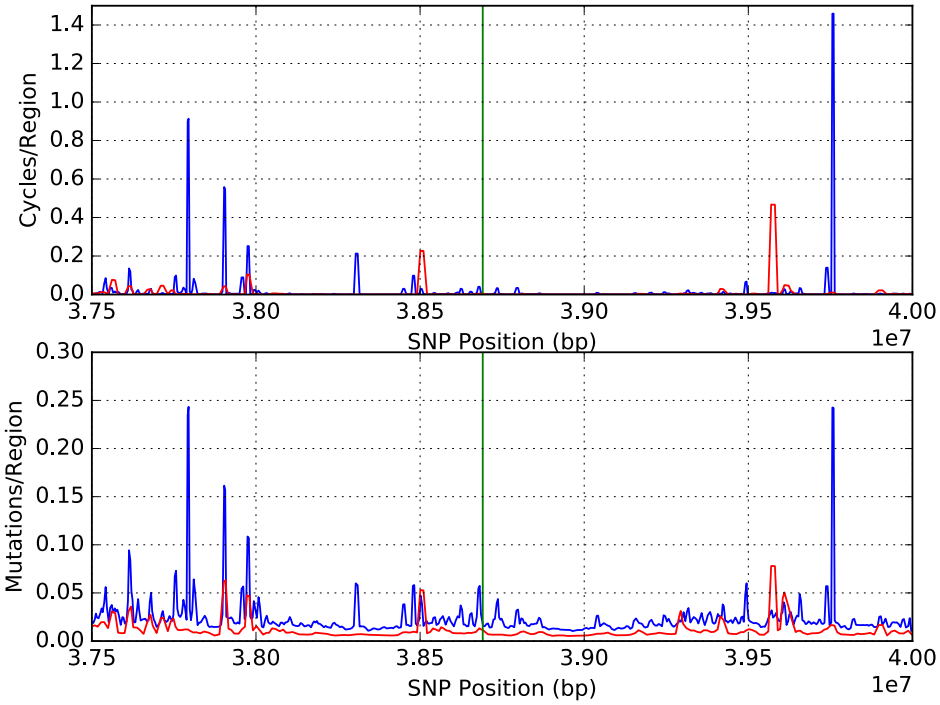


Figure 35.: Linkage map for AFR (blue) and EUR (red). Only individuals with the ancestral allele (green), for the rs2075984 SNP.

Following the examination of the overall map of the region around the selected SNP and the analysis in the ancestral and derived allele, it is possible to conclude that, although the derived allele has lower linkage (Figure 34) which is an expected feature of a derived allele that is by definition younger, the drop of linkage compared with the general map (Figure 33) and ancestral allele (Figure 35) is not so drastic. This also suggests that a further standardization on how strong the selective effect is might be required.

5.7 STARLIKENESS

In principle regions that are under positive selection will display signal of expansion which is reflected phylogenetically on a starlike figure with short sub-structuring in the region. Given this, a starlikeness measure was employed to test for possible detection of signals of positive selection. The measure was tested on the networks of the windows after these were processed into the short-spanning trees. The starlikness measure was scanned across the entire chromosome 22 and again testing it against the two genes used above.

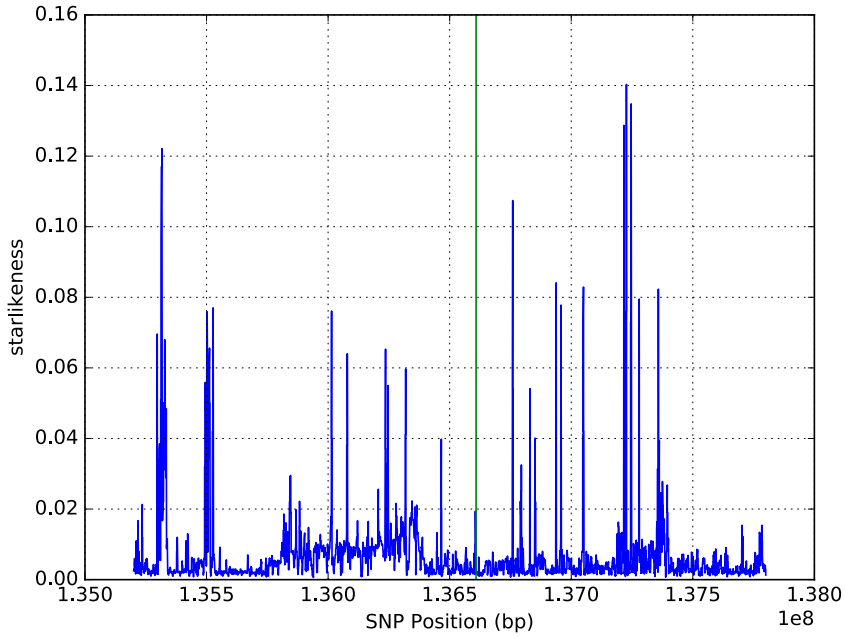


Figure 36.: Plot showing the starlikeness through the region with the SNP that has a direct influence on the LCT gene (green).

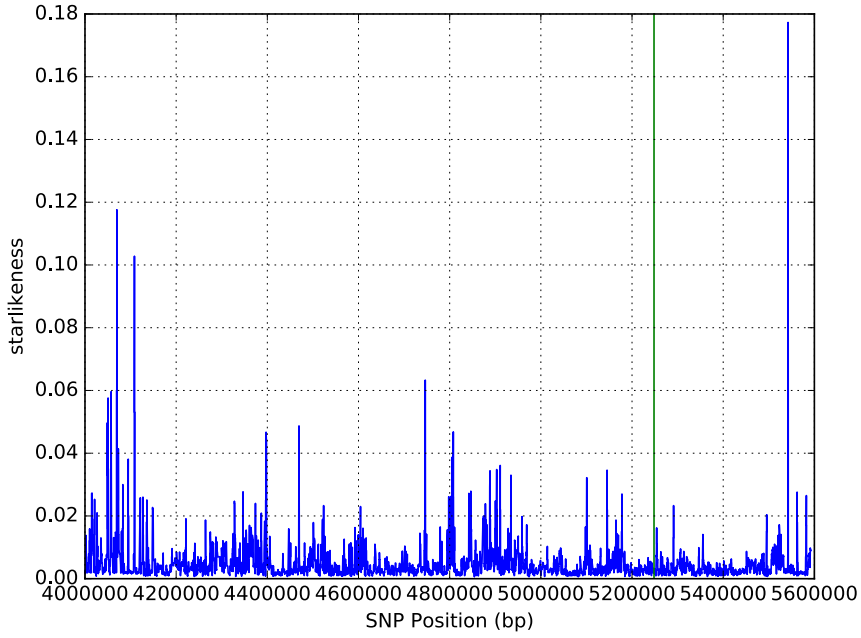


Figure 37.: Plot showing the starlikeness through the region with the SNP that has a direct influence on the HBB gene (green).

