Visualization and Simulation of Variants in Personal Genomes With an Application to Premarital Testing (VSIM)

Thesis by
Azza Thamer Hussain Althagafi

In Partial Fulfillment of the Requirements
For the Degree of
Masters of Science

King Abdullah University of Science and Technology
Thuwal, Kingdom of Saudi Arabia

November, 2018
The thesis of Azza Thamer Althagafi is approved by the examination committee

Committee Chairperson: Robert Hoehndorf
Committee Members: Takashi Gojobori, Mikhail Moshkov
ABSTRACT

Visualization and Simulation of Variants in Personal Genomes With an Application to Premarital Testing (VSIM)

Azza Thamer Althagafi

Interpretation and simulation of the large-scale genomics data are very challenging, and currently, many web tools have been developed to analyze genomic variation which supports automated visualization of a variety of high throughput genomics data. We have developed VSIM an automated and easy to use web application for interpretation and visualization of a variety of genomics data, it identifies the candidate diseases variants by referencing to four databases Clinvar, GWAS, DIDA, and PharmGKB, and predicted the pathogenic variants. Moreover, it investigates the attitude towards premarital genetic screening by simulating a population of children and analyze the diseases they might be carrying, based on the genetic factors of their parents taking into consideration the recombination hotspots. VSIM supports output formats based on Ideograms that are easy to interpret and understand, which makes it a biologist-friendly powerful tool for data visualization, and interpretation of personal genomic data. Our results show that VSIM can efficiently identify the causative variants by referencing well-known databases for variants in whole genomes associated with different kind of diseases. Moreover, it can be used for premarital genetic screening by simulating a population of offspring and analyze the disorders they might be carrying. The output format provides a better understanding of such large genomics data. VSIM thus helps biologists and marriage counsellor to visualize a variety of genomic variants associated with diseases seamlessly.
ACKNOWLEDGEMENTS

"Great indeed has been Allah’s favor upon you” AlQuran, Surah An-Nisa [4:113]

First and foremost, I thank God Almighty for all the grace He has given me. The knowledge, strength, ability to complete this research work and for illuminating the road ahead. Without his blessings, this work would not have been possible.

Then, I would like to thank several people who helped and supported me in my journey towards this degree. Starting with my supervisor Prof Robert Hoehndorf, I would like to express my deepest and sincere gratitude to him for his patience, constant encouragement, inspiration, invaluable guidance and insightful comments. His valuable advice, guidance and careful editing continuously contributed to making this research better. I shall eternally be truly grateful to him for his support and assistance. Furthermore, many thanks go to the National Guards Hospital in Riyadh, for providing us with helpful resources, and cooperation. Special thanks go to Dr. Len Gribble for all the his helpful advice and for supporting me. I also would like to include deep thanks to Ms. Lucy Vick for her valuable information and for being helpful. My sincere thanks also goes to all BORG group members, for their cooperative spirit and the sincere efforts they exerted.

I would like to include a very special and unlimited thanks to my family, for boosting my morale and for their patience during the long hours of work, especially my lovely parents Hassna Althagafi and Thamer Althagafi. I would like to thank them for all their prayers, pushing me toward my dreams, believing in me, supporting and motivating me, and their pure and ultimate love.

Last but not least, I take immense pleasure in thanking all my friends at KAUST who encouraged, supported me, and kept cheering me to forge ahead and keep up the good work.
# TABLE OF CONTENTS

Examination Committee Page 2

Copyright 3

Abstract 4

Acknowledgements 5

List of Abbreviations 8

List of Figures 9

List of Tables 10

## 1 Introduction 11
  1.1 Motivation ................................................. 11
  1.2 Problem Statement and Objectives .......................... 13
  1.3 Thesis Organization ....................................... 14

## 2 Fundamentals 15
  2.1 Basic Genetic Terminology .................................. 15
  2.2 Genetic Inheritance ........................................ 17
    2.2.1 Mendel’s Laws ........................................ 17
    2.2.2 Genetic Diseases ...................................... 18
    2.2.3 Genetic Linkage, Crossing Over and Genetic Recombination 18
  2.3 Genetic Maps and Recombination Rates ...................... 19
  2.4 Genome Annotation of VCF files ........................... 20
    2.4.1 VCF Files Format and VCFtools ......................... 20

## 3 Literature Review 23
  3.1 Interpretation of Genomics Data ........................... 24
  3.2 Simulation Tools .......................................... 28
4 Materials and Methods

4.1 Databases for Variants .................................................... 31
4.2 Pathogenicity Prediction .................................................. 33
4.3 Annotating Variants ....................................................... 33
4.4 Simulation ................................................................. 34
4.5 Visualization .............................................................. 34
4.6 Implementation ............................................................ 35
  4.6.1 Variant Annotation .................................................. 35
  4.6.2 Simulation ........................................................... 37
4.7 Docker Container ......................................................... 40

5 Result

5.1 Annotating and Visualizing Personal Genomic Data .............. 41
5.2 Simulating Child Cohorts and Application to Premarital Testing . 43
5.3 Performance Evaluation .................................................. 44
5.4 Case Study ............................................................... 47
  5.4.1 Finding Causative Variants in Clinical Cases 1 ................. 47

6 Discussion

7 Concluding Remarks

7.1 Summary ................................................................. 52
7.2 Future Research Work .................................................. 53

References

Appendices 64

1This work is part of a collaborative project involving the research groups of Prof Hoehndorf and Prof Gojobori.
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>CADD</td>
<td>Combined Annotation Dependent Depletion</td>
</tr>
<tr>
<td>DIDA</td>
<td>Digenic diseases DAtabase Studies</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome-Wide Association Studies</td>
</tr>
<tr>
<td>HIV</td>
<td>Human Immunodeficiency Virus</td>
</tr>
<tr>
<td>HPO</td>
<td>Human Phenotype Ontology</td>
</tr>
<tr>
<td>LD</td>
<td>Linkage Disequilibrium</td>
</tr>
<tr>
<td>MCAP</td>
<td>Mendelian Clinically Applicable Pathogenicity</td>
</tr>
<tr>
<td>MOI</td>
<td>Mode Of Inheritance</td>
</tr>
<tr>
<td>PharmGKB</td>
<td>Pharmacogenomics Knowledgebase</td>
</tr>
<tr>
<td>RTG</td>
<td>Real Time Genomics</td>
</tr>
<tr>
<td>SNPs</td>
<td>Single Nucleotide Polymorphisms</td>
</tr>
<tr>
<td>VCF</td>
<td>Variant Call Format</td>
</tr>
<tr>
<td>WES</td>
<td>Whole exome sequencing</td>
</tr>
<tr>
<td>WGS</td>
<td>Whole genome sequencing</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

2.1 Genes and traits [1] ................................................................. 15
2.2 Genes, Alleles and Loci on Homologous Chromosomes[2] .......... 16
2.3 Example of VCF File [3] .......................................................... 22

4.1 High-level overview over VSIM work-flow ................................. 35
4.2 The workflow for analyzing genomic sequence data of individuals ... 36
4.3 The workflow for simulation ..................................................... 38

5.1 Visualization of individual genomes ........................................... 42
5.2 Example of Simulation result .................................................... 44
5.3 Simulation time ................................................................. 45
5.4 Linkage Disequilibrium Evaluation Result ................................ 46
5.5 GATK pipeline overview [4] .................................................... 47
5.6 GATK Best Practices for SNP and Indel discovery (variants) [4] ... 48
LIST OF TABLES

4.1 Used Databases ............................................. 33
Chapter 1

Introduction

1.1 Motivation

The motivation for this work is based on two factors. First, almost any human disorder has a genetic component in it. It may vary from 100 percent for monogenic disorders to a much smaller percentage for the complex diseases with multiple contributing factors. Understanding how variation in individuals’ genomes relate to disease risk is essential as it helps us with a better clinical diagnosis of the disease, prognosis of the disease development, new approaches for treatment and the development of new drugs. Second, the rate of consanguineous marriages is high especially within the countries of the Middle East. The rate of consanguineous marriages varies between regions, based on factors including religion, ethnicity, race, and socio-cultural factors [5, 6, 7, 8]. Recent studies have shown that the spread of consanguineous marriages in population is varied, between 33% to 68% in different countries of the Middle East [7]. It reaches up to 58% within Saudi Arabia [9], 50% in in Oman [10] and in the United Arab Emirates [11], and 68% in Egypt [12].

