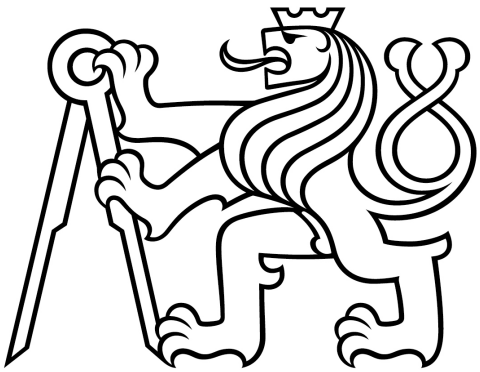
Czech Technical University
Faculty of Electrical Engineering
Department of Cybernetics



Bachelor's thesis

Detection of Selection Pressure on Human Endogenous Retroviruses

Tuan Anh Ho

Supervisor: Mgr. Pačes Jan, Ph.D

Study Programme: Open Informatics
Field of Study: Computer and Information Science
April 2020

BACHELOR'S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Ho Tuan Anh** Personal ID number: **474375**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Cybernetics**

Study program: **Open Informatics**

Branch of study: **Computer and Information Science**

II. Bachelor's thesis details

Bachelor's thesis title in English:

Detection of Selection Pressure on Human Endogenous Retroviruses

Bachelor's thesis title in Czech:

Detekce selekčních tlaků na lidské endogenní retroviry

Guidelines:

Study the recommended literature, familiarize yourself and do sufficient research on the topic. Using the data from 1000 Genomes project (https://www.internationalgenome.org/) and Human Endogenous Retrovirus Database (https://herv.img.cas.cz/) estimate the selection pressures based on site frequency spectrum on human endogenous retroviruses and their vicinity using these methods:
1. Tajima's D (Tajima 1989)
2. Fu and Li's D (Fu and Li 1993)
3. Fu and Li's F (Fu and Li 1993)
4. Fay and Wu's H (Fay and Wu 2000)
5. Zeng et al.'s E (Zeng et al. 2006)
Compare the estimates of the selected loci with the available data from the PopHuman database (https://pophuman.uab.cat/)

Bibliography / sources:

[1] Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics, 123 (1989), pp. 585-595
[2] Fu Y.-X., Li W.-H. Statistical tests of neutrality of mutations. Genetics, 133 (1993), pp. 693-709
[3] Fay J.C., Wu C.-I. Hitchhiking under positive Darwinian selection. Genetics, 155 (2000), pp. 1405-1413
[4] Zeng K., Fu Y.-X., Shi S., Wu C.-I. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics, 174 (2006), pp. 1431-1439

Name and workplace of bachelor's thesis supervisor:

Mgr. Jan Pačes, Ph.D., Ústav molekulární genetiky, AV ČR, Praha

Name and workplace of second bachelor's thesis supervisor or consultant:

doc. Ing. Jiří Kléma, Ph.D., Intelligent Data Analysis, FEE

Date of bachelor's thesis assignment: **06.01.2020** Deadline for bachelor thesis submission: **22.05.2020**

Assignment valid until: **30.09.2021**

Mgr. Jan Pačes, Ph.D.	doc. Ing. Tomáš Svoboda, Ph.D.	prof. Mgr. Petr Páta, Ph.D.
Supervisor's signature	Head of department's signature	Dean's signature

III. Assignment receipt

The student acknowledges that the bachelor's thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the bachelor's thesis, the author must state the names of consultants and include a list of references.

. _____
Date of assignment receipt

_____
Student's signature

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Mgr. Pačes Jan, Ph.D and his colleague RNDr. Edvard Ehler, Ph.D both from the Institute of Molecular Genetics, Czech Academy of Sciences for their assistance and guidance with this thesis. I also want to acknowledge my family and friends, namely Ta Van Duy, Škuthan Jiří, Šatra Šimon, Nguyen Manh Hai, Janoušková Klára, Nguyen Diem Huong and others for their constant support during my studies. Finally, I did this work to honor my deceased father in hopes he would be proud of his son today.

This work was supported by the Institute of Molecular Genetics, Czech Academy of Sciences and ELIXIR CZ research infrastructure project (MEYS Grant No: LM2018131) including access to computing and storage facilities.

Author's statement

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

Prague, date ....

Abstract

The human genome consists of coding and non-coding DNA. The coding DNA comprises about two percent of the genome. It encodes proteins, molecules of functional RNA and regulatory elements. A substantial part of the non-coding DNA is also transcribed to a various extent to RNA molecules of unknown or only partially known function. Large portions of the non-coding DNA consists of repetitive nucleotide sequences. About eight percent of the human genome are endogenous retroelements that are remnants of retroviruses that infected humans in ancient times. Abundance of these loci raises a question about their possible functionality influencing fitness of the host. In this work, we approached this question by estimating selection pressure on these DNA sequences. We used data from the "1000 human genomes project" and applied the methods based on site frequency spectrum. The results are publicly available for the scientific community at the HERVd online database, part of ELIXIR, the European life-sciences Infrastructure.

Keywords DNA, genome, homo sapiens, human endogenous retrovirus, selection pressure

Abstrakt

Lidský genom se skládá z kódující a nekódující DNA. Kódující DNA obsahuje asi dvě procenta genomu. Kóduje proteiny, molekuly funkční RNA a regulační prvky. Podstatná část nekódující DNA se také v různé míře přepisuje na molekuly RNA neznámé nebo pouze částečně známé funkce. Velká část nekódující DNA sestává z repetitivních nukleotidových sekvencí. Asi osm procent lidského genomu je tvořeno endogenními retroelementy, což jsou zbytky retrovirů, které infikovaly lidi ve dávné minulosti. Počet těchto lokusů vyvolává otázku o jejich možném vlivu na fitness hostitele. V této práci řešíme tuto otázku odhadem selekčního tlaku na tyto sekvence. Použili jsme data z projektu "1000 human genome project" a aplikovali metody založené na frekvenčním spektru mutací. Výsledky jsou pro vědeckou komunitu veřejně přístupné v online databázi HERVd, která je součástí ELIXIR, evropská infrastruktura pro biologické vědy.

Klíčová slova DNA, genom, homo sapiens, lidský endogenní retrovirus, selekční tlak

Contents

List of Figures

List of Tables

Acronyms

DNA Deoxyribonucleic acid. 1, 3–6, 11, 15–19, 23, 26, 27

HERVs Human endogenous retroviruses. 1, 25, 27, 29, 31, 32

LD Linkage disequilibrium. 9, 10, 27

nt nucleotides. 3, 5, 27

RNA Ribonucleic acid. 3, 6

SNP Single nucleotide polymorphism. 6, 10, 15

VCF Variant Call Format. 23, 24, 26

Chapter 1

Introduction

1.1 Motivation

The genomes of all living species are under constant pressure from several major evolutionary forces, namely the effects of genetic drift and small population, molecular drive, migration, and selection; the human genome is not different, given the extra layers of cultural adaptations that facilitate their survival. It is the selection that is thought to play a major role in what makes us human and also a highly competitive biological species [1]. As selection occurs, the variation of the human genome in a population fluctuates. A high or low abundance of parts of the genome could indicate the fact that these parts carry significant biological functions for their hosts. Therefore, estimating selection pressures as such could partially help us in finding the answers to whether certain loci carry biological functions.

The mammalian and other vertebrates genomes are interwoven with remnants of ancient retroviruses that colonized their ancestor's germ-lines [2]. These remnants in humans, namely the Human endogenous retroviruses (HERVs), account for about 8% of the human genome. HERVs are relics of old infections that, over the last 100 million years, infected not only the body but the germ-line as well and have become stable components at the crossroads of self and foreign DNA. The HERVs co-evolution with the host has intriguingly brought about the domestication of previously committed retroviral behaviors, providing new cellular functions. As an example, selected HERVs proteins have been co-opted for pregnancy-related reasons [3]. Recently, expressions of HERVs were often correlated with a wide variety of pathologies, including several cancers and autoimmune diseases [4]. Nevertheless, many functions of HERVs may still be unbeknown to us, making it loci of interest.

1.2 Objectives

This work aims to estimate selection pressures on human endogenous retroviruses and their vicinity using site frequency spectrum tests. To accomplish that, we first analyze and afterward implement the methods for estimating these pressures. Having that done, we run our tests on the whole human population data, thus getting our estimates. Finally, we check the correctness of our results by comparing them with online databases - in our case, we compare them with the PopHuman database [5].

1.3 Thesis structure

In chapter 2 we provide an overview of the fundamental biological knowledge necessary for the understanding of our topic. In chapter 3 we analyze the methods used for the calculation of the selection pressures. Chapter 4 gives a preview of other attempts of the calculation. Chapter 5 describes our way of solving this problem, using the related work for showing the correctness among other things. Finally, chapter 6 gives a summary of our work and mentions possible future ideas for improvement.

Chapter 2

Biological background

Throughout this chapter, we address the underlying biological knowledge required to understand the subject of this work. Firstly, we step back to present the big picture, offering a rather brief summary of the human genome with two subsections dedicated to genetic inheritance and mutations. Afterward, we move on to what retroviruses are, how they function, and go through the process of endogenization. Finally, we conclude by giving an overview of evolutionary genetics necessary for our topic.

2.1 The human genome

All living organisms possess a genome containing biological information required to construct and sustain a living being of that particular organism. Organimal genomes are composed of DNA – usually in the forms of double-stranded helices. However, viruses, even though they are not considered as living organisms, can contain Ribonucleic acid (RNA) genomes, either in the form of single-stranded or double-stranded helices [6]. Human beings are diploid like nearly any animal – that is, having two copies of the genome in each of our somatic cells, the cells that form tissues. The DNA is stored in the cell as a chunk called a chromosome. Chromosomes are made from two chromatids – copies of the same DNA, which are needed for the cell division. The human haploid genome (i.e., a single copy) consists of approximately 3.2 billion nucleotides (nt), the fundamental genome building blocks within which information is stored [7]. There exist exactly four DNA nt – adenine, cytosine, guanine and thymine abbreviated as A, C, G, T, named after their acronyms [8].

