Pedro Filipe
Carneiro Venâncio

Aplicação de novos algoritmos bioinformáticos na análise de dados de Next Generation Sequencing (NGS)

Application of new bioinformatic algorithms in Next Generation Sequencing (NGS) data analysis.

DOCUMENTO PROVISÓRIO

Pedro Filipe
Carneiro Venâncio

Aplicação de novos algoritmos bioinformáticos na análise de dados de Next Generation Sequencing (NGS)

Application of new bioinformatic algorithms in Next Generation Sequencing (NGS) data analysis.

DOCUMENTO PROVISÓRIO

"*The greatest challenge to any thinker is stating the problem in a way that will allow a solution*"

— Bertrand Russell

Universidade de Aveiro
2024

Pedro Filipe Carneiro Venâncio

Aplicação de novos algoritmos bioinformáticos na análise de dados de Next Generation Sequencing (NGS)

Application of new bioinformatic algorithms in Next Generation Sequencing (NGS) data analysis.

Relatório de estágio curricular apresentado à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Bioinformática Clínica, especialização em Bioinformática do Genoma , realizado sob a orientação científica da Doutora Gabriela Maria Ferreira Ribeiro de Moura, Professora auxiliar do Departamento de Ciências Médicas da Universidade de Aveiro, e supervisão da Doutora Alexandra Filipa Lopes, membro da entidade de acolhimento Unilabs.

Dedico este trabalho à minha esposa e filho pelo incansável apoio.

o júri / the jury

presidente / president

Prof. Doutor João Antunes da Silva
professor associado da Faculdade de Engenharia da Universidade do Porto

vogais / examiners committee

Prof. Doutor João Antunes da Silva
professor associado da Faculdade de Engenharia da Universidade do Porto

Prof. Doutor João Antunes da Silva
professor associado da Faculdade de Engenharia da Universidade do Porto

Prof. Doutor João Antunes da Silva
professor associado da Faculdade de Engenharia da Universidade do Porto

Prof. Doutor João Antunes da Silva
professor associado da Faculdade de Engenharia da Universidade do Porto

Prof. Doutor João Antunes da Silva
professor associado da Faculdade de Engenharia da Universidade do Porto

agradecimentos / acknowledgements

Agradeço toda a ajuda a todos os meus colegas e companheiros.

Palavras Chave NGS, arquitetura, história, construção, materiais de construção, saber tradicional.

Resumo

Um resumo é um pequeno apanhado de um trabalho mais longo (como uma tese, dissertação ou trabalho de pesquisa). O resumo relata de forma concisa os objetivos e resultados da sua pesquisa, para que os leitores saibam exatamente o que se aborda no seu documento.
Embora a estrutura possa variar um pouco dependendo da sua área de estudo, o seu resumo deve descrever o propósito do seu trabalho, os métodos que você usou e as conclusões a que chegou.
Uma maneira comum de estruturar um resumo é usar a estrutura IMRaD. Isso significa:

- Introdução
- Métodos
- Resultados
- Discussão

Veja mais pormenores aqui:
`https://www.scribbr.com/dissertation/abstract/`

Keywords textbook, architecture, history, construction, construction materials, traditional knowledge.

Abstract

An abstract is a short summary of a longer work (such as a thesis, dissertation or research paper).
The abstract concisely reports the aims and outcomes of your research, so that readers know exactly what your paper is about.
Although the structure may vary slightly depending on your discipline, your abstract should describe the purpose of your work, the methods you've used, and the conclusions you've drawn.
One common way to structure your abstract is to use the IMRaD structure. This stands for:

- Introduction
- Methods
- Results
- Discussion

Check for more details here:
`https://www.scribbr.com/dissertation/abstract/`

Acknowledgement of use of AI tools

Recognition of the use of generative Artificial Intelligence technologies and tools, software and other support tools.

I acknowledge the use of ChatGPT 3.5 (Open AI, https://chat.openai.com) for paraphrasing and translations, the use of Adobe Illustrator (Adobe, https://www.adobe.com/pt/products/illustrator) for image creation and editing, the use of FreePik (FreePik, https://br.freepik.com/) for vector images download, and the use of diagrams.net (diagrams.net, https://app.diagrams.net/) for diagram creation.