There are several papers that report the widespread effects and serious consequences of consanguinity marriage on human health. They mainly focused on its impact on the reproduction of the next generation, rare Mendelian disorders, and childhood mortality [7, 8, 13, 14, 15, 16, 17, 18, 19, 20, 21]. Nevertheless, there is a lack of awareness about consanguinity-related diseases that might be transmitted to the offspring [22]. This results in a high incidence of heredity diseases which are
causing great suffering for the next generation of children as well as pose an economic challenge for the healthcare system.

Genetics Screening test is recommended by government health administrations to inform potential parents about risk for the health of their children. Genetic screens aim to advise the entire population to have a pre-marital screening test before getting married. This test helps to identify the people that are carriers of one or more autosomal recessive disorders, or they have a genetic predisposition that as a result may engender disease in their future children [7, 8, 6].

For example, the “Healthy Marriage Program” is a program by the Saudi Premarital Screening and Genetic Counseling (PMSGC). This program is part of a national project led by the Saudi Ministry of Health administration [23]. A premarital certificate is required for any couples with marriage proposals [24], and screening includes tests for hemoglobinopathies, sickle cell anaemia, thalassemias, and infectious diseases such as Human Immunodeficiency Virus (HIV) [22]. While such programs address important health problems, they are limited to a few diseases and do not address the problem of heritable diseases [25, 23].

There is now a need to develop new ways that could help in better understanding of different variants in individuals and how they relate to diseases of them and their potential children. In the case of pre-marital testing, we want to know from an early stage what are the different diseases or symptom that the offspring might get, and how likely will they develop certain diseases, ideally covering a broad variety of genetically-based diseases. Visually exploring individual’s genomes, and identifying useful information visually, can contribute to interpretation of disease risk in individuals. Through simulation of potential child genomes, this visual exploration can also be extended to aid in pre-marital testing and genetic counseling with respect to the disease risk for heritable diseases.
1.2 Problem Statement and Objectives

We have developed VSIM as a tool for the visualization and interpretation of likely disease-causing variants in individual Whole genome sequencing (WGS) or Whole exome sequencing (WES) data. Given two genomic sequences (WGS or WES) as input, VSIM is further capable to simulate a cohort of child genomes, taking into account the recombination rates and probabilities, thereby accurately considering recombination hot- and cold-spots. We use the information about members of this cohort of child genomes to determine the probabilities that children of the two individuals from which the original genomic sequences were derived will develop a certain disease or phenotype. VSIM can therefore not only be used to interpret and visually explore individual genome sequences but also to contribute to premarital genetic testing. VSIM relies on information about Mendelian diseases from the ClinVar database [26], genetic associations for the complex diseases and risk factors from Genome-Wide Association Studies (GWAS) [27], digenic disease variants from the DIgenic diseases DAtabase Studies (DIDA) [28], as well as pharmacogenomic variants from the Pharmacogenomics Knowledgebase (PharmGKB) [29]. Optionally, VSIM can also include the Mendelian Clinically Applicable Pathogenicity (MCAP) Score [30]. The output of VSIM relies on Ideograms [31] that are easy to interpret and understand, and from which additional information about the variant and its likely phenotypic effect can be accessed. The simulation algorithm underlying VSIM can also be used independently to simulate evolution of a population on a genomic scale, and subsequently investigate co-morbidities of diseases and other traits. VSIM and the related source code is freely available on http://vsim.kaust.edu.sa and https://github.com/bio-ontology-research-group/VSIM.
1.3 Thesis Organization

The thesis is organized as follows: Chapter 2 we will introduce basic fundamental concepts to our work. We describe these concepts for a better understanding of our proposed method. Chapter 3 highlights some of the related work that forms the background for our method. Chapter 4 illustrates a brief overview of the dataset that was used in this thesis work with a general description of the method. This chapter also describes our implementation. Chapter 5 and 6 illustrates and analyses the results obtained by our method. Finally, Chapter 6 concludes our work with a summary of the work that has been accomplished and further discusses potential future work.
Chapter 2

Fundamentals

This chapter covers the fundamentals related to implementation concepts in our method including basic genetic terminology, genome annotation, and Variant Call Format (VCF) files structure. We describe these concepts as we used them in our work.

2.1 Basic Genetic Terminology

DNA is located inside the nuclei of eukaryotic cells and is organized into chromosomes. A chromosome contains a part of the DNA of an organism. DNA contains a sequence of nucleotides that encode instructions for how cells create proteins which then subsequently perform certain functions within the cell or organism, and ultimately lead to the expression of different types of traits and characteristics of the organism [32].

![Diagram of LOCi, Gene, Trait]

Figure 2.1: Genes and traits [1]
A locus is defined as a particular position on one chromosome. Some regions on a chromosome may encode for genes that are associated with particular traits. A gene is anything on our DNA that has a function and includes microRNAs or similar which are not coding for a protein. Genes are related to the express of physical traits of the organism. Figure 2.1 shows an example that illustrates the concepts.

In diploid organisms, chromosomes come in pairs of two, known as a homologous pairs of chromosomes, where one chromosome originates in a female parent and the other originates in the male parent [33].

The set of homologous pairs of chromosomes (see Figure 2.2), together with sex chromosomes, form the genome of an organism which carries the genetic information about the organism’s traits. Genes are usually present on both homologous chromosomes. For instance, if there is a particular gene in a chromosome that encodes for eye color, then its homologous chromosome will usually also contain a gene that is responsible for eye color. These pairs of genes together express the physical traits and they are known as alleles [32]. The alleles do not have to be identical, however: they may differ in single positions, larger regions, or one allele may be completely dysfunctional.

**Figure 2.2:** Genes, Alleles and Loci on Homologous Chromosomes[2]
2.2 Genetic Inheritance

Genetic inheritance is considered as a fundamental principle of genetics in living organisms. It demonstrates how physical traits are passed from one generation to the next. It happens as a consequence of the genetic material in the form of DNA that is passed down from both mother and father to their offspring.

The fundamental understanding of the inheritance process began with the work of the scientist Gregor Mendel. He defined the ‘Laws of Inheritance’ that together provide the foundations for modern genetics. In humans, the genetic material of the father and mother are combined and passed on to the child. That means the offspring will receive a combination of genetic information from the parents. However, there are specific genes from each parent that will dominate the expression of different traits in the offspring [34].