The flow of processing the information encoded in DNA is summarized in the central dogma of molecular biology, shown in figure 2.a. During the mechanisms that process DNA, mistakes can happen, even though very unlikely, ultimately resolving into a greater genetic variety. A variant of a gene is called an allele. As can be seen from figure 2.a, DNA holds instructions for the synthesis of proteins of our cells. These instructions are incorporated within genes. Through evolution, variants of genes contribute to the vast range of commonly detectable human differences. Despite the significance of proteins, protein-coding genes account for only 2% of the genome. Some parts of the rest of the genome are essential for the development of RNA molecules with various purposes, for example, the regulation of genes or the function of chromosomes; yet most of the genome does not have a well described role [7].

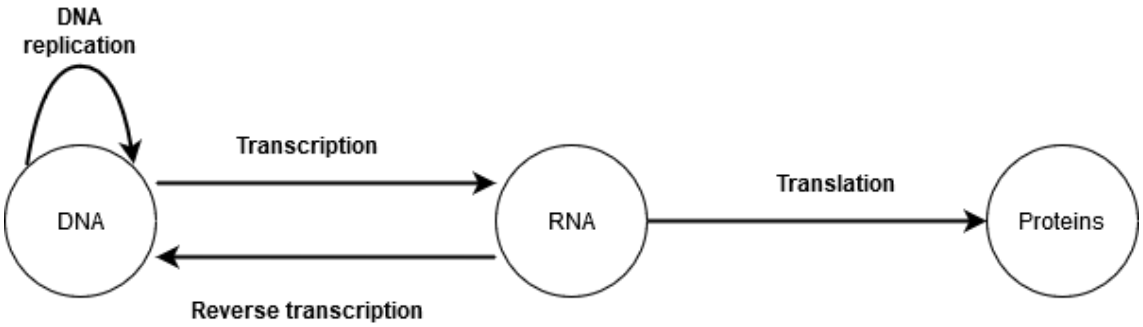


Figure 2.a: The central dogma of molecular biology. [8] RNA sequences can be formed from DNA sequences through a process called transcription. Afterwards in translation, these RNA sequences produce proteins. DNA is maintained through DNA replication. In special cases, RNA can be reversely transcribed into DNA.

2.1.1 Genetic inheritance – mitosis, meiosis

In the course of cell division, genetic information is being passed to the daughter cells. There are two different kinds of cell divisions, specifically mitosis, and meiosis. In somatic cells, the same genetic material in each daughter cell as in the parental cell is being incorporated during mitosis. Nevertheless, in order to produce a gamete (egg or sperm), which comprises only half the diploid complement of genetic material (i.e., is haploid), the advanced division mechanism of meiosis is necessary. Meiosis is, therefore, a vital method of passing the genome from one generation to another [7]. Both the mitosis and meiosis are divided into phases, based on the action happening. Figure 2.b summarizes this process. The fusion of an egg and sperm cell is called a zygote. In diploid cells, some loci of the chromosome copies may differ. When a locus differs, the organism is referred to as heterozygous at that locus and homozygous when it does not differ. Alleles may be dominant or recessive, so when heterozygous, a dominant allele will override the effect of a recessive allele.

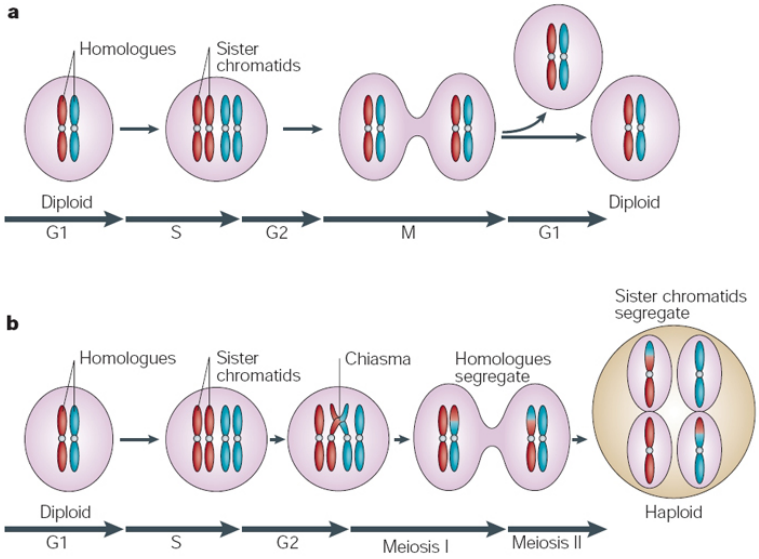


Figure 2.b: (A) Mitosis – Chromosomes of the diploid cells are being divided in the S phase of the cell cycle. During the M phase, the segregation of sister chromatids occurs to create diploid daughter cells. (B) Meiosis – Throughout the premeiotic S phase, two stages of chromosome-segregation, meiosis I. and meiosis II., undergo a single cycle of DNA replication. Homologous chromosomes are separated to opposite poles (shown in red and blue) in the meiosis I. phase. During meiosis II., the formation of non-identical haploid gametes is the result of the segregation of sister chromatids to opposite poles. It should be noted that the cell cycle stages were not drawn to scale [9].

2.1.2 Mutations

An essential part of the evolution are mutations. Any nucleotide change in the DNA sequence is considered a mutation. They can vary in size, either one single nt or large sections of DNA, influencing multiple genes. We classify them based on their origin in the following manner:

- Germ-line mutations – an individual is born with the mutation, inherited from its ancestors. The individual will have these mutations in every cell of his body.
- Acquired mutations – an individual gains these mutations during his lifetime. These mutations do not have to be present in every cell of the body, unlike germ-line mutations. Besides, as long as they happen in somatic cells, they cannot be passed to the next generation [10].

Mutations are also distinguished based on their structural characteristics, as there are three different types.

- Substitution – a change of nucleotides for other. An important case of this type is the so-called Single nucleotide polymorphism (SNP) a single nucleotide change.
- Deletion – an entire deletion of a nucleotide sequence from DNA without any replacements.
- Insertion – an insertion of a nucleotide sequence.

Mutations can be caused by environmental factors, such as radiation or, as previously stated, during the procedures that process DNA. When a mutation occurs, it can affect the host's fitness – how well the host is able to pass his genetic information to its offspring relatively to others. We then refer to the mutation as detrimental, beneficial or neutral, depending on its effect. For the sake of completeness, we give an example of a detrimental and beneficial effect of a mutation.

An example of a detrimental effect of a mutation is the so-called sickle cell disease, a blood disorder that affects the red blood cells that deliver oxygen to other cells. The disease causes the red blood cells to take an unusual shape – the shape of a sickle. This makes them breakdown prematurely, leading to health problems, primarily anemia [11].

On the other hand, the same mutation can become beneficial in some situations. It was shown that in areas where malaria is common, having only one allele with the mutation causing the sickle cell disease (i.e., being heterozygous) increases the fitness of the host. That is, since the one allele does not necessarily lead to sickle cell anemia, but also provides protection against the malaria parasite [12].

2.2 Retroviruses

Retroviruses are a virus group of the family *Retroviridae*, characterized by the fact that their genetic information is carried in the form of RNA [13]. Also, the other main characterization of this virus group is that they have an unusual replicating strategy. In contrast to DNA viruses, where the replication could be as simple as inserting their DNA into the invaded cell, retroviruses do this in two essential phases. First, using an enzyme called reverse transcriptase allows for the generation of a complementary double-stranded DNA from the viral RNA. Subsequently, this complementary DNA is randomly ensued into the DNA genome of the host cell, using the enzyme integrase [14]. When a retroviral genome is integrated into the DNA of the host cell, the genome is then referred to as a provirus. This act, as a

result, changes the genome of the host cell, and thus, as the cell proliferates, the viral genetic information is being replicated. This genetic information is found in all progeny cells at the same chromosomal position and fades with the last clone cell, which can usually only happen after the death of the host. This process is visualized in figure 2.c.

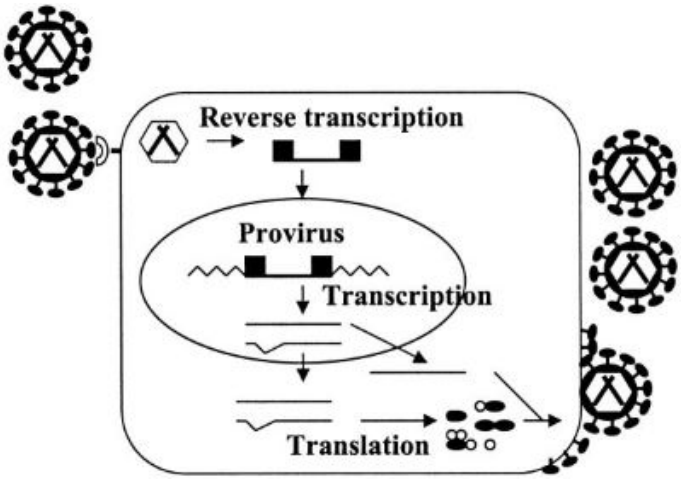


Figure 2.c: The replication cycle of retroviruses [15]. After retroviruses enter a cell, through reverse transcription, they create a DNA copy of their genome, which is then integrated into the DNA of the cell. The provirus may then be transcribed and translated, creating new copies of the retrovirus.

2.2.1 The endogenization process

Two paths can usually be observed by a virus spread: horizontal and vertical. The spread between individuals of the same generation is referred to as horizontal, while vertical spread happens among mothers and their successors [16]. Typically, only the somatic cells are being infected by retroviruses. Yet, sporadically a retrovirus infects a germ-line cell. Thus, any successor produced from an infected germ-line cell is a carrier of the provirus in all of its nucleated cells (at the same chromosomal position). These proviral elements are also transferred onto the next successors, becoming fixed in the gene pool of the host population. In the aftermath of endogenization, the provirus can maintain both paths of the spread mentioned above. The duration of this stage and the proviral frequency obtained in the host population is primarily decided by the effect of the integration on the fitness of the host. Integrations that either cause serious detrimental or pathogenic effects typically do not tend to achieve high allelic frequencies, therefore they do not sustain for long in the host population. However, the allelic frequency is presumed to increase in the event that the integration is neutral or beneficial [2].