Contents

List of Figures

List of Tables

List of Code Snippets

Glossary

NGS Next Generation Sequencing
CLIA Clinical Laboratory Improvement Amendments
ISO International Organization for Standardization
WES Whole Exome Sequencing
WES Whole Genome Sequencing
aCGH Comparative Genomic Hybridization
FISH Fluorescence In Situ Hybridization
MLPA Multiplex Ligation-Dependent Probe Amplification
DNA Deoxyribonucleic Acid
HGP Human Genome Project
RNA Ribonucleic Acid
SWOT Strengths Weeknesses Opportunities and Threats
ES Exome Sequencing
GS Genome Sequencing
GRCh38/hg38 Reference Consortium Human Build 38
BAM Binary Alignment Map
SAM Sequence Alignment Map
CRAM Compressed Reference-oriented Alignment Map
VUS Variants With Unknown Significance

CHAPTER 1

Introduction

"The only source of knowledge is experience." - Albert Einstein

1.1 Internship Context and Framework

This document represents the final report of the internship carried out as part of the Internship Curricular Unit (49991) of the second year of studies of the Master's Degree in Clinical Bioinformatics, with specialization in Genome Bioinformatics, at the University of Aveiro. The internship lasted nine months, starting on November 21st, 2023, and ending on July 19th, 2024, totalling 1296 hours of work.

During this period, the trainee had the opportunity to apply the knowledge acquired throughout the course and to get involved in practical projects related to bioinformatics and genomics. Unilabs, a recognized company in the health area, provided a professional environment where the intern could collaborate with experienced professionals and actively participate in projects relevant to clinical bioinformatics. This report addresses the activities developed during the internship and the contributions to the projects in which the intern was involved.

This introductory section aims to offer an overview of the context in which the internship was carried out, laying the foundations for understanding the activities and results presented throughout the report.

1.2 Project motivation and objectives

Currently, Unilabs uses a genomic intelligence platform that uses natural language processing to analyse new scientific publications of a genetic nature and incorporate them into an always updated knowledge base. This platform is particularly useful in prioritizing sequenced variants, for genetic diagnosis purposes, in their interpretation and in the production of clinical reports, thus enabling the provision of increasingly personalized care.

However, at the time of this internship, Unilabs was in the migration phase to this new platform and, therefore, as a complementary strategy, a new independent software was developed to obtain the necessary metrics for genomic analyses, not directly provided by the aforementioned platform, ensuring compliance with the guidelines and practices recommended for Next Generation Sequencing (NGS). These metrics are important for assessing data quality, i.e., they indicate how well the target regions were covered by sequencing. In the case of the present stage, it was suggested to obtain the sequencing depth. In addition, the depth of coverage directly influences the ability to detect genetic variants: regions with low coverage can result in undetected or underestimated variants. Additionally, coverage metrics are also useful to optimize sequencing protocols, adjusting experimental parameters to ensure adequate coverage of target regions and minimize unnecessary costs.

As will be explained in detail below, the software created and described in this report allows the obtaining of Average Read Depth and Percentage of Coverage at 1x, 10x, 15x, 20x, 30x, 50x, 100x, and 500x per gene and per panel in analysis of gene panels. Additionally, in addition to the presentation of metrics by panel, single gene and exome analysis was also implemented.

1.3 Document Structure

This document is divided into five main chapters, each with several sections, and includes additional content at the end.

The first chapter, **Introduction**, begins with the **Internship Context and Framework**, providing an overview of the internship setting and its relevance. This is followed by a presentation of the **Project Motivation and Objectives**, outlining the purpose and goals of the work. The chapter also includes the **Document Structure**, this section, which explains how the document is organized. The **Characterization of the Host Entity and Work Plan** section describes the host entities, including Unilabs Portugal and Unilabs Genetics, and provides the timeline for the work. Finally, the **Theoretical Framework** section covers important concepts in Genetics and Genomics, Sequencing Methods and Characteristics, and Bioinformatics, including sub-sections on the evolution of genetics, the structure and function of DNA, the discovery of the double helix, DNA packaging, mutations, and genetic variations, as well as Next Generation Sequencing and a SWOT analysis.

The second chapter, **Software Development Process**, is detailed as follows: The **Analysis** section outlines the preliminary work, followed by **Planning**, which covers the project planning phase. The **Design** section describes the design phase, and **Development** covers the creation of the software, including **Environment Preparation**. The chapter also includes sections on **Testing and Validation**, **Optimization**, **Deployment**, **Documentation**, and **Maintenance**, explaining how each phase contributes to the software development lifecycle. This structure of software presentation is based on the SWEBOK Guide [1], which provides a comprehensive overview of software engineering practices.

The third chapter, **Additional Activities During the Internship**, highlights other activities and experiences gained during the internship.