2.2.1 Mendel’s Laws

Gregor Mendel provided three essential conclusions about the principle of genetic inheritance. First, each trait is passed on to the offspring in a stable manner through the ‘units of inheritance’, which are now known as ‘alleles’. The second conclusion is that the offspring inherit one allele for a particular trait from each parent. The third conclusion is that some of the alleles might not be expressed in an individual, but rather, they can pass to the next generation of offspring [35]. These laws can now be expressed through processes of recombination of chromosomes and dominance of particular traits if two different alleles of a gene are present. Mendel’s laws apply to traits that are completely determined by a single gene or locus within a genome.

- The Law of Segregation declares that the two alleles for heritable characteristics or traits separate during gamete formation (meiosis). As a result, each gamete will have only one of the two alleles from parents for each gene [36].
• The Law of Independent assortment declares that a random distribution of alleles occurs. That means each pair of alleles for each trait separate independently from the second pair of alleles during gamete production [36].

• The Law of Dominance states that some alleles are dominant while others are recessive. Organisms with a dominant allele will express the trait encoded on this allele [36].

2.2.2 Genetic Diseases

A genetic disease is a condition that is caused by a change in an individual’s DNA or a particular genotype. One common cause of genetic disease is the loss or change of the function of a gene or allele. Genetic disorders are hereditary, that means they can be passed down from the parents to their children [37].

The inheritance of genetic diseases, and other genetically-determined traits, can follow different inheritance patterns [34]. For example, inheritance patterns for a genetically-determined trait can be classified as recessive, dominant, X-linked, or Y-linked, as well as several other types of inheritance for more complex diseases [34].

2.2.3 Genetic Linkage, Crossing Over and Genetic Recombination

Humans have over 20,000 genes and 22 pairs of chromosomes in addition to two sex chromosomes (X and Y). Several genes will be located on a particular chromosome, and some of these genes will be located closer to each other than others. This physical location can affect the inheritance of genes and the expression of traits. Alleles that are close to each other on one chromosome will tend to remain closer together and may be more likely to be inherited together as a pair [38] while genes that are far apart on the same chromosome or they are on different chromosomes tend to be inherited separately [39]. This phenomenon is known as genetic linkage.
So, when different genes are found on the same chromosome but far apart from each other, they function independently. This is due to the phenomenon of recombination or chromosomal crossover [39] in which homologous chromosomes exchange parts. If genes are nearby on the same chromosome, the cross-over process results in alleles that tend to “stick together” instead of being inherited independently. In that case, the genes are known as linked genes, and genetic linkage can be quantified by determining linkage disequilibrium.

Linkage Disequilibrium (LD) is a nonrandom association of different alleles at various sites. LD occurs when the genotypes at the two loci are not independent of each another. Moreover, LD between two alleles at different loci has been described in different ways [38]. However, all the definitions depend on finding the distinction between the frequencies of gametes that have the two alleles, A and B, at a pair of two different sites, \( p_{AB} \). The definitions also depend on the outcome of the frequencies of the two alleles \( p_A \) and \( p_B \) [38].

\[
D_{AB} = p_{AB} - p_A p_B
\]  

(2.1)

The \( D_{AB} \) represents the coefficient of LD. It describes a particular pair of two alleles (A and B).

### 2.3 Genetic Maps and Recombination Rates

A genetic map, also known as a linkage map, provides information about genomic coordinates of known genes relative to each other. It is derived from the recombination rates between the markers during the process of the crossover of homologous chromosomes. The higher the frequency rate of the recombination between any two genetic markers, the further apart they are from each other on the chromosome. Conversely, the lower the recombination frequency, the smaller the distance between them. So,
the recombination rates are a measure of genetic linkage, and they are used in the production of a genetic linkage map [40].

The sexual dimorphism within humans includes significant variations in the distribution of the recombination rate throughout the human genome and is different in female and male organisms. In the female organism, the crossover rates are reported to be approximately 1.6 fold greater than the recombination rate in male organisms [41, 42, 43]. Consequently, recombination rates are sex-specific [44, 45].

We rely on sex-specific genetic maps based on the crossovers from over 100,000 meioses that are collected from an existing pedigree research studies, with a total of 104,246 meioses (46,327 male and 57,919 female meioses) [46]. The merged dataset consists of 999,007 male and 2,338,628 female recombination events. These recombination events boundaries define a total of 833,754 Single Nucleotide Polymorphisms (SNPs) on the autosomal, and 18,039 intervals SNPs on the X chromosomes. Moreover, the genetic maps which they generated expose the different variations of recombination rates in both sexes throughout the genome at a different resolution.

### 2.4 Genome Annotation of VCF files

DNA or genome annotation is in general known as the process of identifying and labeling specific important regions of genomes such as the position of genes, and all the coding regions in the human genome [47]. Moreover, it aims to characterize the function of the genes. This annotation is performed as a repeated lookup of information in several databases that contain information on particular genomic regions, and there are several tools that can automate annotation.

#### 2.4.1 VCF Files Format and VCFtools

VCF is a generic file format developed for the 1000 Genomes Project and also adopted by many other projects. It stores individual genome sequences in a compressed for-
mat; using a reference genomes, the VCF format stores only the differences of a genome to a reference genome. These differences are variants with respect to an agreed-upon reference. In the VCF file format, DNA sequence variation can further be associated with rich annotations [3].

Several tools have been developed to generate and process VCF files. For example, VCFtools [3] is a set of software tools which implements several utilities to process VCF files, including validation, merging, and comparing.

A VCF file consists of a header line with meta-information such as the reference genome with respect to which sequence variant is recorded, and data lines represent the actual differences between a genome and the reference genome. Figure 2.3 provides an example of a VCF file.

For each record in the file, the information is structured into fields. The field lines name eight mandatory columns as follows:

- #CHROM: chromosome number ordered
- POS: The start position of the variants
- ID: A unique identifier of the variants
- REF: Reference alleles
- ALT: Alternate alleles
- QUAL: a Phred-scaled quality score of the variants
- FILTER: The site-filtering information
- INFO: Additional information regarding the variants.
Figure 2.3: Example of VCF File [3]

The sample-specific information, such as the genotype and the individual sample annotation values, are contained in the ninth column named FORMAT, and the sample IDs columns in the tenth and beyond. Additionally, VCF files can contain an INFO field in which annotations and additional information is located.
Chapter 3

Literature Review

There is a broad range of tools and methods related to analyzing or manipulating genomic information with the potential to help gain a better understanding, diagnoses, treatment and prevention of various kinds of diseases. Consequently, a growing number of studies and applications have been devised to better analyse and understand genomics data.

In this chapter, we will review previous studies and applications to shed light on the effectiveness of the interpretation and simulation of genomic data. We classify the chapter into two main categories: Interpretation of Genomics Data, and Simulation Tools. Each category includes a number of different methods that demonstrate an alternative method or cover important aspects related to our research work.

We first begin with a general overview of the interpretation of genomics data. This includes many of the common approaches used in the area of understanding the causative variants of the Mendelian diseases and complex diseases. Identifying the genes that are associated with those complicated diseases allows us to understand the disease more, and improve diagnostics and therapy. We then introduced a thorough study of several genomics simulation tools. This includes the various features used, such as the required input type, how the user can interact with it, sequencing platforms used, genomic variants simulation, and the output provided.
3.1 Interpretation of Genomics Data

Genetic screening in the clinical laboratories is becoming progressively more common. Many of them use targeted gene sequencing to recognize different variants in a set of genes that are responsible for a specific disease [48, 49]. The key challenging step in all these tests is their ability to find and interpret the genetic variants associated with the disease accurately, and how to assign the likelihood of pathogenicity of the variants [50]. The results of these tests are reported and well documented in a set of databases, and they are classified based on several factors such as their clinical significance, pathogenicity, and the number of genes associated with phenotypes.