2.3 Evolutionary genetics

Evolutionary forces operate on the genome and alter the frequencies of the population's variants. These forces are genetic drift, mutation and recombination, migration, selection [1]. Among these, only selection leads to adaption [17]. In this section, we mainly analyze selection, but we also go through the other forces to give a full picture.

2.3.1 Genetic drift

Genetic drift, sometimes even called random genetic drift, describes the randomness in evolution – when an allele frequency increases or decreases in a population by chance, without taking into account the effect the allele has on the fitness of an individual. This acts through the probabilistic mechanism of chromosome sampling, which will produce the succeeding generation of individuals. It prevails in small and isolated populations [18] as the gene pool is small enough for randomness to make a difference. On the contrary, when a particular allele is shared by a plenty amount of individuals in larger populations, the spread of this allele is almost unavoidable unless it is biologically disadvantageous [19]. Thus, genetic drift commonly occurs after population bottlenecks – events that greatly reduce the population size. Once genetic drift occurs, it continues until ultimately one of two alternatives happens: the affected allele is lost or until this allele is the only one in the population. The genetic diversity of a population is diminished by both alternatives. As might be expected, this evolutionary force plays a major part in our hypotheses, since it can effectively conceal or adjust the effects of selection on loci.

2.3.2 Mutation and recombination

Mutation, as mentioned previously, are an essential part of evolution. They introduce new genetic variability as they happen. It was shown that the mutations occur with a very low probability, known under the term mutation rate. Even though it is very low [7], it is statistically inevitable, due to the size of the genome. The urge to change allele frequency due to mutation is defined as mutation pressure.

Recombination essentially introduces new genetic variability as well, enhancing the ability to adapt. It happens during the process of exchanging genetic information between chromosomes during meiosis. On the genome, there are regions that increase the likelihood of recombination to happen, the so-called recombination hotspots. Therefore, mutations that are close to each other on the chromosome are

characterized as linked, because of the fact that they are usually inherited together (referred to as haplotypes) since it is less likely for recombination to happen in between them. The property, when two alleles are placed on the same chromosome more often than expected if the alleles were divided by random is called Linkage disequilibrium (LD) [1], [7].

2.3.3 Migration

Migration, also called gene flow, is the movement of a species towards more desirable places. As such, migration cannot change the allele frequency in the species. A population would have to be divided into subpopulations confined for a sufficient time to acquire variations between the subpopulations to have an effect on the allelic frequencies.

2.3.4 Selection

Selection was first introduced to us by Charles Darwin in 1859 [20]. As one might know, it is a non-random process that ultimately ends up with an individual adapting to its local environment through selective genetic variations [21]. Nowadays, we differentiate between types of selection [22], as described below.

The different types of selection

First, we call the removal of deleterious mutations as negative or purifying selection, as defined by Kimura [23] and Lewontin [24], since if any mutations are deleterious, they soon disappear from the population. In the same manner, the selection of favorable mutations is referred to as positive selection, also defined by Kimura. These two selections are the extremes to their middle ground, neutral selection, the most common occurrence in evolution.

The usually taught definition of selection is the one describing the force, creating pressure on one's ability to reproduce. In general, this is influenced by competition between other species, parasites, predators, etc. This type of selection is the natural selection or sometimes also referred to as Darwinian selection. A special type of this selection, the sexual selection, is where an individual chooses a mating partner according to its characteristics to reproduce. We typically cannot distinguish between these two when searching for genomic signs of selection. Since according to Nielsen [25], in a model of diploid organisms with bi-allelic loci (that is, only two alleles exist) selection happens if the fitness of the three genotypes (heterozygote, recessive homozygote, dominant homozygote) is not equal. One, therefore, must

search for the SNP frequencies distinguished from the random changes of the genetic drift [1].

In general, we categorize selection into these types (visualized in figure 2.d):

- Stabilizing selection – The stabilization of the most frequent variant in the population. This selection is said to occur, when the population stabilizes on a variant that is not an extreme. It is one of the most common types of natural selection, since most variants do not change in frequency drastically.
- Directional selection – A selection where one of the alleles and its variants is favored, causing the shift over time in the direction of that allele variants.
- Disruptive selection – When a change in the environment increases the rare variant's fitness, this selection occurs. The most frequent variant becomes relatively smaller in comparison to what used to be the rare variants.
- Balancing selection – This selection maintains stable alleles in the population due to the fact that all of them are approximately the same in terms of fitness, leaving all alleles in the populations and, as a result, increasing the variability.

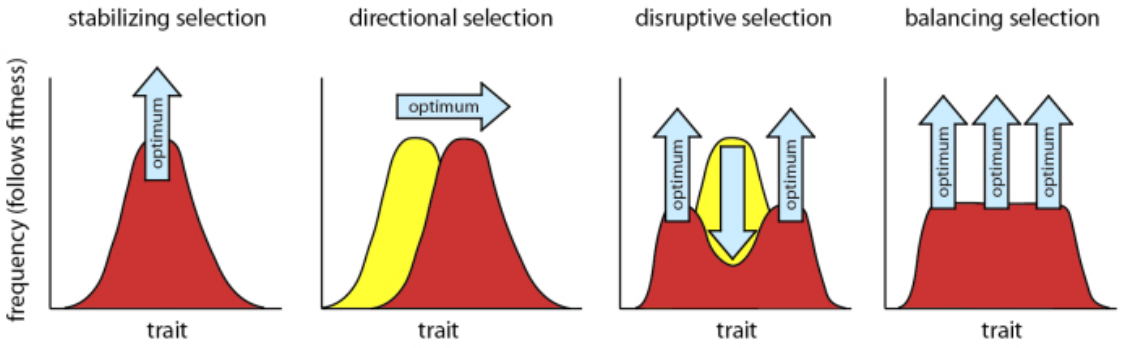


Figure 2.d: Selection types – The red area refers to the current distribution of frequencies of a trait in a population. The yellow area refers to the distribution of the former generations. The arrows labeled with "optimum" indicate the movement of the optimal trait, therefore showing positive selection. The unlabeled arrow shows negative selection [26].

Due to the fact that LD occurs, the simplified bi-allelic model will not represent real situations happening in genomic regions. Since the LD will carry mutations in close proximity to a genomic region acted upon by selection, the frequency of these mutations also changes. This is first introduced by Smith et al. [27] and is called genetic hitchhiking. Genetic hitchhiking is a fundamental part of a process named selective sweep. A selective sweep happens when beneficial mutations fixate

in a population, but thanks to the hitchhiking, linked mutations fixate as well. This reduces the variation in the vicinity of the beneficial mutation. When negative selection eliminates hitchhiked linked mutations, we call it the background selection [28].

These selective sweeps are usually categorized into two types, based on their impact on the population – soft and hard. Hard selective sweep dominates the population, leaving only the selected region in the population – fixating. We show a visualization of a hard selective sweeps for easier understanding in figure 2.e. Soft selective sweeps fixate only partially. Hard selective sweeps can be identified as an excessive amount of variants that were in low frequencies (low-frequency variants).

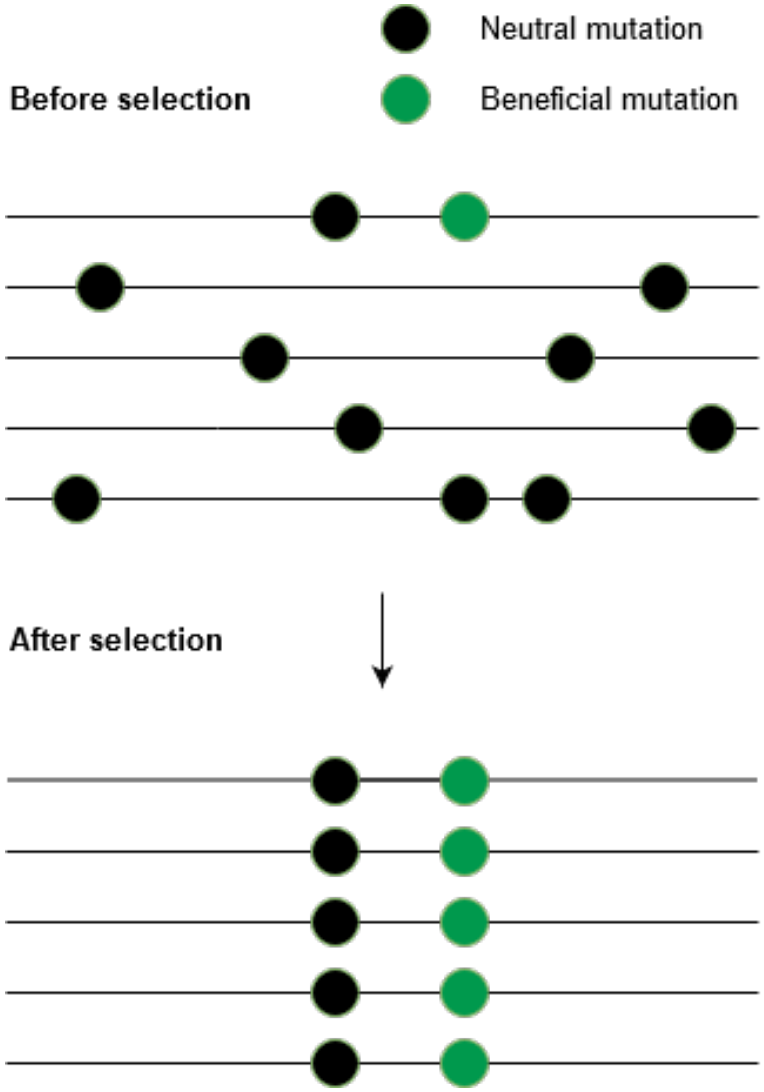


Figure 2.e: A hard selective sweep – Each line represents a DNA sequence, dots represent mutations. When a beneficial mutation (green dot) spreads in the population, linked neutral mutations spread as well. A hard selective sweep then fixates in the population, removing all other variations [29].