The fourth chapter, **Discussion**, provides a comprehensive analysis of the work, interpreting results in the context of existing knowledge and discussing practical applications, limitations, and implications.

The fifth chapter, **Final Remarks**, presents conclusions drawn from the work and suggests possible improvements and new features for future versions of the software.

The document concludes with a **Bibliography** listing all consulted and cited sources, followed by Additional Content which includes supplementary material relevant to the work.

1.4 Characterization of the Host Entity and Work Plan

1.4.1 Unilabs Portugal

Since the beginning of 2006, Unilabs has established solid roots in Portugal. It began its journey with the acquisition of most of the shares of the company "Medicina Laboratorial Dr. Carlos Torres" and since then it has grown steadily, following a strategy of acquiring high-quality laboratories and partners throughout the country.

It has more than 3,500 employees and more than 500 doctors, and operates in more than 1,000 service units, performing more than 25 million medical procedures per year.

In 2017, Unilabs took an important step by acquiring BASE Holding. With this acquisition, it has expanded its service offering to include radiology, affirming its position as a national leader in integrated clinical diagnostics and the provision of Complementary Diagnostic and Therapeutic Means.

Currently, it offers a wide variety of services in various areas, including Clinical Analysis, Pathological Anatomy, Cardiology, Gastroenterology, Medical Genetics, Nuclear Medicine and Radiology. The company maintains its commitment to being close to people, providing answers that contribute to a healthier future. [2]

1.4.2 Unilabs Genetics

Unilabs Genetics (formerly called CGC Genetics) was founded three decades ago and was the first private Medical Genetics laboratory in Portugal. It has been present on the national scene regarding diagnosis through genetic studies. It has a wide selection of tests and is known for its collaborative and thorough approach, meeting the demands of clinicians in a variety of specialist areas.

It is a leader in Europe in Medical Genetics, with special emphasis on rare diseases. It provides accurate diagnoses for public and private healthcare institutions, providing physicians and patients with detailed information about the nature of diseases, prognosis, and treatment options. In addition, the company supports academic institutions, research centres and the pharmaceutical industry with data and knowledge that contributes to the discovery of biomarkers and the development of new drugs.

The Unilabs Genetics laboratory is located in the city of Porto and combines advanced technologies, bioinformatics and artificial intelligence, with a highly qualified team of medical geneticists, specialists in genetic counselling and laboratory technicians. The company follows

the strictest quality and ethics policies, having certifications (Clinical Laboratory Improvement Amendments (CLIA), International Organization for Standardization (ISO) 15189, ISO 9001) that guarantee excellence in its services. [3]

The Table A.1 presents the test catalog provided by Unilabs Genetics.

1.4.3 Timeline

The internship at Unilabs followed a structured timeline, aimed at ensuring the successful completion of all assigned tasks. The timeline was designed to provide a clear framework for the development and implementation of various activities throughout the internship period. This detailed plan served as a guide to ensure that each phase of the project was completed efficiently and on time, while also allowing the necessary flexibility to accommodate any adjustments. The Figure 1.2 outlines the key milestones and deadlines that were met during this internship.

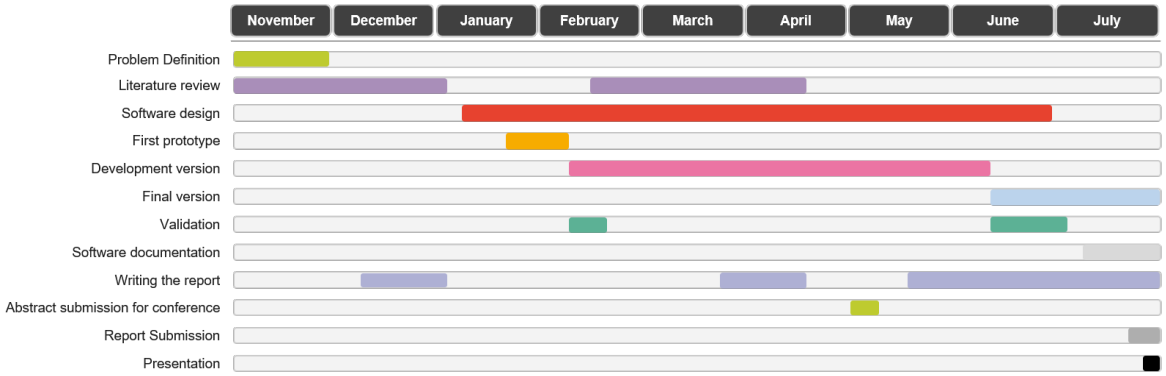


Figure 1.1: Imagem provisória do cronograma do estágio.