Nowadays, with the decreasing cost of the genome sequencing, the laboratories are generating a massive amount of sequence data. Most small laboratories and researcher teams do not have a computing infrastructure, which is required to develop and maintain the genome annotation platforms that are geared toward the processing of high-throughput data [51]. As a consequence to circumvent this problem, different computational programs have been developed to perform automated genome annotations. However, for reliable results, human experts curation is often required.

Variant interpretation is defined as the process of finding straight associations from the variants of individuals to disease phenotypes. That process is considered as a fundamental in the reporting results of clinical and incidental findings, and is also considered as the research endeavours which involve the discovery of variants and finding the results. There are various computational methods which have evolved to assist with the interpretation of variants. Currently, most of the available genome annotation pipelines use existing prediction algorithms. One of the well-known prediction algorithms for variant annotation and scoring is Combined Annotation Dependent Depletion (CADD) [52], which aims to score the deleteriousness of the variants (SNPs) and insertion or deletion variants in the genomes of humans. It combines various annotations into one metric that correlates with allelic diversity. The frame-
work of CADD is based on a machine-learning algorithm (using a Support Vector Machine). The algorithm is trained on 14.7 million simulated variants to differentiate 14.7 million high-frequency human alleles [52].

There are other prediction algorithms useful for variant annotation and scoring variants to support discovery power. Some of these tools are PolyPhen [53], SIFT [54] and GERP [55]. Nevertheless, the current methods suffer from some limitations that restrict their usefulness. First, each annotation method has its specific metric which is rarely comparable. As a result, it is difficult to decide the relative influence of the distinct variant annotations or develop strategies on how to combine them.

Second, the annotation methods have been trained on well-known pathogenic mutations. Those mutations may not be generalizable to a different type of mutation, since they can be subject to ascertainment biases. Third, some studies have shown those broadly used approaches to mis-classify 26–38% known pathogenic mutations [52]. This might cause missing diagnoses of a diseases if the classifiers are used as a reliable tool in a clinical environment. As a consequence of that, the clinicians could not entirely rely on the classification of variants by those pathogenicities likelihood scores. Therefore, these scores are used as only one of several factors in describing a genetic variant and their role in determining a phenotype [50].

In our work, we rely on the MCAP pathogenicity probability score. This score intends to misclassify no more than 5% of pathogenic variants. MCAP can be interpreted with high confidence meaning that if the variant is classified as benign, then it can usually be trusted to be benign. The MCAP implementation uses gradient-boosting trees, which is a supervised machine-learning classifier that excels at analyzing the non-linear interactions between different features. Moreover, it has outperformed existing approaches’ performance in different classification tasks [56, 57]. The MCAP classification features rely on the current pathogenicity likelihood scores (like PolyPhen [53], SIFT [54], etc.) and direct measures of evolutionary conservation.
They provide a novel method of combining the amino acid conservation features with gradient-boosting trees, that can be used with any variant training set. Also, the calculated scores trained on mutations associated with Mendelian diseases which can be instantly utilized by clinicians to interpret and analyze the variants of undetermined significance.

Other studies from different perspectives try to associate genotypes with disease traits, based on machine-learning algorithms, and use genome sequencing to associate genotypes with physical and disease traits. Many models exist for predicting the traits of individuals such as eye color [58], skin color [55, 58, 59, 60], and facial structure [61]. In one study [62] using whole-genome sequencing data, it was shown to be possible to identify individuals by predicting their traits. They assess the benefit of the prediction of different phenotypes for matching phenotypic data to individual-level genotype data. They built a machine-learning model to predict 3D-facial structure, weight, height, body mass index (BMI), skin and eye color. However, for the complex traits, their models demonstrated only small parts of the phenotypic variation.

There are many other computational tools used for prioritizing the potential causative variants, specifically for Mendelian diseases. However, in spite of the DNA sequencing power toward genetic discovery [63] and many computational tools being available, after sequencing affected families, less than 50% of Mendelian disorders are resolved [64]. This is due to the emerging of new phenotypes of diseases, new genes for common disorders, and several unknown disease-causing variants which still need to be discovered [64].

Despite variant prioritization being fundamental for discovering Mendelian diseases, it is just part of more significant causes that involve disease gene prioritization [46]. Disease gene prioritization tools adopt more data to distinguish and prioritize the likely damaged genes associated with a phenotype, using, for example, genotype frequencies, variant allele frequencies, the model of inheritance, patient phenotypes,
and family histories. As opposed to variant prioritization, this is merely identifying the possibly damaging variants. In fact, this is considered as a fundamental difference between both approaches in terms of the underlying algorithms used for both tools. There are several gene prioritization tools and methods [65, 66, 67, 68, 69, 70].

On the other hand, the manner in which these analysis tool results are presented to the users is changing. While the majority of disease gene prioritization or variant prioritization methods were developed by bioinformatics scientists and were hardly accessible to clinicians directly, there is now a significant move towards developing decision support frameworks for variant and disease interpretation that clinicians can directly access, and include web-based or graphical user interfaces as well as supporting information that can be used to justify decisions and provide additional support [71, 72, 73, 74]. We consider these systems, usually using web-based interfaces, as a significant move forward from simply command-line-based analyses results as they show that the predictive methods are now of direct clinical relevance.

Furthermore, the variant and gene prioritization scores in these interactive environments, are only one component of a dynamic approach that aims to diagnose and use population-scale variation sources. A leading examples of these powerful tools are the genome Aggregation Database (gnomAD) [75], the Exome Aggregation Consortium (ExAC) [75], the 1000 Genomes Project [76], disease genotype-phenotype association databases such as ClinVar [26], and the Online Mendelian Inheritance in Man (OMIM) database [77].

Although we have all of these modern advanced tools, finding the disease causation variants remains an ambiguous process. The key is that a variant that affects a gene does not necessarily affect the health of the individual [46]. We need to understand the cascading steps underlying the discovery of the potential variant or gene associated with the diseases, and how the scores combine with other information like the different phenotypes and the family history. Additionally, our study also needs to consider how
they are decided to be significant medically and the evidence that is required for this determination [74].

3.2 Simulation Tools

We rely on genome simulation methods that can simulate genomes of children given the genomes of two parents, using a set of evolutionary rules on how genomes combine in children. Simulation of genomes has been employed for decades, and now mainly focuses on the simulation of genomes in the context of next-generation sequencing [78, 79]. Some of the simulation tools can be used for evaluating software performance [80]; others can be used as a guide for developing new computational tools [81]; and others use machine learning models [82, 83, 84] to study different biological populations, in [82] they study the effect of the crossing and the different mutation types and how this will affect the selective adaptation.

The current computational simulation tools vary in multiple aspects, for example the sequencing technology they use for their underlying representation, their input and output requirements and formats.

There are other studies which focus on the biological evaluation simulation studies, which have an essential impact on greater understanding of the influence of different demographic and evolutionary and situations on sequence variation. Also, they allow researchers to evaluate and create analytical methods in the study of disease-associated genetic factors.

Currently, there are three approaches to simulation genetic data: forward-time, backward-time, and re-sampling [85]. The forward-time approach idea starts from an initial group of individuals and traces its evolution within different genetic models with many generations, and the new individuals are picked from the last generations. The backward-time approach is designed to start from the current observed sample in the current generation and work backwards, i.e., it follows all alleles to a par-
ticular ancestor, and then operates ahead until the current generation, introducing
genetic information like mutations into the created genealogy [86]. The \textit{resampling approach} works with an existing an genomic data set. A standard version of this approach creates individual samples by re-sampling from the existing data like HapMap [87] and 1000 genome project data [76].