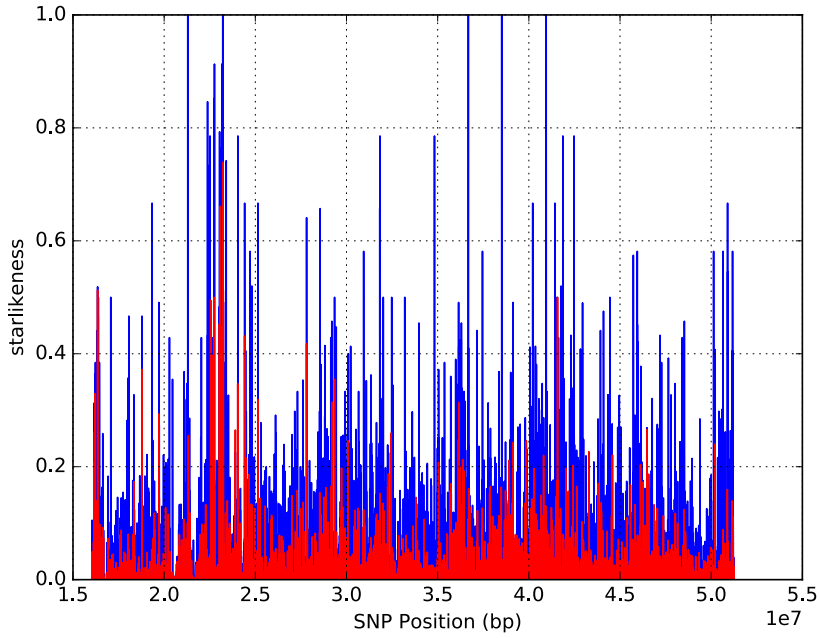


Figure 38.: Plot showing the starlikeness through chromosome 22 for 20 and 30 SNPs, blue and red, respectively.

Overall, the phylogenetic starlikeness was not a good measurement for selection evaluation. The starlikeness is highly variable and provides little evidence of evolutionary episodes. No specific signals was detected on chromosome 22 (Figure 38) and more importantly, no signal was detected around the two SNPs under selection (Figure 36 and figure 37).
The network is based on a limited set of SNPs (20 or 30) which provides little possibility for starlike figures to emerge.

5.8 PHYLOGEOGRAPHIC ANALYSIS OF A HAPLOBLOCK REGION

A fragment that displayed several windows with no cycles or a minimum amount of cycles within several windows with zero cycles was selected. The corresponding region of chromosome 22 is from 42988118bp to 42999883bp. This was one of various fragments detected, and it is actually one of the smallest ones, however the phylogenetic reconstruction of the larger ones was requiring a long time to properly present them in the context and time frame of this thesis.

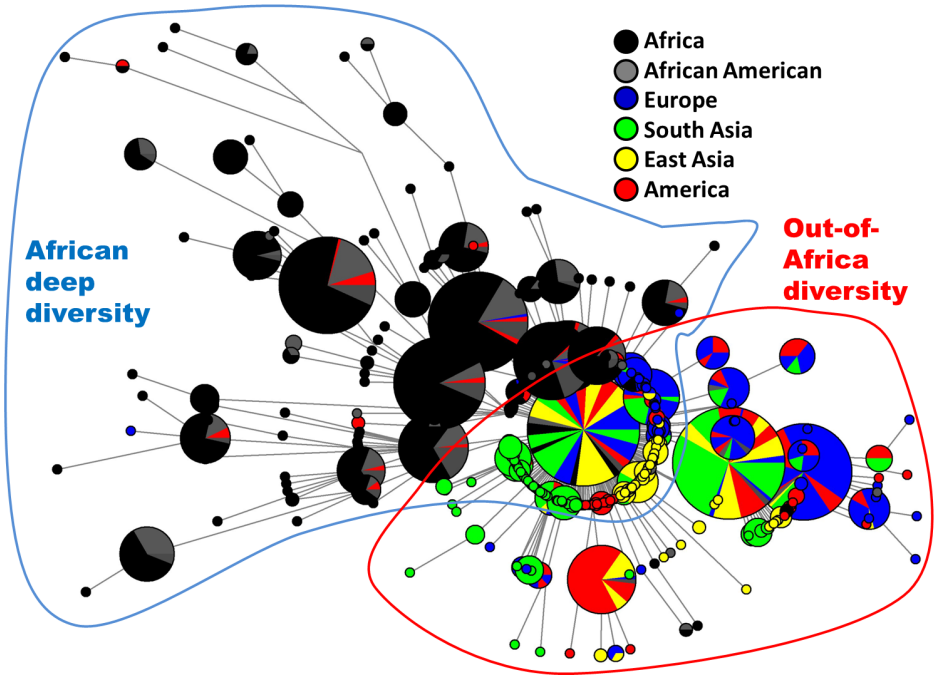


Figure 39.: Phylogenetic network outputted from the Network software GUI. Each circle corresponds to a node (haplotype).

Figure 39 shows the phylogenetic reconstruction of the fragment using the Network software. Even for this large fragment the amount of reticulation (cycles) is minimal. Basically the network displays a much deeper diversity in Africa. The Out-of-Africa diversity is centered in three main haplotypes (large circles with multiple colours) and their near derivatives. One of them is present also in Africa, which could represent the main Out-of-Africa founder for this segment of the genome, in a similar fashion as the mtDNA haplogroup L3 [23]. The two other haplotypes could have already emerged from that haplotype outside Africa.