1.5 Theoretical Framework - Genetics and Genomics

1.5.1 Evolution of genetics over the years

The history of genetics formally began in 1865 with Gregor Mendel's work on plant hybridization. However, the term "genetics" was only coined in 1906 by the English biologist William Bateson to define the new science of heredity. Based on Mendel's laws, genetics introduced groundbreaking concepts such as gene, genotype, and phenotype. By the 1910s, Mendelian genetics merged with the chromosomal theory of inheritance, giving rise to classical genetics. In this framework, the gene was seen as a unit of function, transmission, recombination, and mutation. [4]

This understanding persisted until the 1950s, when DNA was discovered as the material basis of heredity, marking the start of molecular biology. [5]

Following the discovery of DNA as hereditary material, molecular biology began to uncover the complexity of gene function. The fusion of Mendel's ideas with chromosomal theory also provided a more tangible understanding of genes, which could now be physically located on chromosomes. This integration led to significant advances, such as explaining Mendel's laws through cellular mechanisms and discovering genetic recombination. Genetics evolved into

a more institutionalized science, with the establishment of academic chairs and specialized courses worldwide, solidifying its position as a central field in the biological sciences. [5]

The emergence of genomics in the latter half of the 20th century further transformed the field of genetics. The completion of the Human Genome Project (HGP) in 2003 [6], a milestone in genomics, revealed the entire sequence of human DNA, propelling the study of genes beyond individual units to entire genomes. This large-scale approach allowed scientists to explore the intricate network of genes and their interactions, significantly advancing our understanding of complex traits and diseases. Genomics also facilitated the development of personalized medicine, where treatments could be tailored based on an individual's genetic makeup. [5]

The discovery of the Ribonucleic Acid (RNA)-guided CRISPR-Cas9 system has made genome editing easier and more efficient. This breakthrough allows scientists to modify DNA in various cells and organisms with ease, removing previous experimental barriers. Today, CRISPR-Cas9 is widely used in basic research, biotechnology, and the development of new therapies. [7]

Figure 1.2 presents a timeline with some of the major historical milestones in the evolution of genetics over the years.

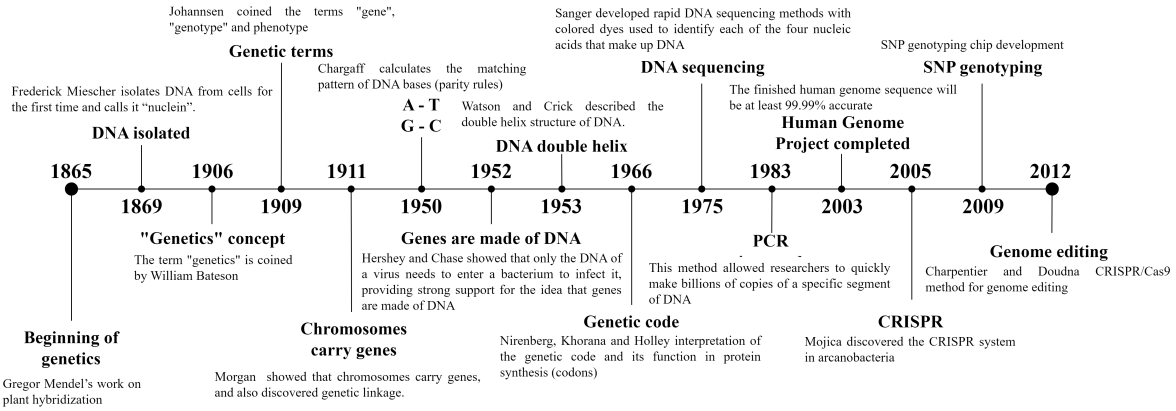


Figure 1.2: Evolution of genetics over the years: A brief timeline with some of the major historical milestones. Image adapted from [8] and [4]

1.5.2 Genetics vs Genomics

Genetics and genomics are both fields of study that explore the roles of genes in living organisms, but they focus on different aspects of heredity and DNA. Genetics is the study of specific genes and their influence on traits and conditions that are passed from one generation to the next. It examines how certain genes cause inherited disorders, such as cystic fibrosis and Huntington's disease. [9]