Chapter 3

Estimating selection pressures based on site frequency spectrum

In this chapter, we tackle the problematics of estimating selection pressures based on site frequency spectrum. We first show the model under which the usual null hypothesis is stated, and afterward, we examine the ideas behind all the tests implemented during the work of this thesis.

3.1 The neutral theory of molecular evolution

Firstly proposed by a Japanese biologist Motoo Kimura in [30], the neutral theory of molecular evolution (hereafter referred to as "the neutral theory") states that "the main cause of evolutionary change at the molecular level – changes in the genetic material itself – is the random fixation of selectively neutral or nearly neutral mutants rather than positive Darwinian selection" [23]. In other words, most of the mutations that happen are neutral. Thus they do not affect the fitness of an individual and, therefore, the genetic drift is the main evolutionary force. The importance of this theory is that it creates a model, under which one can test the selection pressure. To show how the model functions, we first introduce some variables.

3.1.1 Effective population size

The rate at which the genetic drift occurs depends on how well an individual in a generation is able to find a partner and mate. Nevertheless, the proportion of sexes in a population may not be equal. In such a case, the chance of reproduction would

be dependant on one's sex, meaning that a large population would not necessarily be large from the evolutionary point of view. Therefore an ideal population size measure was introduced, the effective population size (N_e). An ideal population has these characteristics [31]:

- An equal number of males and females, all able to reproduce.
- Random mating.
- Throughout generations, there is a constant number of breeding individuals.
- Equal mating probability.

Let us denote N_f as the number of females in the population and N_m as the number of males. The effective population size is then calculated as

$$N_e = \frac{4N_f N_m}{N_f + N_m}, \tag{1}$$

If we consider again a population of only one sex, one can see that $N_e = 0$, which makes sense, as that population is not able to reproduce. Figure 3.a shows the relationship of N_f (without loss of generality, because $N = N_f + N_m$) and N_e for a population of size 1000.

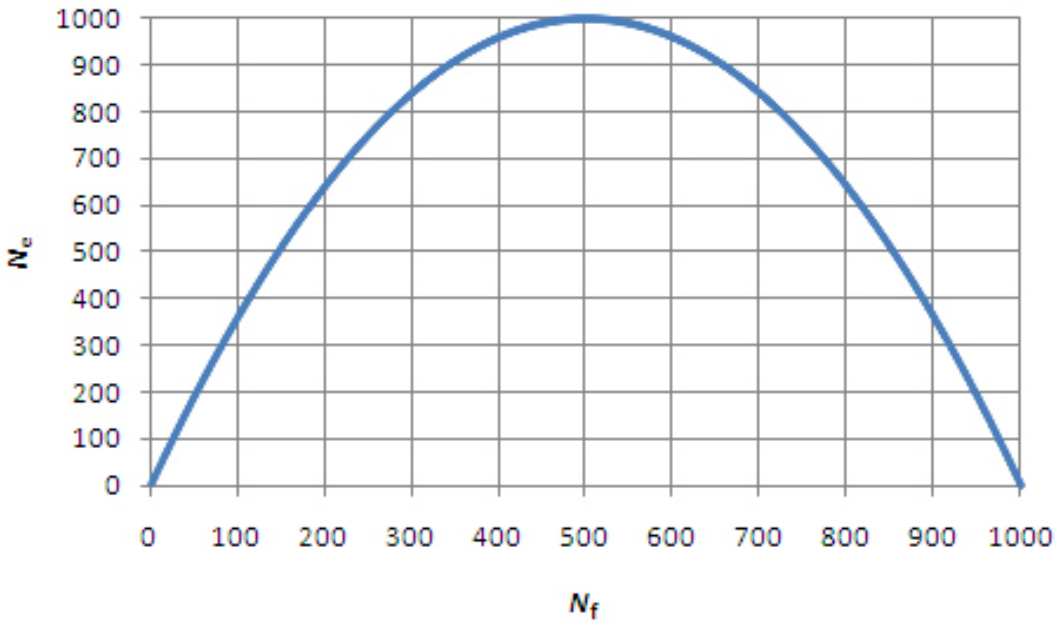


Figure 3.a: The dependency of the number of females in a population of size 1000 to the effective population size [31].

3.1.2 Selection coefficient

As to be expected, the selection coefficient (s) is simply a measure of relative fitness. When $s > 0$, it indicates an advantage, while $s < 0$ indicates a disadvantage. Not only in this thesis is this measure purely theoretical due to the fact that one is not able to measure human being's fitness since some data necessary for its calculation cannot be measured. That is why selection pressures are being calculated, estimating this coefficient in a broad sense.

3.1.3 Effectiveness of selection

Having defined the previous variables, one can now theoretically calculate the effectiveness of selection on a mutation. It is dependant both on the effective population size and the selection coefficient. When $N_e s < 1$ it means that the genetic drift will occur, and the frequency of that mutation will fluctuate due to pure chance, essentially meaning they are neutral [32]. As mentioned earlier, this also means that the genetic drift will prevail in small populations. Otherwise, when $N_e s > 1$, the mutation is under selection.

3.1.4 The null hypothesis

Since the neutral theory states that most of the mutations happening are neutral, if one is able to reject this statement on the sequence under examination, it will show that it is subject to selection. In [30], it was shown that under the neutral model, the rate of substitutions (k) is equal to the rate of mutations (μ). However, in the event of positive selection, the rate of substitution will be higher than the rate of mutation, due to the fact that beneficial mutations have a higher chance of fixating compared to neutral mutations ($k > \mu$). On the other hand, detrimental mutations fixate slower; therefore, the rate of mutation will be higher than the rate of substitution ($\mu > k$). This simple yet powerful fact was used in the development of the tests to detect selection [32].

3.2 Infinite site model

Kimura also developed the infinite site model [33], in which he showed the basics of the development of new alleles thanks to mutation. In this model, allele frequencies were used to estimate a theoretical measure, the genetic diversity (heterozygosity) of an examined population's genomic locus.

This model assumes the following:

- A randomly mating population.
- No occurrence of selection or recombination between DNA sequences.
- The number of sites (i.e. the number of nucleotides) in a DNA sample is infinite, therefore when a new mutation occurs, it must occur on a site where no previous mutation occurred.

Under these circumstances, the measure of genetic diversity of a diploid populations genomic locus is proposed as

$$\theta = 4N_e\mu, \tag{2}$$

where N_e is the effective population size and μ is the mutation rate per generation. For a haploid population, only the coefficient in equation (2) is changed for 2. It should be emphasized that θ cannot be negative. Many estimators based on retrievable data were invented. One of the used estimates is the famous Watterson estimator [34]. Let us consider a sample of n DNA samples. Watterson showed that under the neutral theory

$$E(S) = a_1\theta, \tag{3}$$

S denoting the number of segregating sites (the number of sites that differ e.g. SNP) and

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i}. \tag{4}$$

From this, an estimator of θ, the Watterson estimator, is as follows

$$\hat{\theta}_W = \frac{S}{a_1} \tag{5}$$

An example of the calculation of this estimate is given in the following section. The value of this estimate is in the fact that it uses the number of segregating sites in its calculation. For one, the number of segregating sites can be very easily retrieved from the data, and also, it ignores the frequency of the mutations as it measures only whether a mutation happened or not. The importance of this is shown by Fumio Tajima in [35].

3.3 Tajima's D

Tajima's D, named after the already mentioned Fumio Tajima, is a widely used statistical test used to find the presence of a non-random process in the evolvement of a DNA sequence [35]. It was constructed under the infinite site model and the

neutral theory. Consider a sample of n DNA samples. Let us denote the number of nucleotide differences between i-th and the j-th DNA sequence as π_{ij}. The average number of pairwise nucleotide differences (π) is then given by

$$\pi = \frac{\sum\sum_{i<j} \pi_{ij}}{\binom{n}{2}}. \tag{6}$$

Tajima has then shown in [36] that under the neutral theory $E(\pi) = \theta$, therefore, a new estimator of theta is introduced as

$$\hat{\theta}_\pi = \pi. \tag{7}$$

It should be noted that $\hat{\theta}_\pi$, unlike $\hat{\theta}_W$ does not ignore the frequency of mutants. As both equations (5) and (7) are estimates of θ, and thus should be roughly equal under the null hypothesis, Tajima states: "the remarkable and important difference between the number of segregating sites and the average number of nucleotide differences is the effect of selection". That is because detrimental alleles are usually kept in low frequencies (i.e., they are rare), but the number of segregating sites does not take into account frequencies of mutations, meaning $\hat{\theta}_W$ will be strongly affected unlike $\hat{\theta}_\pi$. This implies the fact that if there are any selective effects, $\hat{\theta}_\pi$ will differ from $\hat{\theta}_W$ by a statistically significant value. One must also account for the size of the sample. Therefore, from the statistical properties of $\hat{\theta}_\pi$ and $\hat{\theta}_W$, the variance was derived for normalization. The test statistics then becomes

$$D = \frac{\pi - \frac{S}{a_1}}{\sqrt{Var(\pi - \frac{S}{a_1})}}. \tag{8}$$

During the derivation of equation (8), it was conducted that the distribution of Tajima's D is not far from the normal distribution, although not normal. Hence, as for the statistically significant value, it is said to be expected for Tajima's D to lie in the interval $[-2, 2]$ 95% of the time. Values of Tajima's D outside of the interval should be examined, for example, using p-values. For the calculation of p-values, we used the rank score method from [37] (http://hsb.upf.edu/). The process is as follows:

1. Calculate the test estimates.
2. Depending on the test, rank each estimate from the lowest to highest or vice versa.
3. Calculate the p-value for each estimate as the estimates rank divided by the number of all estimates.