Our work falls into the first category which is the \textit{forward-time approach}. In the following, we describe the principles that it follows, discussing the primary features of this simulator type.

The forward-time approach works with an initial population and after that, follows its evolution over some generations under different evolutionary situations to create a new generation. This approach achieves the last population once it accomplishes the stopping criterion (for example, a certain number of generations). With this strategy of simulation, the whole ancestral genomics information can be tracked. Also, the properties of population can be analysed at any generation. Moreover, under this framework, the recombination happens between a pair of parental individuals within the current generation in order to generate the next generation of children. The probability of recombination between any two adjacent markers can be designed to be non-uniform or uniform over the chromosome, and for measuring the probabilities between any two markers, Haldane’s [88] mapping functions are usually used. The number of recombination events occurs within a pair of chromosomes commonly follows a Poisson distribution.

The most typical forward-time simulator is simuPOP [89], which is flexible in simulating different chromosome types, such as autosome, chromosome X and Y. Also, it can generate different kinds of samples like case-control, pedigree samples. However, when modeling genetic diseases, simuPOP cannot handle the linked disease loci. Hoggart et al. [90] address this problem by developing a new algorithm, named FREGENE, for managing the impacts of different interactions among multiple sites.
Nevertheless, other works have been designed to address various issues regarding computational time and dealing with the environmental effects. Some of these works are GenomePop [91] that adopts some techniques for improving simulation efficiency, and Nemo [92], GenomeSIMLA [93], and ForSim [94] which incorporate joint effects of genetic and environmental factors into the simulation model.

Concerning the complexity of computational work, the resampling approach is usually considered to be faster than forward and backwards methods. That is essentially because the resampling does not require the modeling of a complex evolutionary process, which is often computationally heavy as it requires more parameters and different cases to consider. Thus, most forward-time simulators are computationally heavier; this is because these kind of simulations track complete ancestral information while the backwards approach maintains only partial information [85].

Overall, the three simulation methods have distinct advantages and weaknesses. However, it seems that there are various coming challenges for further advancing the simulation of population genetics. Some of these challenges are as follows: First, the real population evolution is a remarkably complex process which itself is not sufficiently identified; moreover, the evolution of genomic sequences requires both genetic constraints and complicated environmental factors, that can only be approximated and it has not been considered in many of the simulation algorithms. Finally, improvement in memory usage and computational efficiency should be a focus of future work.
Chapter 4

Materials and Methods

This chapter details the methods and materials we used in our work. Throughout the chapter, detailed analytical methods are described under relevant sub-headings. All the software tools that we used are open-source and freely available for download and usage. We developed Custom Shell scripts for the implementation of analytical methods where no existing software was available. We performed all the statistical analysis either using stand-alone Python and Java scripts, or by the integration Python scripts into Shell code. We used PHP, HTML, CSS, jQuery and JavaScript for web-application development. We have provided a list of the software that we applied in this research in Appendix A on page 65.

4.1 Databases for Variants

We used four databases (see Table 4.1) for annotation. The following is a general overview of each of them.

- ClinVar [26] is a database of genomic variants and the interpretation of their relevance to diseases. It identifies the relationships among medically important variants and phenotypes. The variations contained in this database are in VCF format, and ClinVar contains a mixture of variations asserted to be pathogenic as well as those known to be non-pathogenic, with regards to their clinical significance. However, our work focused on the pathogenic and likely pathogenic variants. Therefore, as a result of this restriction, we obtained 84,536 variants
out of 396,647 SNPs.

- Genome-Wide Association Studies (GWAS) [27] is a statistical method that considers the associations between SNPs and genetic factors across the whole genome to correlate with particular traits or disorders, and it can find these associations for many traits simultaneously. This takes full advantage of all the SNPs and uses them as sign posts for different phenotypes or traits. The GWAS Catalog [95] now contains over 2,500 unique SNP–trait associations, i.e., associations between single nucleotide variants and phenotypes or diseases. In the GWAS Catalog, we find information about variants (in particular their genomic position) and an association with (usually) complex diseases (a complex disease is a disease that is multi-factorial and may, for example, be associated with many variants, each of which modify the disease. We used 69,460 variants from the GWAS Catalog in our work.

- DIDA [28] is a database that provides comprehensive information on the genes and associated genetic variants which are associated with digenic diseases (a disease follows digenic inheritance if particular genotypes in exactly two genes explain the disease or phenotype in a patient [96]). DIDA includes 213 digenic combinations which are composed of 364 distinct variants, that are involved in 44 digenic diseases [28].

- PharmGKB [29] investigates the association of genetic variation and drugs efficiency. PharmGKB contains pharmacogenetic information related to 3,070 variants. From PharmGKB, variants can be associated with different drug responses.
### Table 4.1: Used Databases

<table>
<thead>
<tr>
<th>Databases</th>
<th>Purpose</th>
<th>Source</th>
<th>Number of Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClinVar</td>
<td>Mendelian Diseases</td>
<td>NCBI</td>
<td>84,536</td>
</tr>
<tr>
<td>GWAS Catalog</td>
<td>Complex Diseases</td>
<td>EBI</td>
<td>69,460</td>
</tr>
<tr>
<td>PharmGKB</td>
<td>Drug response</td>
<td>Pharmgkb.org</td>
<td>3,070</td>
</tr>
<tr>
<td>DIDA</td>
<td>Digenic diseases</td>
<td>DIDA</td>
<td>364</td>
</tr>
</tbody>
</table>

#### 4.2 Pathogenicity Prediction

A pathogenic variant is a genetic variant that enhances an individual’s likelihood to develop a particular disorder. The development of the disease symptoms is more likely to appear in the individual when such a variant (or mutation) is inherited. MCAP [30] is a pathogenicity predictor for the rare missense variants within the human genome. It is tuned to the high sensibility that is required in a clinical context. Moreover, it combines pathogenicity scores of several other tools (including Polyphen-2 [53], SIFT [54] and CADD [52]) within novel machine learning models and additional features. We used MCAP to predict pathogenicity in all the variants in a VCF file. For that we used ANNOVAR database [97], which is an efficient software tool which uses the up-to-date information in order to annotate genetic variants identified from diverse genomes.

#### 4.3 Annotating Variants

The annotation algorithm for our tool VSIM uses VCF files directly. The VCF file must include at least the following fields for each variant: chromosome number (#CHROM), Position (POS), Reference Alleles (REF) and Alternate Alleles (ALT), and information (INFO). The variants are annotated with information related to the databases which are described in more detail above.
4.4 Simulation

We implemented the simulation based on the Real Time Genomics (RTG) simulation tool [98]. RTG provides a blueprint platform for genomic analysis. RTG tools is delivered as an executable with multiple commands, executed through a command line interface. RTG, among others, supports the generation of child genomes from two VCF files that represent parents, and contains parameters that allow the specification of the number of recombinations per chromosome as well as the addition of random novel mutations in children. However, RTG’s simulation algorithm for child genomes only supports completely random recombinations of VCF files, while it is well known that recombination does not occur randomly in living organisms [99].