In contrast, genomics is a more recent field that encompasses the study of all the genes within an organism, referred to as the genome, and how these genes interact with each other and the environment. Unlike genetics, which focuses on individual genes, genomics uses advanced technologies like bioinformatics and high-performance computing to analyze vast amounts of genetic data. This comprehensive approach is crucial for studying complex diseases,

such as cancer and diabetes, which result from the interplay between multiple genes and environmental factors. While both fields contribute to advancements in health and disease treatment, genomics represents a broader, more holistic view of genetic influence. [10]

1.5.3 Structure and function of DNA, RNA, and proteins

Nucleic acids, specifically DNA and RNA, are fundamental molecules in biological systems, playing key roles in storing and transmitting genetic information. DNA, which exists primarily as a double-stranded helix, encodes the instructions necessary for the growth, development, and reproduction of all living organisms. RNA, on the other hand, serves multiple purposes, including acting as a messenger that carries genetic information from DNA to the ribosomes for protein synthesis. The structural complexity and functional versatility of these molecules underscore their importance in the central dogma of molecular biology, which describes the flow of genetic information from DNA to RNA to proteins. [11]

1.5.4 Molecular Structure of DNA

The molecular structure of DNA is a polymer composed of repeating units called nucleotides. Each nucleotide consists of three components: a five-carbon sugar (deoxyribose), a phosphate group, and a nitrogenous base. The nitrogenous bases are categorized into two groups: purines (adenine and guanine) and pyrimidines (cytosine and thymine). The nucleotides are linked together by phosphodiester bonds, forming a sugar-phosphate backbone that gives DNA its structural integrity. DNA molecules are double-stranded, with two complementary strands running in opposite directions (antiparallel orientation). These strands are held together by hydrogen bonds between specific base pairs: adenine pairs with thymine, and guanine pairs with cytosine. This complementary base pairing is crucial for the accurate replication and transmission of genetic information. [11]

The Figure 1.3 shows a representation of the molecular structure of DNA and the base pairing between adenine (A), thymine (T), cytosine (C), and guanine (G).

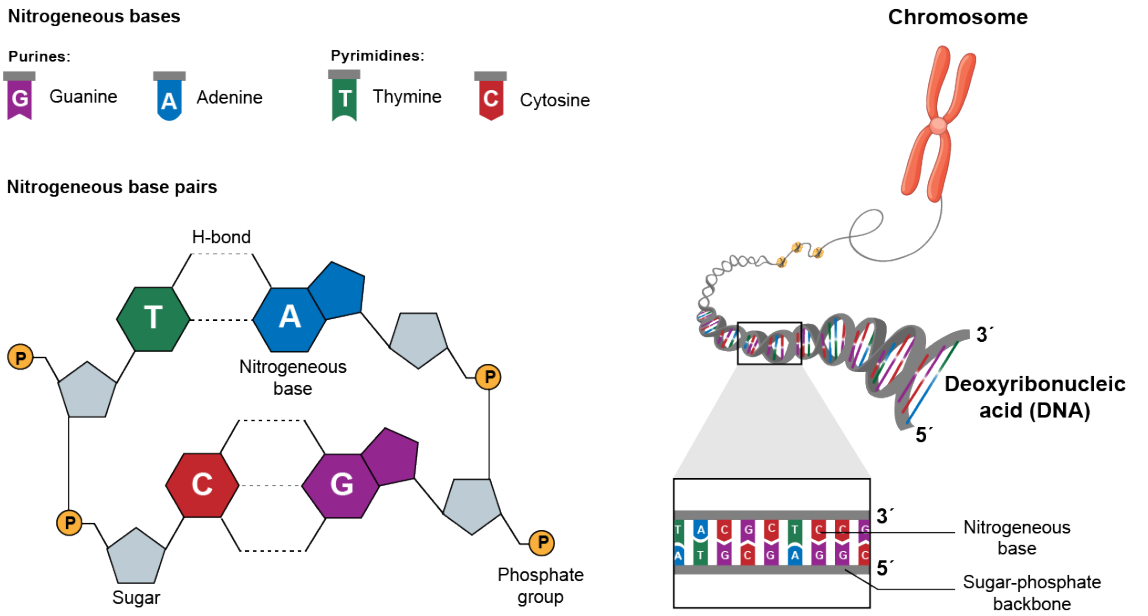
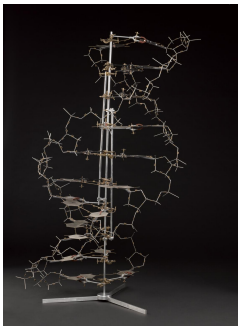