4. For visualization purposes, transform each p-value x as $-log_{10}(x)$. Thus, low p-values will become large values.

One can construct biological implications from the analysis of the values of Tajima's D, shown in table 3.1. When Tajima's D is around 0, one cannot reject the null hypothesis of the DNA sample evolving at random. For positive Tajima's D not close to 0, we can say there is a shortage of low-frequency alleles. This could be caused, for one, by a balancing selection, since it maintains the stable (intermediate-frequency) alleles – for example, distinct subpopulations that are genetically very different are all being selected. Alternatively, for two, a population shrinkage, which can lead to the same example – having one population, a shrinking event occurring leading to genetically very different subpopulations. For negative Tajima's D not close to 0, we say there is a large number of low-frequency alleles. The biological implications could be: a sudden population expansion – there was not enough time for the population to acquire variability – or a selection removing variation such as the stabilizing selection.

Tajima's D values	Analysis
$D = 0$, therefore $\pi = \frac{S}{a_1}$	Cannot reject the null hypothesis, meaning the sample is not under selection
$D < 0$, therefore $\pi < \frac{S}{a_1}$	Average number of rare alleles is higher than the number of mutations, implying a large number of rare alleles
$D > 0$, therefore $\pi > \frac{S}{a_1}$	Average number of rare alleles is lower than the number of mutations, implying a shortage of rare alleles

Table 3.1: Analysis of possible Tajima's D results

For the purpose of example calculation of Tajima's D, we include the following formulas for the calculation of the variance. Subsequent equations used to calculate other tests that do not help to show the idea of the test are provided in appendix A.

$$a_2 = \sum_{i=1}^{n-1} \frac{1}{i^2}, \tag{9}$$

$$b_1 = \frac{n+1}{3(n-1)}, \tag{10}$$

$$b_2 = \frac{2(n^2+n+3)}{9n(n-1)}, \tag{11}$$

$$c_1 = b_1 - \frac{1}{a_1}, \tag{12}$$

$$c_2 = b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2}, \tag{13}$$

$$d_1 = \frac{c_1}{a_1}, \tag{14}$$

$$d_2 = \frac{c_2}{a_1^2 + a_2}. \tag{15}$$

Given these, Tajima's D is then

$$D = \frac{\pi - \frac{S}{a_1}}{\sqrt{d_1 S + d_2 S(S-1)}}. \tag{16}$$

subjects	0 1 2 3 4 5 6 7 8 9
subject 1	GAAAA AAAAA
subject 2	AGAAA AAAAA
subject 3	AAGAA AAAAA
subject 4	AAAGA AAAAA
subject 5	AAAAG AAAAA

Table 3.2: Simple DNA sequences, where the sites are labeled with numbers from 0 to 9, used for the example calculation of Tajima's D.

Consider the DNA sequences with its site labeling from table 3.2. We have 5 DNA sequences, hence $n = 5$. The number of segregating sites is $S = 5$, because there are 5 variants of sites - the sites labeled from 0 to 4. Sites 5 to 9 are not segregating, because there are no variants. Comparing pairwise each subject, that is $\binom{5}{2} = 10$ combinations, where $\pi_{ij} = 2$ for $\forall i, j \in \{1, 2, 3, 4, 5\} \wedge i \neq j$ (e.g. subject 1 and subject 2 differ at sites 0 and 1), we have from equation (6)

$$\pi = \frac{2+2+2+2+2+2+2+2+2+2}{10} = 2$$

Next, equation (4) and equations from (9) to (15) are calculated, all dependant on the number of sequences. Results are shown in table 3.3.

a_1	2.0833
a_2	1.4236
b_1	0.5
b_2	0.3667
c_1	0.02
c_2	0.0226
d_1	0.0096
d_2	0.0039

Table 3.3: Values of equations from (9) to (15) and equation (4) used in the example calculation of Tajima's D.

From that, using equation (15)

$$D = \frac{2 - \frac{5}{2.0833}}{\sqrt{0.0096 \times 5 + 0.0039 \times 5 \times 4}} = -1,1268$$

It should be noted that for the calculation of this test and for all the following tests, the sites that did not differ (sites 5-9) were not needed in the calculation. This fact is used in the manner of DNA sequence storing, mentioned in the next chapter.

3.4 Fu & Li's D* and F*

In the aftermath of the publication of Tajima's D, new ideas to test the neutrality of mutations have been proposed, but from different views. In [38], it was proposed to test neutrality with the usage of the coalescent theory, although along the same lines as Tajima's D. This model uses the information collected from the ancestors of the sampled population.

Genealogy, a study of family lineages, uses tree diagrams to show an individual's family tree. One can map these family trees to genes, sometimes called gene genealogies, meaning how genes acquire mutations over time, becoming other genes. An example of such a tree is visualized in figure 3.b.

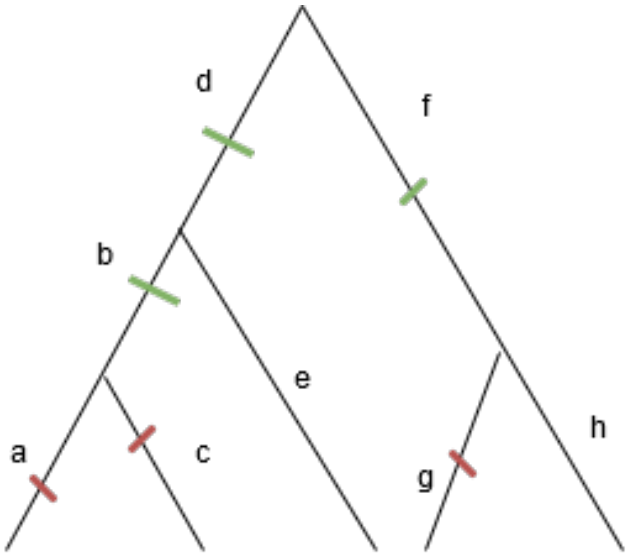


Figure 3.b: A gene genealogy example with labels. External branches are labelled as a,c,e,g,h, internal branches as b,d,f, with both external (in red) and internal (in green) mutations.

Different terminologies are used in these trees as opposed to graph theory. The term "edge" in graph theory is here referred to as "branch". Further, an "external branch" is an edge between a node and a leaf. An "internal branch" connects two nodes, of which none are leaves. From this, external mutations are mutations that happened on an external branch and internal mutations on internal branches.

The main idea of the Fu & Li's tests is using the fact that the number of external mutations is equal to the number of internal mutations under neutrality. Again, we analyze detrimental mutations. In a genealogical tree, since the detrimental mutations are acquired recently in the sample, they assemble in the external branches. These mutations are then called singleton mutations, or simply, singletons – also defined as alleles that appear only once in the sample population. Not all singletons have to be in the external branches, though [39]. For a population of size n, Fu and Li then suggest the following tests:

$$D_o = \frac{\eta - a_1\eta_e}{\sqrt{Var(\eta - a_1\eta_e)}}, \tag{17}$$

$$F_o = \frac{\hat{\theta}_\pi - \eta_e}{\sqrt{Var(\hat{\theta}_\pi - \eta_e)}}, \tag{18}$$

where η is the number of all mutations and η_e is the number of mutations in the external branches. To get the number of external mutations, however, an outgroup is required to infer the number of external mutations accurately. An outgroup is a

group of organisms distantly related to the examined population (e.g., chimpanzees for humans). Fu and Li expected that an outgroup may not always be available, thus suggesting these tests without the need for an outgroup. Denote η_s as the number of singletons. Then the tests are

$$D^* = \frac{\frac{n}{n-1}\eta - a_1\eta_s}{\sqrt{Var(\frac{n}{n-1}\eta - a_1\eta_s)}}, \tag{19}$$

$$F^* = \frac{\frac{n-1}{n}\eta_s - \hat{\theta}_\pi}{\sqrt{Var(\frac{n-1}{n}\eta_s - \hat{\theta}_\pi)}}. \tag{20}$$

As has already been mentioned, these tests are along the lines of Tajima's D, showing very similar results. Nonetheless, certain selection scenarios, such as selective sweeps, are more prone to Fu & Li's tests than to Tajima's D, due to them generating an excessive amount of singletons.

In practice, formulas from [40] are used for the implementation of Fu & Li's D* and Fu & Li's F*, also included in appendix A.

3.5 Fay & Wu's H

Let ξ_i be the number of mutations found i times in a sample population of size n . In [41] Fu showed that

$$E(\xi_i) = \frac{\theta}{i}. \tag{21}$$

This fact allowed for many other possible approaches of estimating θ as well as a revision of already known estimators at that time, being

$$\hat{\theta}_W = \frac{1}{a_1}\sum_{i=1}^{n-1}\xi_i, \tag{22}$$

$$\hat{\theta}_\pi = \binom{n}{2}^{-1}\sum_{i=1}^{n-1} i(n-i)\xi_i. \tag{23}$$

One of the new approaches was introduced by Fay and Wu in [42] as

$$\hat{\theta}_H = \binom{n}{2}^{-1}\sum_{i=1}^{n-1} i^2\xi_i, \tag{24}$$

because of the fact that it is strongly influenced by high-frequency mutations. Hence, a comparison of $\hat{\theta}_\pi$ and $\hat{\theta}_H$ would grant a way for the identification of an excessive amount of high-frequency mutations. Very few high-frequency mutations are found in a population under neutrality. However, directional selection leads to a selective sweep that increases the frequency of linked high-frequency mutations as the sweep almost completes or in adjacent regions after fixation [7]. This is what the following test could imply.

$$H = \hat{\theta}_\pi - \hat{\theta}_H \tag{25}$$

A problem of this test is that the variance of $\hat{\theta}_H$ was not easily obtainable. Therefore, a normalization was not possible. This was later tackled by Zeng et al.'s in [43] with the proposal of a new estimator of θ as

$$\hat{\theta}_L = \frac{1}{n-1}\sum_{i=1}^{n-1} i\xi_i, \tag{26}$$

of which the variance was easier to obtain. Since it followed that

$$\hat{\theta}_H = 2\hat{\theta}_L - \hat{\theta}_\pi, \tag{27}$$

the normalized version of the Fay & Wu's H was written as

$$H = \frac{\hat{\theta}_\pi - \hat{\theta}_L}{\sqrt{Var(\hat{\theta}_\pi - \hat{\theta}_L)}}. \tag{28}$$

3.6 Zeng et al.'s E

With the introduction of $\hat{\theta}_L$, Zeng et al. also proposed a new test, known as the E test or Zeng et al.'s E, as

$$E = \frac{\hat{\theta}_L - \hat{\theta}_W}{\sqrt{Var(\hat{\theta}_L - \hat{\theta}_W)}}. \tag{29}$$

As previously outlined, $\hat{\theta}_W$ is strongly affected by low-frequency mutations, and now, $\hat{\theta}_L$ is strongly affected by high-frequency mutations. A negative E will then signify an excess of low-frequency mutations, occurring immediately after a sweep, again. The exclusive feature of this test is then the indication of a selective sweep [44], in combinations with the other tests.