6

DISCUSSION AND FUTURE WORK

In this thesis, an absolutely novel approach for the detection of haploblocks and recombination hotspots was created, implemented and tested. Such approach is a drastic change in relation to previous measures of linkage as a phylogenetic approach, based on network construction. Windows of variation across the genome corresponding to 20 or 30 SNPs were established and a network of the window was established. Recombination can create cycles in a phylogenetic reconstruction, which itself can be a direct measure of recombination rate in a region. A second measure based on the number of mutations was developed so that cycles with more mutations in the edges would have more weight (since they are more securely real recombination signals). Nevertheless, both measures provided very similar results.

The validation of the new methodology was performed by comparing the new genetic maps with widely used heatmaps based on traditional measures of LD (r^2 in this case). Broadly regions displaying high or low recombination with the new methodology also displayed high or low linkage in the corresponding heatmap. Although some exceptions were found, further investigation is required on the reasons for those events. Nevertheless, the objective is not to replicate the results from traditional LD as this methodology has very different properties and it would be expected that they behave differently in different situations. The simulations showed how promising the methodology might be, but overall it is required a higher level of complexity in the analysis in terms of variation in the region and even number of individuals. The approach also allowed to recognize the PAR1 within the X-chromosome region where a high rate of recombination takes place. The X-chromosome analysis also showed that, although we are dependent on the previous phasing of the chromosomes, this process is reliable and it does not affect the results in a significant way.

The network reconstruction on windows of diversity was also used to test a measure of how star-like the reconstruction was, which could be used to scan the genome for selection. Such approach did not prove successful which is not unexpected considering that each window was only based on 20 or 30 SNPs. However, in terms of detection of selection, the

general approach can show powerful results when considering the separation of the data and analysis in two alleles in a similar fashion as the extended haplotype homozygosity (EHH) test. In many ways such approach can be more powerful than a homozygosity test used in the detection of selective sweeps, since an early recombination close to the SNP of interest can disrupt by chance part of the homozygosity, while in this case it would account as a single event in the whole area. While some level of standardization might be required, the approach could be highly relevant in genomic studies.

The project started as an approach to detect haploblock whose phylogenetic and phylogeographic approach could provide powerful insights into the history of human populations. Results were showed for a relatively short fragment where the main aspects of human evolution were shown (modern human origins and Out-of-Africa migration), but detailed aspects can be extracted from several stretches of the genome. For future work, the combination with the chimpanzee homologous sequences will have the potential for dating, but a further complex Bayesian approach needs to be developed in order to properly distribute private mutations between the two pairs of homologous chromosomes in an individual, since that is a random process in the phasing algorithms.

This was a very ambitious Masters' project that proved successfully in launching a new approach that can be important in the future. It requires further refining and mostly extend the approach to the remaining genome, that can be a challenge computationally-wise. This first approach on the chromosome 22 and chromosome X generated data with over 1.5TB in size. Another further development should be the testing of the methodology in genomes with higher resolution, since the 1000genomes data is easily accessible but it mostly provides low-coverage sequencing. Access to another database, namely the Simons Diversity Project [45] in the future will be required. Nevertheless, it is always a positive point that the approach proved generally successful in a lower quality data anticipates further quality in a high quality dataset.

This project was also important in establishing a series of tools for manipulating Human Genomic data including a further development of the VDE, that was developed in the previous year. One of the novel features to be added is the capacity to build a heatmap based on LD measurements, since the pipeline developed here is substantially more efficient than the available tools in the literature.

BIBLIOGRAPHY

[1] L. D. Brooks, "Developing a Haplotype Map of the Human Genome for Finding Genes Related to Health and Disease," 2001. [Online]. Available: http://www.genome.gov/10001665

[2] A. Chakravarti, "Perspectives on human variation through the lens of diversity and race," *Cold Spring Harbor Perspectives in Biology*, vol. 7, no. 9, pp. 1–14, 2015.

[3] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin, "The variant call format and VCFtools," *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, 2011. [Online]. Available: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btr330

[4] T. Brown, "Chapter 1, The Human Genome," in *Genomes*, 2nd ed. Oxford: Wiley-Liss, 2002.

[5] P.-Y. Kwok, *Single Nucleotide Polymorphisms*, 2002, vol. 212.

[6] B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S. W. Scherer, S. Tavaré, P. Deloukas, M. E. Hurles, and E. T. Dermitzakis, "Relative impact of nucleotide and copy number variation on gene expression phenotypes," *Science*, vol. 315, no. 5813, pp. 848–853, 2007. [Online]. Available: http://science.sciencemag.org/content/315/5813/848

[7] J. Monson-Miller, D. C. Sanchez-Mendez, J. Fass, I. M. Henry, T. H. Tai, and L. Comai, "Reference genome-independent assessment of mutation density using restriction enzyme-phased sequencing." *BMC genomics*, vol. 13, no. 1, p. 72, 2012.

[8] A. Griffiths, J. Miller, and D. Suzuki, "Somatic versus germinal mutation," in *An Introduction to Genetic Analysis*, 7th ed. New York, NY, USA: W. H. Freeman, 2000.

[9] N. Li and M. Stephens, "Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data," *Genetics*, vol. 165, no. 4, pp. 2213–2233, 2003.

[10] J. D. Wall and J. K. Pritchard, "Haplotype blocks and linkage disequilibrium in the human genome," *Nature Reviews Genetics*, vol. 4, no. 8, pp. 587–597, 2003. [Online]. Available: http://www.nature.com/doifinder/10.1038/nrg1123

[11] P. C. Sabeti, D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E. S. Lander, "Detecting recent positive selection in the human genome from haplotype structure," *Nature*, vol. 419, no. 6909, pp. 832–837, oct 2002.

[12] Z. Gompert and C. A. Buerkle, "Analyses of genetic ancestry enable key insights for molecular ecology," *Molecular Ecology*, vol. 22, no. 21, pp. 5278–5294, 2013. [Online]. Available: http://doi.wiley.com/10.1111/mec.12488

[13] B. Y. J. M. Smith and J. Haigh, "The hitch-hiking effect of a favourable gene," pp. 23–35, 1974.

[14] M. Slatkin, "Linkage disequilibrium — understanding the evolutionary past and mapping the medical future," *Nature Reviews Genetics*, vol. 9, no. 6, pp. 477–485, 2008. [Online]. Available: http://www.nature.com/doifinder/10.1038/nrg2361

[15] R. C. Lewontin, "The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models." *Genetics*, vol. 49, no. 1, pp. 49–67, 1964.