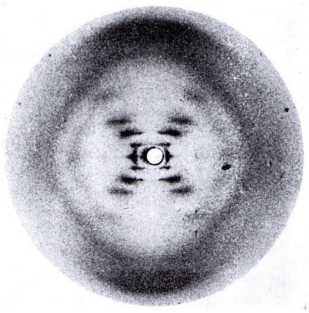
Figure 1.3: Representation of DNA and its constituents. In the top left corner, the nitrogenous bases are illustrated, categorized into Purines (Guanine and Adenine) and Pyrimidines (Thymine and Cytosine). Below this, on the left side, the paired nitrogenous bases are shown, along with their respective hydrogen bonds and the sugar-phosphate backbone connections. On the right side of the image, a chromosome is depicted unraveling, revealing the DNA structure. [12]

1.5.5 Discovery of the Double Helix

The three-dimensional structure of DNA, famously known as the double helix, was elucidated by James Watson and Francis Crick in 1953. Their discovery was informed by Rosalind Franklin's X-ray diffraction images, which revealed the helical structure of DNA. Watson and Crick proposed that the two strands of the helix are wound around each other, with the sugar-phosphate backbone on the outside and the nitrogenous bases on the inside. The helical structure is right-handed, with ten base pairs per turn of the helix. The stability of the double helix is largely due to the hydrogen bonds between the complementary base pairs and the hydrophobic interactions between the stacked bases. This discovery not only explained how DNA could carry genetic information but also provided insights into how it could be replicated during cell division. [11]



(a) Double helix model of DNA



(b) X-ray diagram of DNA

Figure 1.4: A figure with the reconstruction of the Watson-Crick's double helix model of DNA in 1.4a built by Science Museum Group Collection [13] and Franklin's X-ray diagram of the B form of sodium thymonucleate (DNA) fibres in 1.4b, published in Nature on 25 April 1953 [14]

1.5.6 DNA Packaging in Eukaryotic Cells

In eukaryotic cells, DNA is not free-floating within the nucleus; instead, it is highly organized and compacted into structures known as chromosomes. This compaction is achieved through the association of DNA with histone proteins, forming nucleosomes, which are the basic unit of chromatin. Each nucleosome consists of a segment of DNA wrapped around a core of histone proteins. These nucleosomes are further coiled and folded into higher-order structures, eventually forming the condensed chromosomes visible during cell division. The packaging of DNA into chromatin is essential for fitting the large eukaryotic genome into the limited space of the nucleus. Furthermore, chromatin structure plays a crucial role in gene regulation, as regions of tightly packed chromatin (heterochromatin) are generally transcriptionally inactive, while loosely packed regions (euchromatin) are more accessible to transcriptional machinery. [11]

1.5.7 Mutations and genetic variations

Mutations are fundamental drivers of genetic diversity, occurring when changes happen in the DNA sequence. These alterations can range from small-scale mutations, such as the substitution, insertion, or deletion of one or a few nucleotides, to large-scale mutations that involve significant chromosome segments or entire genes. Chromosome mutations specifically impact either individual nucleotides or larger chromosome fragments. They can involve deletions, insertions, inversions, translocations, or even gene duplications. On the other hand, genome mutations refer to changes in the number of whole chromosomes or sets of chromosomes and are studied separately. [15]

Genetic variation, on the other hand, refers to the observable differences between individuals within a population, which arise from variations in their genotypes. While mutations are the source of new genetic variation, genetic variation encompasses the broader concept of how different alleles at specific loci contribute to diversity within a population. This variation is shaped and refined by evolutionary forces such as natural selection, gene flow, and genetic

drift. In cultivated plants, human-directed selection can also manipulate genetic variation to enhance desirable traits. [15]

1.6 Theoretical Framework - Sequencing Methods and Characteristics

Even though there are several methods for sequencing DNA, the most widely used technique nowadays is Next Generation Sequencing. This section focuses only on NGS and its applications, including gene panels, exome sequencing, and genome sequencing.