Chapter 4

Related work

This chapter first clarifies what data were used for the estimation of selection pressure were accessed. Afterward, we take a look at several other frameworks and packages for estimating selection pressure using the neutrality tests.

4.1 The 1000 genomes project

The start of genome sequencing evoked a large number of projects, one of them being the 1000 genomes project, that started from 2008 to 2015 [45]. This project's goal was to find most of the genetic variants in the human population. As a result, the data of human genome variants were made available for the public through freely accessible databases.

The project was divided into 3 phases. The data from the final phase consists of variants from 2504 people from 26 different populations (shown in figure 4.a), stored in a special text file format - Variant Call Format (VCF), an effective way of storing genetic variants, since most of the human DNA is shared, therefore, only variants need to be stored.

4.2 VCFtools

Due to the rise in popularity of data storing using the VCF format, projects for the parsing of the data files were developed. Among these projects, VCFtools [46] became extensively used. VCFtools allows an easy summary, filtering, and all sorts of analyses of VCF data files, including the Tajima's D test. It does not, however, include the rest of the neutrality tests we used. For the purpose of correctness,

VCFtools was used as one of our reference calculations of Tajima's D, among other uses in our pipeline (see chapter 5).

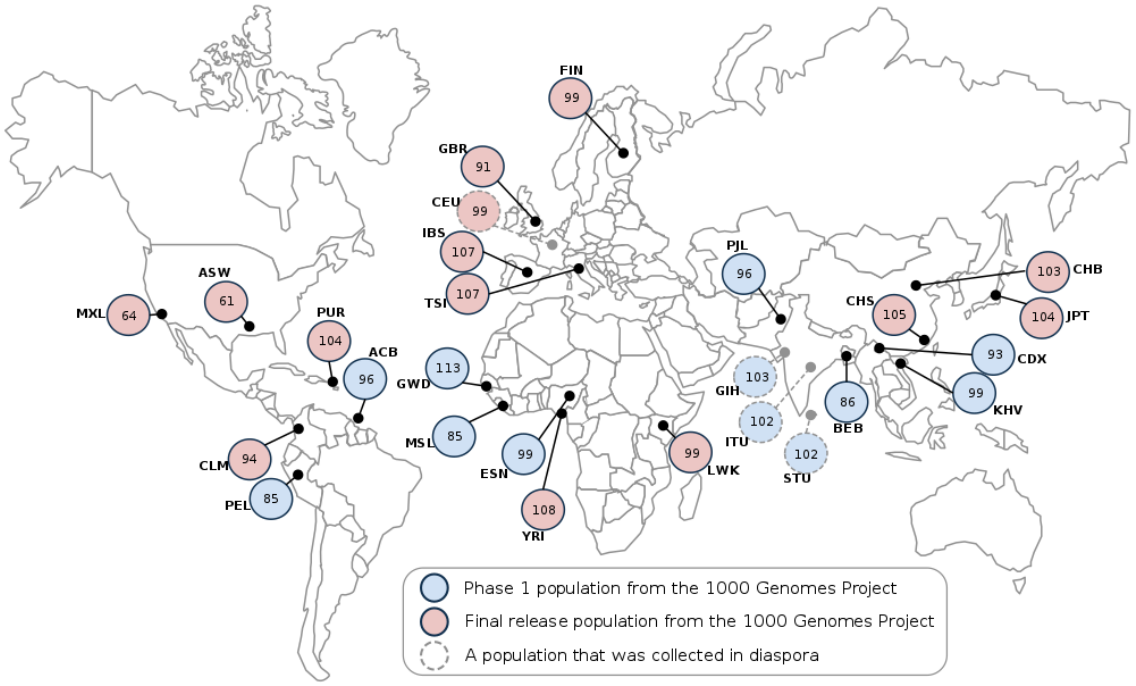


Figure 4.a: The populations from which the genetic variants were acquired [47].

4.3 PopGenome

PopGenome [48] is an R software package, widely used for the calculation of population genomic data analyses. With the use of other packages of R, one can directly use a VCF file for the analyses. This package involves the implementations of all tests mentioned. However, for an unknown reason, when we tried to calculate the Fay & Wu's H and Zeng et al.'s E tests, the value was not shown. Hence, we use PopGenome for reference calculations of Tajima's D, Fu & Li's D*, and Fu & Li's F*. Also, during the trial, substantial RAM requirements were demanded to calculate these tests of only relatively small sized loci. Our own developed solution does not demand such memory requirements.

4.4 The 1000 Genomes Selection Browser 1.0

The 1000 genomes selection browser 1.0 [37] is an example of the end goal of this project. It is a visualized online database of estimates of selection pressures on whole chromosomes, without the Zeng et al.'s E test. The tests, however, were calculated based on populations. Our estimates are meant to be calculated on the whole population of the human species, no matter the population of origin. Also, it

was not calculated using the final phase of the 1000 genomes project data, therefore, it is not up-to-date.

4.5 PopHuman

PopHuman, as mentioned in the introduction, is a population genomics online database with which we will compare some of our estimates in order to check the correctness of our estimates. In our work, we calculated estimates in different window sizes than it was done in PopHuman. Window size used in our work was based on the size of average HERVs, which is several thousands of nucleotides.

Chapter 5

Implementation and results

We now provide information about the structure of the data and go through the selection pressure estimation process. After that, we compare our estimates with other sources and conclude with a few current usages of our estimations.

5.1 The data

As mentioned in the previous chapter, the publicly available data of the human genome variants is freely accessible from the official site of the 1000 genomes project (https://www.internationalgenome.org/). The variants of the genomes are divided into VCF files according to the chromosome of the variant. A file containing the annotation of the data (i.e., the population of origin of the individual, sex of the individual, etc.) is also included. As even the compressed genome variant files are about 15 Gigabytes large, for time efficiency reasons, we use an already popularized and highly cited tool for the VCF file parsing, likewise mentioned in the previous chapter, VCFtools.

5.2 Data filtering

Due to the nature of our goal, one cannot do estimations on DNA sequences that contain insertion and deletion mutations (hereafter referred to as "indels"). Data analyses of sequences including indels are much more complex and are approached in different ways. Also, the VCF file contains additional information, such as metadata, identification numbers etc., which we do not need for our calculation. Hence, we use the VCFtools filtering option to remove indels and create an output file containing only the counts for each allele on each position. The filtered data is a tab-separated

text file, containing five columns – the chromosome number, the position of the variant on the chromosome, number of alleles, number of chromosomes (that is, 5008 for a diploid population of size 2504) and the allele count as can be seen in table 5.1.

CHR	POS	N_ALLELES	N_CHR	{ALLELE:COUNT}
22	16050075	2	5008	A:5007 G:1
22	16050115	2	5008	G:4976 A:32
22	16050213	2	5008	C:4970 T:38
22	16050319	2	5008	C:5007 T:1
22	16050527	2	5008	C:5007 A:1
22	16050568	2	5008	C:5006 A:2
22	16050607	2	5008	G:5003 A:5
22	16050627	2	5008	G:5006 T:2
22	16050646	2	5008	G:5007 T:1
22	16050655	2	5008	G:5007 A:1

Table 5.1: An example of the VCFtools output, shown in a table. First ten variants of the tab-separated text file created from using the counts option of VCFtools with indel filtering, as described.

5.3 Data processing

Having the frequency text files, one can easily calculate all estimators mentioned in chapter 3, thus also the estimates of the selection pressures. The selection pressure estimates are usually calculated in windows since the neutrality tests measure the selection pressure of given the DNA sequences. Hence, from the whole chromosome, one takes a region of the selected window size out and calculates the estimate on that particular region. As same as the 1000 genomes selection browser 1.0, we use a window size of 30000 nt, but also of 5000 nt, because of the size of HERVs. We also consider a bi-allelic model, because it has been standardized and used among other packages, for example, the mentioned PopGenome package, even though it has been said that the model does not represent real situations happening in genomic regions due to the LD. That means, for the calculation of our tests, we use only sites that have exactly two alleles. The way we present our output is similar to the VCFtools output, meaning four tab-separated columns – the chromosome number, the starting position of the window, the number of sites included in the calculation, and the value of the test. An example of that is shown in table B.2. The implementation of the tests was done in Python with the use of an external library, NumPy, and the

whole process of estimation was done using the computational resources of ELIXIR infrastructure.

CHR	BIN_START	N_SITES	D
22	16050000	79	2.51434622116
22	16080000	12	0.984047060837
22	16110000	16	1.60635258458
22	16140000	40	2.01096917906
22	16170000	16	0.655183029176
22	16200000	46	2.24605070249
22	16230000	46	2.84298117536
22	16260000	42	2.86850790983
22	16290000	22	2.37278613932
22	16320000	23	1.58701517644

Table 5.2: An example of our output file, shown in a table. First ten values of Tajima's D with bins of size 30000 of the tab-separated text file created from our implementation of Tajima's D.

5.4 Comparison

To show the correctness of our implementation, we downloaded several randomly chosen chromosomal loci with the available estimates from the PopHuman database, ran our implementations, the R software PopGenome package and VCFtools tests on the loci, and compared them with each other. The PopHuman database has estimates on window sizes of 10000 and 100000. We took five CEU population samples at random of size 10000 from the PopHuman database and performed the comparisons; the results are presented in the table 5.3.