[16] W. G. Hill and A. Robertson, "Linkage disequilibrium in finite populations," *Theoretical and Applied Genetics*, vol. 38, no. 6, pp. 226–231, 1968. [Online]. Available: http://dx.doi.org/10.1007/BF01245622

[17] S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotim, A. Adeyemo, R. Cooper, and R. Ward, "The structure of haplotype blocks in the human genome," *Science*, vol. 296, no. 2002, pp. 2225–2229, 2012.

[18] G. McVean, S. Myers, S. Hunt, P. Deloukas, D. Bentley, and P. Donnelly, "The Fine-Scale Structure of Recombination Rate Variation in the Human Genome," *Science*, vol. 304, no. 2004, pp. 581 – 4, 2004.

[19] A. H. Mangs and B. J. Morris, "The Human Pseudoautosomal Region (PAR): Origin, Function and Future," *Current Genomics*, vol. 8, pp. 129–136, 2007.

[20] A. Flaquer, G. A. Rappold, T. F. Wienker, and C. Fischer, "The human pseudoautosomal regions: a review for genetic epidemiologists," *Eur J Hum Genet*, vol. 16, no. 7, pp. 771–779, 2008.

[21] A. G. Hinch, N. Altemose, N. Noor, P. Donnelly, and S. R. Myers, "Recombination in the Human Pseudoautosomal Region PAR1," *PLoS Genetics*, vol. 10, no. 7, 2014.

[22] T. D. Petes, "Meiotic recombination hot spots and cold spots." *Nature reviews. Genetics*, vol. 2, no. 5, pp. 360–369, 2001.

[23] P. Soares, F. Alshamali, J. B. Pereira, V. Fernandes, N. M. Silva, C. Afonso, M. D. Costa, E. Musilová, V. MacAulay, M. B. Richards, V. Černý, and L. Pereira, "The expansion of mtDNA haplogroup L3 within and out of Africa," *Molecular Biology and Evolution*, vol. 29, no. 3, pp. 915–927, 2012.

[24] K. J. Kim, H.-J. Lee, M.-H. Park, S.-H. Cha, K.-S. Kim, H.-T. Kim, K. Kimm, B. Oh, and J.-Y. Lee, "SNP identification, linkage disequilibrium, and haplotype analysis for a 200-kb genomic region in a Korean population." *Genomics*, vol. 88, no. 5, pp. 535–40, 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16919420

[25] S. Shifman, "Linkage disequilibrium patterns of the human genome across populations," *Human Molecular Genetics*, vol. 12, no. 7, pp. 771–776, 2003. [Online]. Available: http://hmg.oxfordjournals.org/content/12/7/771.full

[26] M. C. Campbell and S. A. Tishkoff, "African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping." *Annual review of genomics and human genetics*, vol. 9, pp. 403–433, 2008.

[27] P. Triska, P. Soares, E. Patin, V. Fernandes, V. Cerny, and L. Pereira, "Extensive admixture and selective pressure across the sahel belt," *Genome biology and evolution*, vol. 7, no. 12, pp. 3484–3495, 2015.

[28] S. a. Tishkoff and S. M. Williams, "Genetic analysis of African populations: human evolution and complex disease." *Nature reviews. Genetics*, vol. 3, no. 8, pp. 611–621, 2002.

[29] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang, "Five Years of GWAS Discovery," *The American Journal of Human Genetics*, vol. 90, no. 1, pp. 7–24, 2012. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0002929711005337

[30] W. S. Bush and J. H. Moore, "Chapter 11: Genome-wide association studies." *PLoS computational biology*, vol. 8, no. 12, p. e1002822, 2012. [Online]. Available: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002822

[31] J. Hey, "What's so hot about recombination hotspots?" *PLoS Biology*, vol. 2, no. 6, pp. 730–733, 2004.

[32] R. D. Hernandez, J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton, G. McVean, G. Sella, and M. Przeworski, "Classic Selective Sweeps Were Rare in Recent Human Evolution," *Science*, vol. 331, no. 6019, pp. 920–924, 2011. [Online]. Available: http://science.sciencemag.org/content/331/6019/920

[33] J. M. Smith and J. Haigh, "The hitch-hiking effect of a favourable gene." *Genetical research*, vol. 89, no. 5-6, pp. 391–403, dec 2007.

[34] B. F. Voight, S. Kudaravalli, X. Wen, and J. K. Pritchard, "A map of recent positive selection in the human genome," *PLoS Biol*, vol. 4, no. 3, 03 2006. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pbio.0040072

[35] P. C. Sabeti, P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E. H. Byrne, S. A. McCarroll, R. Gaudet, S. F. Schaffner, and E. S. Lander, "Genome-wide detection and characterization of positive selection in human populations," *Nature*, vol. 449, no. 7164, pp. 913–918, oct 2007. [Online]. Available: http://dx.doi.org/10.1038/nature06250

[36] C. A. Mulcare, M. E. Weale, A. L. Jones, B. Connell, D. Zeitlyn, A. Tarekegn, D. M. Swallow, N. Bradman, and M. G. Thomas, "The T Allele of a Single-Nucleotide Polymorphism 13.9 kb Upstream of the Lactase Gene (LCT) (C−13.9kbT) Does Not Predict or Cause the Lactase-Persistence Phenotype in Africans," pp. 1102–1110, jun 2004.