1.6.1 Next Generation Sequencing - Gene Panels, Exome, and Genome Sequencing

Next Generation Sequencing technologies have revolutionized genomic medicine, enabling rapid advancements in clinical diagnostics, therapeutic decision-making, and disease prediction. Unlike traditional sequencing methods such as Sanger sequencing, which are limited by low throughput and high cost [16], NGS allows for the massively parallel sequencing of DNA, significantly increasing throughput and reducing costs by several orders of magnitude. As a result, clinical laboratories now have the capability to analyze nearly complete exomes or genomes of individuals, thereby improving the diagnostic process for a wide range of genetic conditions. [17]

NGS is characterized by the use of clonally amplified or single molecule templates, which are sequenced in parallel, providing an unprecedented ability to analyze large amounts of DNA efficiently. The adoption of NGS in clinical settings is growing rapidly, with three primary applications gaining prominence: disease-targeted gene panels, Exome Sequencing (ES), and Genome Sequencing (GS). [17]

Disease-targeted gene panels

Disease-targeted gene panels focus on a set of known disease-associated genes, allowing for greater depth of coverage and increased analytical sensitivity. This targeted approach improves the detection of heterozygous variants, mosaicism, or low-level heterogeneity, particularly in applications related to mitochondrial diseases or oncology. Targeted panels also allow laboratories to leverage desktop sequencers, reducing the costs of sequencing and data storage. However, follow-up techniques such as Sanger sequencing may be required to fill gaps in the data caused by regions of low coverage. [17]

Exome Sequencing

Exome Sequencing targets the coding regions of the genome, which constitute approximately 1-2% of the entire genome but harbor around 85% of known disease-causing mutations. [18] ES is particularly valuable in "detecting variants in known disease-associated genes as well as for the discovery of novel gene-disease associations" [17], with clinical studies demonstrating a diagnostic success rate of approximately 20%. [19] Despite this potential, ES faces challenges related to coverage variability, as certain exonic regions may not be captured or sequenced with sufficient depth to make a sequence call, leading to the need for additional sequencing techniques to confirm findings. [17]

Genome Sequencing

Genome Sequencing offers a more comprehensive approach by covering both coding and non-coding regions of the genome. This technique simplifies the preparation of samples for sequencing by eliminating the need for pre-sequencing enrichment strategies. While GS holds promise for identifying regulatory variants and structural variants that are outside of coding regions, it remains the most expensive NGS technology and typically provides lower average depth of coverage. Nonetheless, as technology advances, these limitations are expected to diminish, making GS a more accessible option for clinical diagnostics. [17]

NGS involves three major components: sample preparation, sequencing, and data analysis. These steps are interrelated, and the quality of each influences the overall outcome of the sequencing process. Below is an overview of these essential stages according to the guidelines of [17].

1.6.2 Next Generation Sequencing - Overview

Sample preparation

The NGS process begins with the extraction of genomic DNA from a biological sample, typically from a patient. The quality and quantity of this DNA are critical to successful sequencing. Laboratories must specify the required sample type and amount based on their validation data. Sample mix-ups must be prevented through robust processes, as even minor errors can significantly affect results. [17]

For specific NGS applications like targeted panels or ES, enrichment strategies are employed to focus on a subset of genomic regions. Enrichment ensures that only the regions of interest are sequenced, improving efficiency and reducing costs. Enrichment can be achieved through various methods, including multiplex PCR and hybridization-based capture. [17]

Library Generation and Barcoding

Library generation is the process of preparing DNA fragments of a specific size (100-500 base pairs) for sequencing. These fragments are tagged with adapter sequences on both ends, which are essential for downstream sequencing steps. The fragmentation of DNA can be performed using different methods, each with its advantages and limitations. In most NGS workflows, PCR amplification is used to amplify the DNA library before sequencing. [17]

Barcoding is a crucial step in library preparation, where each sample is tagged with a unique sequence identifier. This allows multiple samples to be pooled together in a single sequencing run, reducing the cost per sample. Barcoding is typically integrated into the adapter sequences or added during a PCR enrichment step. [17]

Target Enrichment

In many NGS applications, particularly targeted panels and ES, only specific regions of the genome are sequenced. Target enrichment methods are used to isolate these regions before sequencing. Target enrichment strategies include PCR-based methods (e.g., single or multiplex PCR) and hybridization-based capture. While PCR-based methods are suitable for

smaller panels, hybridization-based approaches are preferred for larger-scale applications like exome sequencing. [17]

Sequencing Platforms

NGS platforms are designed to perform millions of parallel chemical reactions, allowing them to sequence vast amounts of DNA simultaneously. These platforms utilize different sequencing chemistries, such as sequencing by synthesis, sequencing by ligation, and ion sensing. [20] The choice of sequencing platform depends on various factors, including sequence capacity, read length, run time, cost, and accuracy. [17]

Each platform has its own strengths and weaknesses. For example, some platforms excel at producing longer reads, while others are optimized for high-throughput sequencing at a lower cost per sample. The selection of a platform is influenced by the specific clinical or research application. [17]