Chr	Position start	Our	R PopGenome	VCFtools	PopHuman
22	30800000	-0.457218017865	-0.457218	-0.457218	-0.399
10	66700000	0.730934542079	0.730935	0.736635	0.855
6	78390000	0.61169510819	0.6116951	0.611695	0.612
12	60950000	0.348557429955	0.348557	0.348557	0.276
3	96950000	-2.38428711298	-2.384287	-2.38429	-2.38429

Table 5.3: Comparison of our calculation of Tajima's D values with three other references.

We found that the PopHuman database probably has an error in their implementation of Tajima's D calculations. Since other sources show similar results, each independent of one another, we conclude our estimate of the Tajima's D test is correct. We do not, however, check other tests with the PopHuman database, since the database shows incorrect results. The Fay & Wu's H and Zeng's E could not be compared, because no other available resource is known to us, although we believe they are correct, due to the fact that they are calculated similarly to Tajima's D. For Fu & Li's D* and Fu & Li's F*, they were compared on the same loci as Tajima's D, but only with the R PopGenome package, tables included in appendix B. As the values differ only insignificantly, we conclude our estimates as correct. The reason for the difference could be due to the way VCFtools calculates the allele counts, very occasionally showing an allele that does not exist, increasing the allele count, which makes a bi-allelic site not bi-allelic.

5.5 Use of our estimations

Our results will be used in an online database, HERVd [49] (https://herv.img.cas.cz/), created by the Institute of Molecular Genetics, Czech Academy of Sciences. The database will visualize our results of HERVs and their vicinity for other scientific purposes in a similar way as shown in figure 5.a. Also, our results are currently used in a different diploma thesis of a student from the University of Chemistry and Technology, Prague [50].

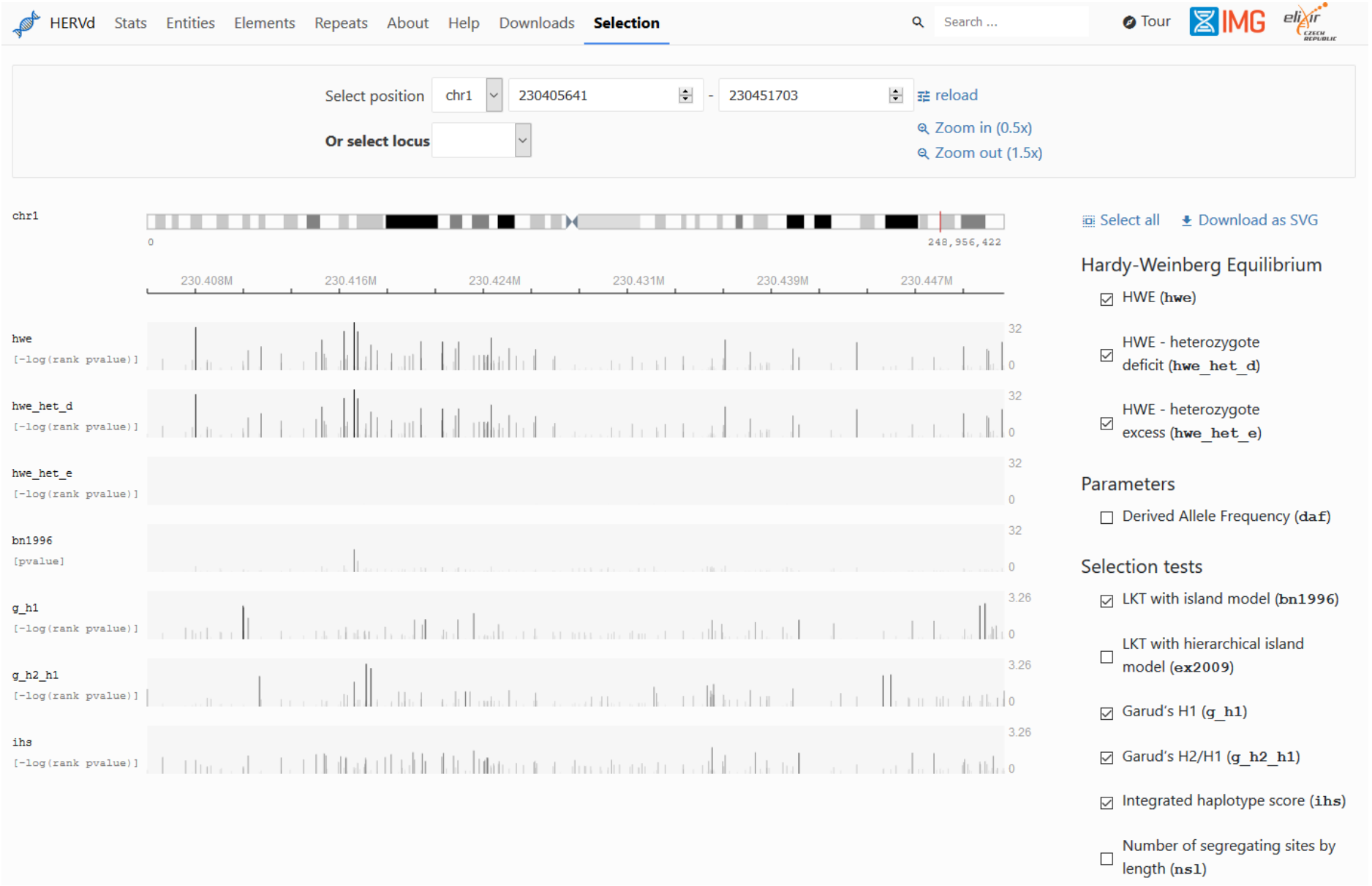


Figure 5.a: A visualization of a few of the currently available estimates in the HERVd database.

Chapter 6

Conclusion

The aim of this thesis was to estimate selection pressures based on site spectrum frequency of HERVs and their vicinity. The motivation is that HERVs are account for a relevant amount of the human genome, were correlated with a variety of pathologies and many of their functions are yet unknown to us. Showing any signs of selection on them could indicate biological functions.

We first provide the reader with basic biological knowledge for the understanding of our task. Afterward the explanation of the tests for estimating selection pressure based on site frequency spectrum was given, including the null hypothesis under which these tests are stated.

Next, we analyzed other attempts of estimating these pressures, showing a few examples. From these attempts, we used some to confirm the correctness of our estimates. Then we explained our pipeline of the actual estimation which was followed by the comparison of the estimates with available data. Thus, we implemented these methods in python scripts, creating a framework containing neutrality tests. We concluded that our tests are implemented correctly, from the information available to us. Results are being made publicly available through the HERVd database (herv.img.cas.cz/), one of the key worldwide known databases focused on retroviral research. HERVd is database supported by ELIXIR (https://elixir-europe.org/), a large European bioinformatics infrastructure. This allows our results to be recognized and easily accessible by scientific community.

6.1 Future work

Future work could involve providing a way of access to the framework for the public. Also, the implementation of other approaches of detecting selection pressure could be considered, as our tests are only based on site frequency spectrum. A combination of all such tests could provide a convincing sign of a biological function on HERVs.

Bibliography

[1] E. Ehler, *Selection pressure on human genome detected in the proximity of forensic microsatellite markers*, 2017.

[2] N. Bannert and R. Kurth, "The evolutionary dynamics of human endogenous retroviral families", *Annu. Rev. Genomics Hum. Genet.*, vol. 7, pp. 149–173, 2006.

[3] N. Grandi and E. Tramontano, "Human endogenous retroviruses are ancient acquired elements still shaping innate immune responses", *Frontiers in immunology*, vol. 9, p. 2039, 2018.

[4] W. E. Johnson, "Endogenous retroviruses in the genomics era", *Annual review of virology*, vol. 2, pp. 135–159, 2015.

[5] S. Casillas, R. Mulet, P. Villegas-Mirón, S. Hervas, E. Sanz, D. Velasco, J. Bertranpetit, H. Laayouni, and A. Barbadilla, "PopHuman: the human population genomics browser", *Nucleic acids research*, vol. 46, no. D1, pp. D1003–D1010, 2018.

[6] T. A. Brown, *The Human Genome*, Jan. 1970. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK21134/.

[7] M. Jobling, E. Hollox, M. Hurles, T. Kivisild, and C. Tyler-Smith, "HUMAN EVOLUTIONARY GENETICS Second Edition", *HUMAN EVOLUTIONARY GENETICS, SECOND EDITION*, pp. 601–608, 2014.

[8] N. Saitou, *Introduction to evolutionary genomics*. Springer, 2013.

[9] C. O'Connor, *Meiosis, Genetic Recombination, and Sexual Reproduction*, 2008. [Online]. Available: https://www.nature.com/scitable/topicpage/meiosis-genetic-recombination-and-sexual-reproduction-210/.

[10] *What is a gene mutation and how do mutations occur? - Genetics Home Reference - NIH*. [Online]. Available: https://ghr.nlm.nih.gov/primer/mutationsanddisorders/genemutation.

[11] *Sickle cell disease - Genetics Home Reference - NIH*. [Online]. Available: https://ghr.nlm.nih.gov/condition/sickle-cell-disease.

[12] L. Luzzatto, "Sickle cell anaemia and malaria", *Mediterranean journal of hematology and infectious diseases*, vol. 4, no. 1, 2012.

[13] T. E. of Encyclopaedia Britannica, *Retrovirus*, Mar. 2019. [Online]. Available: https://www.britannica.com/science/retrovirus.

[14] J. M. Coffin, *The Place of Retroviruses in Biology*, Jan. 1997. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK19382/.

[15] W.-S. Hu and V. K. Pathak, "Design of Retroviral Vectors and Helper Cells for Gene Therapy", *Pharmacological Reviews*, vol. 52, no. 4, pp. 493–512, 2000, ISSN: 0031-6997. eprint: http://pharmrev.aspetjournals.org/content/52/4/493.full.pdf. [Online]. Available: http://pharmrev.aspetjournals.org/content/52/4/493.