4.5 Visualization

Interpretation of large-scale genomics data is complex, and it is challenging to make information in genomic data accessible. One way to make complex information accessible is through visualization in which certain types of information are organized and structured visually. Genomes and chromosomes are commonly represented visually through the use of “ideograms”, i.e., a schematic representation of chromosomes, which is used to show the relative size of the chromosomes and their characteristic banding patterns.

We use the Ideogram.js [100] JavaScript library for visualizing chromosomes, and we overlay the visual representation of each chromosome with the information obtained from annotating variants in a VCF file.

After obtaining the information related to all the chromosomes’ positions, the next step is to parse the genomic features (chromosome, annotation, start and stop) from a generic feature format (GFF) file which is in the NCBI Homo Sapiens Annotation Release. This data is consolidated and formatted into a unified JSON structure file.
Then, this file, e.g. ID325476.json, represents the final output of the visualization viewer data. Additionally, it contains all the data that is used by the client side in Ideogram.js.

4.6 Implementation

The implementation is divided into two parts, the first one is variant annotation of personal genomics data, and the second part is the simulation. Figure 4.1 provides an overview of our overall workflow. In the following, we describe the principles that we follow in both parts.

4.6.1 Variant Annotation

For testing, we analyzed sequence data of the human genomes collected from 1000 Genomes project [101]. The variants are represented in VCF files (using the VCFv4.0 format) [3]. The VCF files have all the individual’s information for each chromosome. By using Jvarkit [102] we extracted individual VCF files from the main VCF file for all the chromosomes. We then concatenated the chromosomes list for each individual using VCFtool [3] to generate a single VCF file containing all the information of one individual. Figure 4.2 summarizes the algorithm that we followed.
We used the four databases for annotation (dated 30 Aug, 2018), using the reference genome of “GRCh37” as the main genomic variant set. However, since some of the databases we use (ClinVar, GWAS, and DIDA) update regularly, we identify variants from these databases remotely for the annotation.

For identifying the candidate disease variants with ClinVar database, we downloaded Mode Of Inheritance (MOI) for diseases that are included in OMIM from the Human Phenotype Ontology (HPO) phenotype database. As a result, we obtained a total of 6,843 MOI records that were classified ‘Recessive’, ‘Dominant’ and ‘Others’. For ClinVar, we annotated the variants with at least one disease from OMIM, with pathogenic or likely pathogenic clinical significance.

Moreover, along with MOI, we used zygosity of the variant as a guide to decide if the person will get a certain disease. Zygosity information is not provided in ClinVar, but rather in the given VCF file of the individual [3] and it is represented in the genotype (GT) field. For example, in the VCF file a heterozygous variant will have a genotype value 0|1, while the homozygous variant will have 1|1. The pathogenic variant disease was associated with a variant based on genotype information and MOI. For instance, for a specific variant, if the disease mode of inheritance is recessive,
and the zygosity of the variant in the VCF file is 0|1, then this person will carry the diseases associated with that variant. And if the disease mode of inheritance is dominant, and the zygosity of the variant in the VCF file is 0|1 or 1|1, then this person is likely to develop the diseases associated with that variant.

As an example, the pathogenic variant rs1801265 in DPYD is associated with (OMIM:274270), and this OMIM is recessive. If the genotype with this VCF file matches with the position of the rs1801265 and has the genotype 0|1 or 1|1, then the person will carry this disease. On the other hand, the pathogenic variant rs3214759 in CRYGB is associated with (OMIM:615188) and this OMIM is dominant. Then, if the genotype with this VCF file matches with the position of the rs1801265 and has the genotype 1|1, then the person is likely to develop this disease.

However, ClinVar is the only database that uses the zygosity information, so, the information from the remaining databases was retrieved by finding the match with the chromosome, position, REF alleles and ALT alleles. We further assigned an M-CAP score [30] to each variant in the input VCF file and filtered the most likely pathogenic variants.

4.6.2 Simulation

We used the RTG tool for generating simulated children. However, since the RTG tool does not consider the varying recombination rates at different chromosomal positions, we implemented our own simulation algorithm on top of the RTG tools. Figure 4.3 summarizes the algorithm that we followed.

We used a set of precomputed recombination rate maps for human genome build 37 [40]. Since we need to find the recombination probability that helps with deciding the number of recombinations per chromosome, we converted the recombination rate Map (with recombination rate measured in cM) to recombination probability using formula 4.1 [103]. After obtaining recombination probabilities for each chromosomal
position in the human genome, we normalized the value and calculated the Cumulative Distribution Function (CDF), then we randomly drew from the probability distribution \( n \) and \( m \) times (\( n \) and \( m \) are parameters that determine the number of cross-overs per chromosome in males and females respectively).

\[
d = 50 \ln \left( \frac{1}{1 - 2 \Pr[\text{recombination}]} \right)
\]

\[
\Pr[\text{recombination}|\text{linkage of } d \text{ cM}] = \sum_{k=0}^{\infty} \Pr[2k + 1 \text{ crossovers}|\text{linkage of } d \text{ cM}] \quad (4.1)
\]

\[
e^{-d/100} \frac{(d/100)^{2k+1}}{(2k+1)!} = e^{-d/100} \sinh(d/100) = \frac{1 - e^{-2d/100}}{2}
\]

Where ‘\( \sinh \)’ is the ‘hyperbolic sine function’. The recombination probability is roughly “\( d \)/100 for small values of “\( d \)” and approaches 50% as “\( d \)” goes to infinity [103].

In order to run, the simulator requires two VCF files as an input (representing the mother and father genotype information). Then, the algorithm combines them into a single VCF file. After the combination of VCF files, we generated a population of
simulated children (the default number is 100) using our simulation algorithm that accounts for the recombination probabilities.

**Algorithm 1: Simulations Algorithm**

**Input**: - Two VCF files represent Father and Mother genomics data.  
  - Real genetic maps for male and female with CDF.  
**Output**: Summary statistics of different diseases and symptoms might be associated with the offspring.  
**START**: Merge the two input file using VCFtool  
  - \((N, M) \leftarrow\) crossover locations (Mother and Father crossover)  
  // chosen randomly according to a CDF.  
  1 for \(i := 1 \rightarrow \text{NumberOfChildren}\) do  
  2    for \(i := 1 \rightarrow \text{NumberOfVariants}\) do  
  3      if \(\text{VariantsPosition}(N||M)\) then  
  4        Perform the crossover in the recombination positions;  
  5      else  
  6        Go to the next position;  
  7        create : Child;  
  8    - Matches with ClinVar DB for Mendelian Diseases Variants  
  9    - Matches with GWAS DB for Complex Disease Variants  
 10    - Matches with DIDA DB for Digenic Disease Variants  
 11    - Matches with PharmGKB DB for Drug Response  
 12    Result : \((\text{VCFfile})\) \(\leftarrow\) Combine all the results and count how many children are affected with specific disease variants  
 13    return \((\text{VCFfile with Summary Statistics});\)

For premarital testing, we can then repeat the step of annotating variants in the resulting simulated children, as before, and generate summary statistics of how many individuals in the simulated cohort are carrying certain disease-associated variants. These summary statistics (and individuals within the simulated cohort) can then be visualized similarly to our visualization of individual VCF files. Algorithm 1 illustrates the procedure that we followed for the simulation.
4.7 Docker Container

For further availability of the work and to overcome the problem with different operating system requirements to install dependencies, we use containerization, specifically a Docker container [104]. Docker is an open-source project aimed to automate the deployment of different software inside containers. The containers serve as an additional layer of abstraction from the environment in which the application runs, and it provides automation of OS-level virtualization on Linux.