[37] S. A. Tishkoff, F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, J. S. Silverman, K. Powell, H. M. Mortensen, J. B. Hirbo, M. Osman, M. Ibrahim, S. A. Omar, G. Lema, T. B. Nyambo, J. Ghori, S. Bumpstead, J. K. Pritchard, G. A. Wray, and P. Deloukas, "Convergent adaptation of human lactase persistence in Africa and Europe," *Nat Genet*, vol. 39, no. 1, pp. 31–40, jan 2007. [Online]. Available: http://dx.doi.org/10.1038/ng1946

[38] C. I. Fernández, N. Montalva, M. Arias, M. Hevia, M. L. Moraga, and S. V. Flores, "Lactase non-persistence and general patterns of dairy intake in indigenous and mestizo chilean populations," *American Journal of Human Biology*, vol. 28, no. 2, pp. 213–219, 2016. [Online]. Available: http://dx.doi.org/10.1002/ajhb.22775

[39] G. Band, Q. S. Le, L. Jostins, M. Pirinen, K. Kivinen, M. Jallow, F. Sisay-Joof, K. Bojang, M. Pinder, G. Sirugo, D. J. Conway, V. Nyirongo, D. Kachala, M. Molyneux, T. Taylor, C. Ndila, N. Peshu, K. Marsh, T. N. Williams, D. Alcock, R. Andrews, S. Edkins, E. Gray, C. Hubbart, A. Jeffreys, K. Rowlands, K. Schuldt, T. G. Clark, K. S. Small, Y. Y. Teo, D. P. Kwiatkowski, K. A. Rockett, J. C. Barrett, C. C. A. Spencer, and M. G. E. N. ¶, "Imputation-Based Meta-Analysis of Severe Malaria in Three African Populations," *PLoS Genet*, vol. 9, no. 5, pp. 1–13, 2013.

[40] M. G. E. Network, "Reappraisal of known malaria resistance loci in a large multicenter study," *Nat Genet*, vol. 46, no. 11, pp. 1197–1204, nov 2014. [Online]. Available: http://dx.doi.org/10.1038/ng.3107

[41] T. O. Apinjoh, J. K. Anchang-Kimbi, C. Njua-Yafi, A. N. Ngwai, R. N. Mugri, T. G. Clark, K. A. Rockett, D. P. Kwiatkowski, and E. A. Achidi, "Association of candidate gene polymorphisms and TGF-beta/IL-10 levels with malaria in three regions of Cameroon: a case-control study." *Malaria journal*, vol. 13, p. 236, 2014.

[42] V. D. Mangano, Y. Kabore, E. C. Bougouma, F. Verra, N. Sepulveda, C. Bisseye, F. Santolamazza, P. Avellino, A. B. Tiono, A. Diarra, I. Nebie, K. A. Rockett, S. B. Sirima, and D. Modiano, "Novel Insights Into the Protective Role of Hemoglobin S and C Against Plasmodium falciparum Parasitemia." *The Journal of infectious diseases*, vol. 212, no. 4, pp. 626–634, aug 2015.

[43] H. G. S. ConsortiumInternational, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 7011, pp. 931–945, oct 2004. [Online]. Available: http://dx.doi.org/10.1038/nature03001

[44] A.-S. Malaspinas, M. C. Westaway, C. Muller, V. C. Sousa, O. Lao, I. Alves, A. Bergström, G. Athanasiadis, J. Y. Cheng, J. E. Crawford, T. H. Heupink, E. Macholdt, S. Peischl, S. Rasmussen, S. Schiffels, S. Subramanian, J. L. Wright, A. Albrechtsen, C. Barbieri, I. Dupanloup, A. Eriksson, A. Margaryan, I. Moltke, I. Pugach, T. S. Korneliussen, I. P. Levkivskyi, J. V. Moreno-Mayar, S. Ni, F. Racimo, M. Sikora, Y. Xue, F. A. Aghakhanian, N. Brucato, S. Brunak, P. F. Campos, W. Clark, S. Ellingvåg, G. Fourmile, P. Gerbault, D. Injie, G. Koki, M. Leavesley, B. Logan, A. Lynch, E. A. Matisoo-Smith, P. J. McAllister, A. J. Mentzer, M. Metspalu, A. B. Migliano, L. Murgha, M. E. Phipps, W. Pomat, D. Reynolds, F.-X. Ricaut, P. Siba, M. G. Thomas, T. Wales, C. M. Wall, S. J. Oppenheimer, C. Tyler-Smith, R. Durbin, J. Dortch, A. Manica, M. H. Schierup, R. A. Foley, M. M. Lahr, C. Bowern, J. D. Wall, T. Mailund, M. Stoneking, R. Nielsen, M. S. Sandhu, L. Excoffier, D. M. Lambert, and E. Willerslev, "A genomic history of Aboriginal Australia," *Nature*, vol. 538, no. 7624, pp. 207–214, oct 2016. [Online]. Available: http://dx.doi.org/10.1038/nature18299

[45] S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J. P. Spence, Y. S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I. G. Romero, A. R. Jha, D. M. Behar, C. M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O. L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M. S. Abdullah, A. Ruiz-Linares, C. M. Beall,

A. Di Rienzo, C. Jeong, E. B. Starikovskaya, E. Metspalu, J. Parik, R. Villems, B. M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J. T. S. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M. F. Hammer, T. Kivisild, W. Klitz, C. A. Winkler, D. Labuda, M. Bamshad, L. B. Jorde, S. A. Tishkoff, W. S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Pääbo, J. Kelso, N. Patterson, and D. Reich, "The Simons Genome Diversity Project: 300 genomes from 142 diverse populations," *Nature*, vol. 538, no. 7624, pp. 201–206, oct 2016. [Online]. Available: http://dx.doi.org/10.1038/nature18964

[46] L. Pagani, D. J. Lawson, E. Jagoda, A. Mörseburg, A. Eriksson, M. Mitt, F. Clemente, G. Hudjashov, M. DeGiorgio, L. Saag, J. D. Wall, A. Cardona, R. Mägi, M. A. W. Sayres, S. Kaewert, C. Inchley, C. L. Scheib, M. Järve, M. Karmin, G. S. Jacobs, T. Antao, F. M. Iliescu, A. Kushniarevich, Q. Ayub, C. Tyler-Smith, Y. Xue, B. Yunusbayev, K. Tambets, C. B. Mallick, L. Saag, E. Pocheshkhova, G. Andriadze, C. Muller, M. C. Westaway, D. M. Lambert, G. Zoraqi, S. Turdikulova, D. Dalimova, Z. Sabitov, G. N. N. Sultana, J. Lachance, S. Tishkoff, K. Momynaliev, J. Isakova, L. D. Damba, M. Gubina, P. Nymadawa, I. Evseeva, L. Atramentova, O. Utevska, F.-X. Ricaut, N. Brucato, H. Sudoyo, T. Letellier, M. P. Cox, N. A. Barashkov, V. Škaro, L. Mulahasanovic´, D. Primorac, H. Sahakyan, M. Mormina, C. A. Eichstaedt, D. V. Lichman, S. Abdullah, G. Chaubey, J. T. S. Wee, E. Mihailov, A. Karunas, S. Litvinov, R. Khusainova, N. Ekomasova, V. Akhmetova, I. Khidiyatova, D. Marjanović, L. Yepiskoposyan, D. M. Behar, E. Balanovska, A. Metspalu, M. Derenko, B. Malyarchuk, M. Voevoda, S. A. Fedorova, L. P. Osipova, M. M. Lahr, P. Gerbault, M. Leavesley, A. B. Migliano, M. Petraglia, O. Balanovsky, E. K. Khusnutdinova, E. Metspalu, M. G. Thomas, A. Manica, R. Nielsen, R. Villems, E. Willerslev, T. Kivisild, and M. Metspalu, "Genomic analyses inform on migration events during the peopling of Eurasia," *Nature*, vol. 538, no. 7624, pp. 238–242, oct 2016. [Online]. Available: http://dx.doi.org/10.1038/nature19792