[16] Y. Chen, J. Evans, and M. Feldlaufer, "Horizontal and vertical transmission of viruses in the honey bee, Apis mellifera", *Journal of invertebrate pathology*, vol. 92, no. 3, pp. 152–159, 2006.

[17] M. Wade, "Evolutionary Genetics", in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., Fall 2008, Metaphysics Research Lab, Stanford University, 2008.

[18] S. Wright, "Evolution in Mendelian populations", *Genetics*, vol. 16, no. 2, p. 97, 1931.

[19] T. E. of Encyclopaedia Britannica, *Genetic drift*, Jan. 2020. [Online]. Available: https://www.britannica.com/science/genetic-drift.

[20] C. Darwin and W. F. Bynum, *The origin of species by means of natural selection: or, the preservation of favored races in the struggle for life*. Penguin Harmondsworth, 2009.

[21] T. R. Gregory, "Understanding natural selection: essential concepts and common misconceptions", *Evolution: Education and outreach*, vol. 2, no. 2, p. 156, 2009.

[22] M. Pigliucci, "An extended synthesis for evolutionary biology", *Annals of the New York Academy of Sciences*, vol. 1168, no. 1, pp. 218–228, 2009.

[23] M. Kimura, *The neutral theory of molecular evolution*. Cambridge University Press, 1983.

[24] R. C. Lewontin *et al.*, *The genetic basis of evolutionary change*. Columbia University Press New York, 1974, vol. 560.

[25] R. Nielsen, "Molecular signatures of natural selection", *Annu. Rev. Genet.*, vol. 39, pp. 197–218, 2005.

[26] L. Loewe, *Negative selection*. [Online]. Available: https://www.nature.com/scitable/topicpage/negative-selection-1136/.

[27] J. M. Smith and J. Haigh, "The hitch-hiking effect of a favourable gene", *Genetics Research*, vol. 23, no. 1, pp. 23–35, 1974.

[28] B. Charlesworth, M. Morgan, and D. Charlesworth, "The effect of deleterious mutations on neutral molecular variation.", *Genetics*, vol. 134, no. 4, pp. 1289–1303, 1993.

[29] R. C. McCoy and J. M. Akey, "Selection plays the hand it was dealt: evidence that human adaptation commonly targets standing genetic variation", *Genome biology*, vol. 18, no. 1, p. 139, 2017.

[30] M. Kimura, "Evolutionary rate at the molecular level", *Nature*, vol. 217, no. 5129, pp. 624–626, 1968.

[31] R. Kliman, *Genetic Drift and Effective Population Size*. [Online]. Available: https://www.nature.com/scitable/topicpage/genetic-drift-and-effective-population-size-772523/.

[32] L. Duret, *Neutral Theory: The Null Hypothesis of Molecular Evolution*. [Online]. Available: https://www.nature.com/scitable/topicpage/neutral-theory-the-null-hypothesis-of-molecular-839/.

[33] M. Kimura, "The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations", *Genetics*, vol. 61, no. 4, p. 893, 1969.

[34] G. Watterson, "On the number of segregating sites in genetical models without recombination", *Theoretical population biology*, vol. 7, no. 2, pp. 256–276, 1975.

[35] F. Tajima, "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.", *Genetics*, vol. 123, no. 3, pp. 585–595, 1989.

[36] F. Tajima, "Evolutionary relationship of DNA sequences in finite populations", *Genetics*, vol. 105, no. 2, pp. 437–460, 1983.

[37] M. Pybus, G. M. Dall'Olio, P. Luisi, M. Uzkudun, A. Carreno-Torres, P. Pavlidis, H. Laayouni, J. Bertranpetit, and J. Engelken, "1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans", *Nucleic acids research*, vol. 42, no. D1, pp. D903–D909, 2014.

[38] Y.-X. Fu and W.-H. Li, "Statistical tests of neutrality of mutations.", *Genetics*, vol. 133, no. 3, pp. 693–709, 1993.

[39] B. Haubold and T. Wiehe, *Introduction to computational biology: an evolutionary approach*. Springer Science & Business Media, 2006.

[40] K. L. Simonsen, G. A. Churchill, and C. F. Aquadro, "Properties of statistical tests of neutrality for DNA polymorphism data.", *Genetics*, vol. 141, no. 1, pp. 413–429, 1995.

[41] Y.-X. Fu, "Statistical properties of segregating sites", *Theoretical population biology*, vol. 48, no. 2, pp. 172–197, 1995.

[42] J. C. Fay and C.-I. Wu, "Hitchhiking under positive Darwinian selection", *Genetics*, vol. 155, no. 3, pp. 1405–1413, 2000.

[43] K. Zeng, Y.-X. Fu, S. Shi, and C.-I. Wu, "Statistical tests for detecting positive selection by utilizing high-frequency variants", *Genetics*, vol. 174, no. 3, pp. 1431–1439, 2006.

[44] B. Walsh and M. Lynch, *Evolution and selection of quantitative traits*. Oxford University Press, 2018.

[45] 1. G. P. Consortium *et al.*, "A global reference for human genetic variation", *Nature*, vol. 526, no. 7571, pp. 68–74, 2015.

[46] 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, *et al.*, "The variant call format and VCFtools", *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, 2011.

[47] T. K. Oleksyk, V. Brukhin, and S. J. O'Brien, "The Genome Russia project: closing the largest remaining omission on the world Genome map", *GigaScience*, vol. 4, no. 1, s13742–015, 2015.

[48] B. Pfeifer, U. Wittelsbürger, S. E. Ramos-Onsins, and M. J. Lercher, "PopGenome: an efficient Swiss army knife for population genomic analyses in R", *Molecular biology and evolution*, vol. 31, no. 7, pp. 1929–1936, 2014.

[49] J. Pačes, A. Pavlíček, R. Zika, V. V. Kapitonov, J. Jurka, and V. Pačes, "HERVd: the human endogenous retroviruses database: update", *Nucleic acids research*, vol. 32, no. suppl_1, pp. D50–D50, 2004.

[50] M. Šatrová, "Detection of selection pressure on human endogenous retroviruses", University of Chemistry and Technology Prague, Intitute of Informatics and Chemistry, 2020.

Appendix A

Formulas for the neutrality tests

Here we include the formulas for variance calculation of the tests not mentioned in chapter 3. We also assume that n denotes the number of chromosomes and S denotes the number of segregating sites.

A.1 Fu & Li's tests

Fu & Li's tests are

$$D^* = \frac{\frac{S}{a_1} - \frac{n-1}{n}\eta_s}{\sqrt{u_{D*}S + v_{D*}S^2}}, \tag{IA.1}$$

$$F^* = \frac{\hat{\theta}_\pi - \frac{n-1}{n}\eta_s}{\sqrt{u_{F*}S + v_{F*}S^2}}, \tag{IA.2}$$

where

$$v_{D*} = \frac{\frac{a_2}{a_1^2} - \frac{2}{n}(1 + \frac{1}{a_1} - a_1 + \frac{a_1}{n}) - \frac{1}{n^2}}{a_1^2 + a_2}, \tag{IA.3}$$

$$u_{D*} = \frac{\frac{n-1}{n} - \frac{1}{a_1}}{a_1} - v_{D*}, \tag{IA.4}$$

$$v_{F*} = \frac{\frac{2n^3+110n^2-255n+153}{9n^2(n-1)} + \frac{2(n-1)a_1}{n^2} - \frac{8a_2}{n}}{a_1^2 + a_2}, \tag{IA.5}$$

$$u_{F*} = \frac{\frac{4n^2+19n+3-12(n+1)(a_1+\frac{1}{n})}{3n(n-1)}}{a_1} - v_{F*}, \tag{IA.6}$$

taken from [40].

A.2 normalized Fay & Wu's H

The normalized version is as follows:

$$H = \frac{\hat{\theta}_\pi - \hat{\theta}_L}{\sqrt{Var(\hat{\theta}_\pi - \hat{\theta}_L)}}, \tag{IA.7}$$

where

$$\begin{aligned} Var(\hat{\theta}_\pi - \hat{\theta}_L) = \frac{n-2}{6(n-1)}\hat{\theta}_W + \\ \frac{S(S-1)18n^2(3n+2)(a_2+\frac{1}{n^2})}{9n(n-1)^2(a_1^2+a_2)} - \\ \frac{S(S-1)(88n^3+9n^2-13n+6)}{9n(n-1)^2(a_1^2+a_2)} \end{aligned} \tag{IA.8}$$

A.3 Zeng et al.'s E

For Zeng et al.'s E test we have

$$E = \frac{\hat{\theta}_L - \hat{\theta}_W}{\sqrt{Var(\hat{\theta}_L - \hat{\theta}_W)}}, \tag{IA.9}$$

where

$$\begin{aligned} Var(\hat{\theta}_L - \hat{\theta}_W) = \left(\frac{n}{2(n-1)} - \frac{1}{a_1}\right)\hat{\theta}_W + \\ \left(\frac{a_2}{a_1^2} + 2\left(\frac{n}{n-1}\right)^2 a_2 - \frac{2(na_2 - n + 1)}{(n-1)a_1} - \frac{3n+1}{n-1}\right)\frac{S(S-1)}{a_1^2+a_2} \end{aligned} \tag{IA.10}$$

Appendix B

Comparison tables of Fu & Li's D* and Fu & Li's F*

Chr	Position start	Our	R PopGenome
22	30800000	-1.50836591443	-1.508366
10	66700000	-0.648409118634	-0.6027495
6	78390000	-0.025115547475	-0.02511555
12	60950000	-1.20950242353	-1.209502
3	96950000	-5.36817052126	-5.368171

Table B.1: Comparisons of Fu & Li's D* values.

Chr	Position start	Our	R PopGenome
22	30800000	-1.2683483372	-1.254678
10	66700000	-0.0391790868531	0.006035406
6	78390000	0.296512655405	0.3046096
12	60950000	-0.653563886317	-0.6410901
3	96950000	-5.07100680328	-5.04352

Table B.2: Comparisons of Fu & Li's F* values.