This containerization and decoupling from operating system resources allows the VSIM tool web service to be deployed easily and consistently, regardless of the target environment. As a consequence, our tool can be used within protected environments like hospitals, which can then easily use it with their patients.
Chapter 5

Result

VSIM implements a web-based simulation and visualization tool and runs on Apache web server. The communication between the client-side layer and the server-side layer takes place based on JavaScript. It aims to support genetic counseling and interpretation of genomic sequences. VSIM performs two main operations: first, VSIM is able to annotate and visualize personal genomes available in the VCF file format [3] in order to support visual exploration of variants and other genomic aberrations that may have an impact on health. Second, given two VCF files for two potential parents, VSIM can simulate a population of children, based on accurately accounting for recombination probabilities across the human genome, and then allows visual exploration of the simulation results. One of the main applications of the second feature of VSIM is genetic counseling and premarital genetic testing.

5.1 Annotating and Visualizing Personal Genomic Data

VSIM accepts a VCF file as input, annotates the variants in the VCF file, and then visualizes the results on a chromosomal ideogram.

Annotation of variants falls into five categories: known Mendelian disease variants (using the information from the ClinVar database [26]); disease-associated variants derived from GWAS studies (using the information from the GWAS Catalog [27]); variant combinations in digenic disease (using the information provided by the DIDA database [28]); pharmacogenomic variants (from the PharmGKB database [29]); and
predicted pathogenic variants (using the M-CAP pathogenicity score [30]).

VSIM then generates chromosomal views based on chromosomal ideograms and shows the chromosomal positions at which a functional variant has been found. Different categories of variants are shown in different colors, and it is possible to filter variants by their type (whether they are Mendelian disease variants, pharmacogenomic variants, etc.). Users are able to obtain additional information about variants when selecting a single variant, and can follow a hyper-link to a website with additional information and evidence about the type of variant. Figure 5.1 provides an example of the output.
5.2 Simulating Child Cohorts and Application to Premarital Testing

VSIM is further capable of simulating cohorts of potential child genomes when given two VCF files as input, and then using this simulated cohort to estimate the probability of encountering particular genetically based diseases in potential children (as well as the co-morbidities between the diseases). For this purpose, VSIM uses a map of genome-wide recombination rates for the human genome [40] which provides a global (i.e., not population-specific) estimate of recombination rates, distinguished by male and female genomes. The recombination rate is derived from 3.3 million crossovers from 104,246 meioses (57,919 female and 46,327 male meioses) [40].

Using the two input VCF files, the recombination rates and a parameter that determines the number of cross-overs per chromosome, VSIM simulates a population of potential children while considering the recombination probabilities; therefore, the population of children will account for, at least partially, linkage disequilibrium and the resulting correlation between risk-conveying or causative genomic positions. We annotate all genomes in the simulated cohort of children using the same annotation procedure and annotation sources used by VSIM, and we use the percentage of children within the population that carry a particular functional variant, to estimate the likelihood that children develop a particular disease. While the likelihood could be estimated directly using Mendel’s laws from the two parent genomes, our simulation approach will give more accurate probabilities in the case of complex, digenic, and oligogenic diseases, and further allows the estimation of co-morbidities in the child population. Figure 5.2 provides an example of the simulation result.
5.3 Performance Evaluation

The time it takes to analyze (i.e., annotate and visualize) a single whole genome depends on the size of the VCF file. For a VCF file with 3 million SNPs, it takes approximately 10 minutes to generate the final output. The most time consuming part is annotating the original VCF file with the MCAP prediction score. VSIM annotates a single variant on average in $1.4 \times 10^{-4}$ seconds. The experiments were run on a computer with an Intel Core i7 processor running at 2.5 GHz using 16 GB of RAM, running in OS X version 10.11.6.

When applying VSIM for determining the likelihood of children having or carrying a particular disease, the simulation time not only depends on the size of the VCF file but also depends on the number of simulated children. Figure 5.3 shows the performance benchmarks for a different number of simulated children. The time linearly increases with the number of simulations to perform, and the generation of simulated genomes can easily be parallelized.

We follow another procedure to evaluate our algorithm. Starting with a randomly generated population, Algorithm 2 illustrates the procedure that we follow to deter-
Algorithm 2: Simulation Evaluation Algorithm

Input : Population (VCF files) generated randomly from 1000 Genomes Project.
Output: Pearson correlation score for each generation.
Initialize: new population with 0

for $i := 1 \rightarrow \text{size}(\text{population})$ do
  mother ← random-select(\text{population}) ;
  father ← random-select(\text{population}) ;
  child = crossover(mother, father) ;
  mutate (child) with a small random probability ;
  add (child) to new-population ;
  $list_1$ ← Calculate LD of the current population ;
  $list_2$ ← Find the matched LD value with the real population ;
  (Correlation, generation) ← Calculate the Personal correlation between the
  the two list, $\text{pearsonr}(list_1, list_2)$
  population = new population ;
return (Correlation, generation) ;

Figure 5.3: Simulation time
mine how quickly our algorithm is able to generate a Linkage Disequilibrium (LD) that is similar to the one we find in a human population. We generate a random population, and starting from this population we simulate a population of children (i.e., we move one generation forward). The algorithm stops after a certain number of generations. We then measure LD in the resulting population. In the initial population consisting of completely random genomes there is no LD present; our algorithm introduces LD due to non-random recombination rates, so that after several generations the LD should approximate the LD found in a human population.

For measuring the LD between sites, we used VCFtools [3], with –hap-r2 argument. Since the program must perform pairwise site comparisons, and we have more than 30 million sites, this analysis can be time-consuming, so we reduced the number of comparisons by using a window within 50,000 base pairs of one another. Figure 5.4 shows the correlation value for the first seven generations. The correlation increases as we move from one generation to the next, and the correlation with human LD emerges after only a few generations, and since it is a stochastic process it is bound to fluctuate a bit.

![Figure 5.4: Linkage Disequilibrium Evaluation Result](image)
5.4 Case Study

5.4.1 Finding Causative Variants in Clinical Cases

In the scope of genome sequence analysis, we focus on a total of 106 samples belonging to 19 Saudi families, provided by the National Guards Hospital in Riyadh, Kingdom of Saudi Arabia. In these families, we often observe congenital marriage(s) with at least one affected child and phenotypes of unknown rare genetic disorder(s). Those samples have been sequenced in the KAUST core laboratories, applying WGS with 30X% coverage on an Illumina Hiseq 4000 next generation sequencing machine.

Our approach to finding the potential causative variants for a given patient is as follows:

- First, we used Genome Analysis Toolkit (GATK) [4] to find the variants in the genome.

The GATK mostly focused on variants discovery. The overall pipeline as we can see in figure 5.5 covers the method that we need to follow to go from raw sequence all the way to having VCF files with variant calls. They provide GATK best practices (Figure 5.6), that show detailed recommendations that they suggest is the best way to run the analysis. In general, there are three main phases. First, the pre-processing and cleanup of the raw data in order to be acceptable for the variant analysis. Second, variant discovery, which is the step where we distinguish noise from the actual variation. The third panel

---

1 This work is part of a collaborative project involving the research groups of Prof Hoehndorf and Prof Gojobori.
Figure 5.6: GATK Best Practices for SNP and Indel discovery (variants) [4]

provides an evaluation to refine the variant calls and the genotypes, and decide which part we need to keep and which part we need to filter out.