[47] R. Bao, L. Huang, J. Andrade, W. Tan, W. a. Kibbe, H. Jiang, and G. Feng, "Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing," *Libertas Academica*, vol. 13, pp. 67–82, 2014.

[48] S. El-Metwally, T. Hamza, M. Zakaria, and M. Helmy, "Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges," *PLoS Computational Biology*, vol. 9, no. 12, 2013.

[49] P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. Hsi-Yang Fritz, M. K. Konkel, A. Malhotra, A. M. Stütz, X. Shi, F. Paolo Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. K. Lam,

X. Jasmine Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E.-W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, and J. O. Korbel, "An integrated map of structural variation in 2,504 human genomes," *Nature*, vol. 526, no. 7571, pp. 75–81, 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26432246

[50] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[51] O. Delaneau and J. Marchini, "Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel." *Nature communications*, vol. 5, p. 3934, jun 2014.

[52] O. Delaneau, J. Marchini, and J.-F. Zagury, "A linear complexity phasing method for thousands of genomes," *Nat Meth*, vol. 9, no. 2, pp. 179–181, feb 2012. [Online]. Available: http://dx.doi.org/10.1038/nmeth.1785

[53] The 1000 Genomes Project Consortium, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, 2010. [Online]. Available: http://www.nature.com/doifinder/10.1038/nature09534

[54] J. R. Miller, S. Koren, and G. Sutton, "Assembly algorithm for Next-Ganeration Sequencing data," *Genomics*, vol. 95, no. 6, pp. 315–327, 2010.

[55] "A Deep Catalog of Human Genetic Variation." [Online]. Available: http://www.1000genomes.org/

[56] J.-l. Gailly and M. Adler, "The gzip home page," 2003. [Online]. Available: http://www.gzip.org/

[57] Exome Aggregation Consortiuml, "Analysis of protein-coding genetic variation in 60,706 humans," *bioRxiv*, oct 2015. [Online]. Available: http://biorxiv.org/content/early/2015/10/30/030338.abstract

[58] "NHLBI Exome Sequencing Project (ESP) Exome Variant Server." [Online]. Available: http://evs.gs.washington.edu/EVS/

[59] W. Allcock, J. Bresnahan, R. Kettimuthu, M. Link, C. Dumitrescu, I. Raicu, and I. Foster, "The globus striped gridftp framework and server," in *Proceedings*

of the 2005 ACM/IEEE Conference on Supercomputing, ser. SC '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 54–. [Online]. Available: http://dx.doi.org/10.1109/SC.2005.72

[60] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham, "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses," *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0002929707613524

[61] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly, "Haploview: Analysis and visualization of LD and haplotype maps," *Bioinformatics*, vol. 21, no. 2, pp. 263–265, 2005.

[62] H. J. Bandelt, P. Forster, B. C. Sykes, and M. B. Richards, "Mitochondrial portraits of human populations using median networks." *Genetics*, vol. 141, no. 2, pp. 743–753, 1995. [Online]. Available: http://www.genetics.org/content/141/2/743.short

[63] D. H. Huson, R. Rupp, and C. Scornavacca, *Phylogenetic Networks: Concepts, Algorithms and Applications*. New York, NY, USA: Cambridge University Press, 2011, vol. 1.

[64] S. Purcell, "PLINK(v1.06)." [Online]. Available: http://pngu.mgh.harvard.edu/purcell/plink/

[65] J. Casbon, "PyVCF - A Variant Call Format Parser for Python," 2012. [Online]. Available: https://pyvcf.readthedocs.org/

[66] M. Alves, J. Alves, R. Camacho, P. Soares, and L. Pereira, "From networks to trees," *Advances in Intelligent and Soft Computing*, vol. 154 AISC, pp. 129–136, 2012.

[67] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numer. Math.*, vol. 1, no. 1, pp. 269–271, Dec. 1959. [Online]. Available: http://dx.doi.org/10.1007/BF01386390

[68] N. A. Rosenberg and M. Nordborg, "Genealogical trees, coalescent theory and the analysis of genetic polymorphisms." *Nature reviews. Genetics*, vol. 3, no. 5, pp. 380–390, may 2002.

[69] N. Rosenberg and M. Feldman, "The relationship between coalescence times and population divergence times," *Devolopments in Theorical Populations Genetics*, p. 35, 2002.

[70] J. Saillard, P. Forster, N. Lynnerup, H. J. Bandelt, and S. Nørby, "mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion." *American journal of human genetics*, vol. 67, no. 3, pp. 718–26, 2000.

[71] J.-M. Chen, D. N. Cooper, N. Chuzhanova, C. Ferec, and G. P. Patrinos, "Gene conversion: mechanisms, evolution and human disease," *Nat Rev Genet*, vol. 8, no. 10, pp. 762–775, oct 2007. [Online]. Available: http://dx.doi.org/10.1038/nrg2193

[72] R. D. Hernandez, "A flexible forward simulator for populations subject to selection and demography." *Bioinformatics (Oxford, England)*, vol. 24, no. 23, pp. 2786–2787, dec 2008.