- Second, we use ANNOVAR [97] to annotate and identify whether these variants belong to coding or non-coding regions and assign their MAF values based on the data from 1000GP [76], gnomAD [75] and ExAC [75] databases and filter out the variants having a MAF value >0.01 (we are interested in the rare variants).

- Third, we utilize the pedigree information provided by the hospital and filter out the variants which appear in the unaffected family members. This step is rather challenging since each family is actually a different case; that requires filter tuning for the possible mode of inheritance scenarios, given the pedigree information. For example, in the case where both parents do not have any phenotypes but their children may have the same or different phenotypes, we should consider a recessive case instead. For a family where only one of the parents has phenotypes and they have an affected child, then the disease could be either dominant or recessive.

- After that, we employ PVP, which is a high-performing variant prioritization
tool, [105], utilizing the phenotype information to prioritize the variants of the patients.

- Finally, our clinician interprets the potential variant list from PVP, based on her background knowledge and the existing knowledge in the ontologies/databases/literature, to identify the causative variants. Furthermore, the clinician will also be able to visualize the PVP output using VSIM in the future.

Due to ethical issues $^2$, we cannot share the variants or analysis results of these patients in this thesis.

---

$^2$This work has been reviewed and approved by the KAUST Institutional Bioethics Committee under reference number 17IBEC08
Chapter 6

Discussion

With the new advances in DNA sequencing, we now have access to a large amount of information about DNA and the potential effects of genomic variants. This information can be used to identify the causes of particular phenotypes, for example, variants that are present in tumor cells but are absent in the normal cells, including mutation and changes in the gene expression. Moreover, it can then be used to identify the mutations associated with different diseases, discover new molecular targets for therapy, and to identify molecular profiles that are associated with a particular clinical outcome. However, this information can be hidden in the amount of information resulting from next generation sequencing and difficult to access. Identification and visualization of specific variants that could possibly be the cause of a disease will facilitate the understanding of the personal genomes and make a selected amount of information accessible to clinicians and genetic counselors.

We developed VSIM in the form of a web application tool that aims to provide an interpretation of the individual genomes. As an extended application of that, it can be used for interpretation of genetic testing results for couples considering marriage. VSIM can help to predict, and provide a general overview of, the potential diseases that might be associated with their children. To achieve this goal, VSIM simulates a population of children, and then provides statistics relevant to their genetic information. This information is represented in the form of chromosomal views.

The main limitation of our work, however, is the identification of variants based on the limited number of databases. Moreover, the sex linkage (X-Linked for female,
and Y-Linked for male) diseases, that are passed down through families through one of the X or Y chromosomes, are not considered. For future work, we need to integrate additional information regarding a different set of diseases and consider uncommon types of inheritance in humans.

Using VSIM, we can visually analyse the variants associated with personal genome data in a reasonable amount of time. VSIM can also be used as premarital testing, as it provides a thorough summary for the parents that helps them know in advance the different risks associated with their children.
Chapter 7

Concluding Remarks

7.1 Summary

VSIM is an easy to use and fast interpretation and visualization tool for variants in individual whole exome sequence data. Importantly, it can also be used as a premarital screening test. VSIM takes the VCF file of an individual as input to interpret genetic information of one person, or two VCF files of the mother and the father as an input file to the simulation model, together with the pre-downloaded database information to capture information regarding Mendelian disease, complex diseases, Digenic diseases, and Pharmacogenomics Knowledgebase for drug response. Also, VSIM predicts the pathogenic variants based on the MCAP score. Moreover, through the use of an RTG update simulation model, the “childsim” command is used to simulate an individual’s genotype information from a VCF, containing the two parent genotypes generated. The new output VCF contains all the existing variants and samples with a new column for the new child sample. VSIM then aims to identify the variants that are causes of potential diseases in an individual, or the resulting simulated samples. We visualize the results of the analysis using ideograms that provide clear formatting of the resulting information.
7.2 Future Research Work

More research is needed for better understanding of different characteristics of variants and how they are associated with the diseases, and this thesis work can be extended in many directions. We list a few potential future plans as follows:

- Increase number of databases that are used as a reference for annotation.

- Use a machine-learning approach for better analysis and identification of the causative variants.

- Improve the website functionality, in terms of the input file formats, speed and efficiency.

- Further improvement could be achieved by applying the simulation to several real affected families and improving the tool accordingly, in terms of predicting causative variants that are not reported in any of the well known disease databases.

- The simulation algorithm can also be used to simulate population-wide genetic epidemiology and study how diseases manifest themselves or look for the evolution of certain traits, because our algorithm is more generic than the one we now use for premarital testing application, as it has many more applications.
REFERENCES


## APPENDICES

### A Appendix A: List of free software and tools used in the research

<table>
<thead>
<tr>
<th>Tool</th>
<th>Rationale</th>
<th>Software Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bedtools</td>
<td>Bedtools utilities are tools for a wide-range of genomics analysis tasks, in form of command lines. In this work, we use Bedtools <strong>intersect</strong> that extract the intersection interval between two files, that is, between one or multiple VCF files, with the diseases databases.</td>
<td><a href="https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html">https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html</a></td>
</tr>
<tr>
<td>VCFtools</td>
<td>VCFtools is a package that was developed for working with VCF files, such as the files that were generated by the 1000 Genomes Project. This tool provides a convenient way to work with complex genetic variation data in the form of VCF files. We use this tool for various tasks, for example, <strong>vcf-merge</strong> to mate the genetic information of mother and father.</td>
<td><a href="http://vcftools.sourceforge.net/">http://vcftools.sourceforge.net/</a></td>
</tr>
<tr>
<td><strong>PHP</strong></td>
<td>This development language is used primarily because of its integration with the Hypertext Markup Language (HTML) in the VSIM website. It ensures quick prototyping and development.</td>
<td><a href="http://php.net/">http://php.net/</a></td>
</tr>
<tr>
<td><strong>Ideogram.js</strong></td>
<td>Ideogram.js is a JavaScript library which allows tight control in the formatting and visualization of genomics data in a chromosomal layout.</td>
<td><a href="https://github.com/eweitz/ideogram">https://github.com/eweitz/ideogram</a></td>
</tr>
<tr>
<td><strong>RTG</strong></td>
<td>RTG Tools contains utilities for manipulating multiple VCF files, as well as advantages for processing other NGS data formats. We use RTG in the simulation task and update the functionality of simulating a population of children and take into consideration the recombination rates.</td>
<td><a href="https://www.realtimegenomics.com/products/rtg-tools">https://www.realtimegenomics.com/products/rtg-tools</a></td>
</tr>
<tr>
<td><strong>Jvarkit</strong></td>
<td>Jvarkit is Java utilities for Bioinformatics. In this work for the testing examples from the 1000 Genome Project, the individual VCF files are combined, so we use biostar130456 tool for obtaining individual VCF files from the main VCF file.</td>
<td><a href="http://lindenb.github.io/jvarkit/">http://lindenb.github.io/jvarkit/</a></td>
</tr>
<tr>
<td><strong>ANNOVAR</strong></td>
<td>ANNOVAR is a software tool to utilize information to functionally annotate genetic variants. We used table_annovar.pl, with hg19 human genome and mcap protocol.</td>
<td><a href="http://annovar.openbioinformatics.org/en/latest/">http://annovar.openbioinformatics.org/en/latest/</a></td>
</tr>
</tbody>
</table>
B Papers Submitted and Under Preparation

• Azza Althagafi, Robert Hoehndorf, “VSIM: Visualization and simulation of variants in personal genomes with an application to premarital testing”, Under Preparation.