[73] S. Forrest, "Genetic algorithms: principles of natural selection applied to computation," *Science*, vol. 261, no. 5123, pp. 872–878, 1993. [Online]. Available: http://science.sciencemag.org/content/261/5123/872

[74] E. Dawson, G. Abecasis, S. Bumpstead, Y. Chen, and S. Hunt, "A first-generation linkage disequilibrium map of human chromosome 22," *Nature*, vol. 418, no. August, pp. 544–548, 2002.

[75] H. Shi, H. Zhong, Y. Peng, Y.-L. Dong, X.-B. Qi, F. Zhang, L.-F. Liu, S.-J. Tan, R. Z. Ma, C.-J. Xiao, R. S. Wells, L. Jin, and B. Su, "Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations," *BMC Biology*, vol. 6, p. 45, oct 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2605740/

[76] R. Stanyon, M. Sazzini, and D. Luiselli, "Timing the first human migration into eastern asia," *Journal of Biology*, vol. 8, no. 2, p. 18, 2009. [Online]. Available: http://dx.doi.org/10.1186/jbiol115

[77] P. Mellars, K. C. Gori, M. Carr, P. A. Soares, and M. B. Richards, "Genetic and archaeological perspectives on the initial modern human colonization of southern asia," *Proceedings of the National Academy of Sciences*, vol. 110, no. 26, pp. 10 699–10 704, 2013. [Online]. Available: http://www.pnas.org/content/110/26/10699.abstract

A

SUPPORT MATERIAL

Throughout this thesis several files were created. All the rdf, out, info, VCF and many other files are in an external disk, due to being a large amount of data that requires a lot of disk space. Scripts, figures and tables are in the `dissertation_pg27668` folder. The scripts folder contains all scripts created, complementary files, subsets of VCF files for testing and a folder with text files that contain the cycles SNP ID of each population from chromosome 22. The `figures_and_tables` folder contains several folders with the output of the Cycles/Region measure and the starlikeness measure, namely the tab-delimited files and the respective plots.

B

PYTHON CODE

The remaining methods created throughout this work are not listed, but were equally important for the parsing and transformation of the files. Scripts containing those functions are in the scripts folder.
The forthcoming 5 methods are some relevant methods created. The first method (`cycle_numpy`) will retrieve all cycles and its' SNP ID (without duplicates) for each population in study.

```
1
# G its a unweighted graph object from network X
# B its a numpy matrix, which only considers connected nodes
def cycle_numpy(G, B):
    list_cycle = []
    nodos = G.nodes() # list with nodes
    B_1 = B[1]
    B_0 = B[0]
    for i in range(len(B_0)):
        v1 = B_0[i] # node 1
        v2 = B_1[i] # node 2
        C = np.where(B_0 == v2)[0] # neighbors of node 2
        mut = G[nodos[v1]][nodos[v2]]['mutation'][0] # SNP ID of edge v1 - v2
        for j in C: # course through all neighbors of node 2
            v3 = B_1[j] # node 3
            if v3 != v1: # node 3 cannot be node 1
                mut2 = G[nodos[v2]][nodos[v3]]['mutation'][0]
                # mut2 is the SNP ID of edge v2 - v3
                D = np.where(B_0 == v3)[0] # neighbors of node 3
                for k in D: # course through all neighbors of node 3
                    v4 = B_1[k] # node 4
                    mut3 = G[nodos[v3]][nodos[v4]]['mutation'][0]
                    # mut3 is the SNP ID of edge v3 - v4
                    E = np.where(B_0 == v4)[0] # neighbors of node 4
                    for l in E: # course through all neighbors of node 3
                        v5 = B_1[l] # node 5
                        mut4 = G[nodos[v4]][nodos[v5]]['mutation'][0]
                        # mut4 is the SNP ID of edge v4 - v5
                        # Check if the first edge (SNP ID) is equal to the third;
```

```
                        # and the second to the fourth
                        # Also node 5 must be equal to the first node (node 1)
                        #  to complete the cycle
                        if mut == mut3 and mut2 == mut4 and v1 == v5:
                            # do not consider duplicates
                            # or passing through the same cycle again
                            if (mut, mut2) not in list_cycle:
                                if (mut2, mut) not in list_cycle:
                                    list_cycle.append((mut, mut2))
    return np.array(list_cycle)
```

Listing B.1: Python function `cycle_numpy`

The second method (`create_tree_branch`) will be needed for the fifth method (`build_starlikeness_path`), that creates a "path" of the data that is needed to calculate the starlikeness.

```

# G its a weighted graph object from network X
# From the parsed out file a dictionary (dic_taxa)
# is retrieved with the node name (key)
# and the corresponding samples (list)
# nodos is the list with nodes
# center_edges is a numpy matrix with the neighbors of the center node
# This method creates  dictionary with nodes as keys and a corresponding
# list of tuples with the number of samples and the mutation number
# between center node and its neighbors
def create_tree_branch(B, G, center_edges, dic_taxa, nodos):
    phylo_dic = {}
    arvore = []
    B_1 = B[1]
    B_0 = B[0]
    for i in center_edges[0]:
        phylo_dic[nodos[B_1[i]]] = []
        mut_num = G[nodos[B_0[i]]][nodos[B_1[i]]]['mutation']
        try:
            num_samp = len(dic_taxa[nodos[B_1[i]]])
            if num_samp > 1:
                num_samp += 1
        except KeyError:
            num_samp = 0
        phylo_dic[nodos[B_1[i]]].append((nodos[B_1[i]], len(mut_num), num_samp))
        arvore.append(nodos[B_1[i]])
        # arvore will be a list of neighbors of the center node
    return arvore, phylo_dic
```

Listing B.2: Python function to retrieve the information of only the neighbors of the center node (root)

Method 3 (`parse_out_file`) does the parsing of the out file from the Network Software, retrieving all the information from the graph created by this software.

```
# This method parses the out file from the Network Sofware
# dic is a dictionary with each line as key and a list
# with the information from the respective line
def parse_out_file(name, path):
    dic = {}
    num = 0
    map_char = []
    dic_taxa = {}
    list_taxon = []
    with open(path + '/' + str(name), "r") as out:
        lines = out.readlines()
        for j in lines:
            if "Link" in j:
                num += 1
                link = j.split(' ')
                dic[num] = []
                for k in link:
                    if len(k) >= 1 and k != '\n':
                        dic[num].append(k.split(',')[0])
            elif 'Mapping of character' in j:
                mapping = j.splitlines()
                mapping = mapping[0].split(' ')
                mapping = [x for x in mapping if x != '']
                map_char.append(mapping[3])
                # map_char contains the mutation position
                # when the window has 20 SNPs but cycles that form a cube
                # the mutation position has another number higher than 20
                # ex: 23;5
                # the 23rd mutation position corresponds to the real 5th position
            elif 'Tax.' in j:
                taxa = [x for x in j.split(' ') if x != '']
                try:
                    dic_taxa[taxa[1]].append(taxa[4].split('\n')[0])
                except KeyError:
                    dic_taxa[taxa[1]] = [taxa[4].split('\n')[0]]
            elif 'Mapping of taxon' in j and 'mv' not in j:
                taxon = [x for x in j.split(' ') if x != '']
                list_taxon.append(taxon[3].split(';')[0])
        for tax in list_taxon:
```

```
            for kt in dic_taxa.keys():
42              if tax in dic_taxa[kt] or tax == kt:
                    bol = True
44                  break
                else:
46                  bol = False
            if not bol:
48              dic_taxa[tax] = [tax]
    return dic, dic_taxa, map_char
```

Listing B.3: Python function to parse the out file

`build_weighted_graph` method builds a weighted graph for the calculation of the Dijkstra path.

```
1
# This method builds a weighted graph from the output
3 # of parse_out_file
def build_weighted_graph(dic, dic_taxa, map_char):
5     num_mut = 0
    G = nx.Graph()
7     for key in dic.keys():
        ind_at = int(dic[key].index('at'))
9       mutations = len(dic[key][ind_at + 1:])
        num_mut += mutations
11      if dic[key][2] == 'a' or dic[key][2] in [str(x) for x in range(10)]:
            if dic[key][4] == 'mv': # mv = median vector
13              G.add_node(dic[key][4] + dic[key][5])
                node1 = dic[key][4] + dic[key][5]
15              if dic[key][8] == 'mv':
                    G.add_node(dic[key][8] + dic[key][9])
17                  node2 = dic[key][8] + dic[key][9]
                else:
19                  G.add_node(dic[key][8])
                    node2 = dic[key][8]
21          else:
                G.add_node(dic[key][4])
23              node1 = dic[key][4]
                if dic[key][7] == 'mv':
25                  G.add_node(dic[key][7] + dic[key][8])
                    node2 = dic[key][7] + dic[key][8]
27              else:
                    G.add_node(dic[key][7])
29                  node2 = dic[key][7]
            muts = []
31          for value in dic[key][ind_at + 1:]:  # ['6', '7'] or ['3']
                if int(value) > 20:
```

```
                    vals = map_char[int(value) - 1].split(';')  # ['23', '16']
    23 -> 16
                    muts.append(str(vals[1]))
                else:
                    muts.append(value)
            try:
                # number of samples of the haplotype
                s1 = len(dic_taxa[node1])
            except KeyError:
                s1 = 0
            try:
                s2 = len(dic_taxa[node2])
            except KeyError:
                s2 = 0
            if s1 == 0 or s2 == 0:
                # Nodes that do not have samples (median vectors)
                # will have a higher edge weight
                G.add_edge(node1, node2, mutation=muts, weight=2)
            else:
                G.add_edge(node1, node2, mutation=muts, weight=1)
    return G
```

Listing B.4: Python function to build a weighted graph object using NetworkX

Method to build the path were the coursing will have to be done to calculate the starlikeness of the graph.

```
# G its a weighted graph object from network X
# no is the center node of the network
# B is a matrix with only connected nodes after the
# calculation of the Dijkstra path
# It accepts as input the phylo_dic from create_tree_branch
# This method will append for each node his neighbors
# (only the neighbors that are further from the center node)
# and successive neighbors with the number of samples
# and mutation position of the haplotype
def build_starlikeness_path(B, G, arvore, dic_taxa, no, nodos, phylo_dic, k):
    dic_neig = {}
    B_1 = B[1]
    B_0 = B[0]
    for i in range(len(arvore)):
        new_l = [arvore[i]]
        while len(new_l) > 0:
            x = new_l.pop(0)
            elem = nodos.index(x)
            neig = np.where(B_0 == elem)
            for n in neig[0]:
```

```
                if nodos[B_1[n]] != no:
                    new_l.append(nodos[B_1[n]])
                    mut_num = G[nodos[B_0[n]]][nodos[B_1[n]]]['mutation']
                    try:
                        num_samp = len(dic_taxa[nodos[B_1[n]]])
                        if num_samp > 1:
                            num_samp += 1
                    except KeyError:
                        num_samp = 0
                    phylo_dic[arvore[i]].append((nodos[B_1[n]], len(mut_num),
num_samp))
                    last_l = [nodos[B_1[n]]]
                    dic_neig[(nodos[B_1[n]], len(mut_num), num_samp)] = []
                    key_n = (nodos[B_1[n]], len(mut_num), num_samp)
                    while len(last_l) > 0:
                        y = last_l.pop(0)
                        elem_y = nodos.index(y)
                        neig_y = np.where(B_0 == elem_y)
                        for p in neig_y[0]:
                            if nodos[B_1[p]] != phylo_dic[arvore[i]][0][0] and
nodos[B_1[p]] != k \
                                    and nodos[B_1[p]] != phylo_dic[arvore[i
]][-1][0] \
                                    and nodos[B_1[p]] not in [x[0] for x in
dic_neig[key_n]]:
                                last_l.append(nodos[B_1[p]])
                                mut_num = G[nodos[B_0[p]]][nodos[B_1[p]]]['
mutation']
                                try:
                                    num_samp = len(dic_taxa[nodos[B_1[p]]])
                                    if num_samp > 1:
                                        num_samp += 1
                                except KeyError:
                                    num_samp = 0
                                dic_neig[key_n].append((nodos[B_1[p]], len(
mut_num), num_samp))
                                phylo_dic[arvore[i]].append((nodos[B_1[p]], len(
mut_num), num_samp))
                                k = y
                            else:
                                try:
                                    new_l.pop(0)
                                except IndexError:
                                    pass
    return dic_neig
```

Listing B.5: Python function to find the path where the starlikeness calculation will course