1.7 Theoretical Framework - Bioinformatics

1.7.1 Next Generation Sequencing - Data Analysis

NGS generates an enormous amount of sequence data, necessitating the development of sophisticated data analysis pipelines. These pipelines are designed to process and interpret the raw sequencing data, transforming it into meaningful genomic information. NGS data analysis can be divided into four primary steps: Base Calling, Read Alignment, Variant Calling, and Variant Annotation. The bioinformatics analysis stage of NGS is essential for converting unprocessed sequencing data into biologically and clinically significant findings. This section provides a bioinformatics overview of these steps and their importance in the NGS workflow. [17]

Base Calling

Finding the nucleotide at each place in a sequencing read is known as **Base Calling**. In order to ensure that the raw data gathered during sequencing is transformed into a sequence of nucleotides, this step is usually included into the software of the sequencing device. [17]

For example, the Illumina NovaSeq 6000 Sequencing System uses two-channel sequencing approach, requiring only two images to represent data for all four DNA bases, with one image capturing information from the red channel and the other from the green channel. [21]

An 'N' designation, or 'no call' is used when a cluster fails to meet quality filters, registration is unsuccessful, or the cluster is not properly captured in the image. The process involves extracting intensity data for each cluster from both the red and green images and comparing these intensities to identify four distinct populations. Each of these populations is associated with one of the DNA bases, and the base calling process assigns each cluster to the corresponding population. [21]

Figure 1.5 illustrates the intensity data for each cluster from both the red and green channels in 2-channel sequencing, as used in the NovaSeq 6000 Sequencing System.

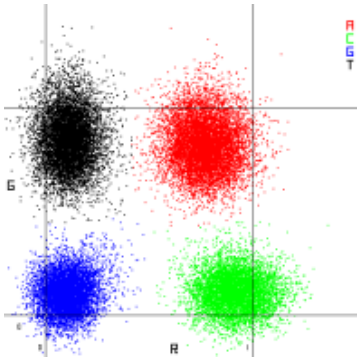


Figure 1.5: Visualization of cluster intensities in 2-Channel Sequencing (NovaSeq 6000 - Illumina). The image shows the intensity data for each cluster from both the red and green channels, with each cluster representing a different DNA base. Image from [21].

Table 1.1 outlines how the signal intensities from red and green fluorescence channels are used to identify DNA bases (A, C, G, T) during sequencing. Each base is associated with a unique combination of signal intensities from these two channels. The red and green channels either show intensity, marked as "on" (1), or no intensity, marked as "off" (0), at specific cluster locations where DNA bases are being read.

When both the red and green channels show intensity (1,1), this indicates the presence of adenine (A). The simultaneous detection of both red and green signals suggests that the cluster is emitting light in both spectrums, and this combination is uniquely attributed to adenine.

For cytosine (C), the red channel is on (1), but the green channel is off (0). This means that the cluster emits light in the red spectrum only, and the absence of green emission differentiates it as cytosine.

When neither the red nor the green channel shows intensity (0,0), the system identifies guanine (G). The lack of any signal at a known cluster location suggests that no fluorescence is being emitted, and this corresponds to guanine.

Finally, thymine (T) is identified when the red channel is off (0) and the green channel is on (1). In this case, the cluster is emitting light in the green spectrum only, and the absence of red fluorescence distinguishes it as thymine.

This method of combining red and green fluorescence signals allows the sequencing system to effectively differentiate between the four DNA bases by associating specific patterns of signal intensity with each base.

Table 1.1: Base Calls in 2-Channel Sequencing (NovaSeq 6000 - Illumina). Table from [21]

Base	Red Channel	Green Channel	Result
A	1 (on)	1 (on)	Clusters that show intensity in both the red and green channels.
C	1 (on)	0 (off)	Clusters that show intensity in the red channel only.
G	0 (off)	0 (off)	Clusters that show no intensity at a known cluster location.
T	0 (off)	1 (on)	Clusters that show intensity in the green channel only.

Read Alignment/Mapping

Base calling is followed by demultiplexing, the process of separating reads from different samples based on their unique barcodes. After demultiplexing, the next step is store reads in a FASTQ file, a text-based format that includes a read identifier, the nucleotide sequence, a separator, and a quality score string. The quality scores are calculated using the Phred scale, which represents the likelihood of a base call being correct, with higher scores indicating greater confidence. For a given base error probability, P, the Phred quality score, Q, is calculated in Equation 1.1.

$$Q = -10 \log_{10} P \tag{1.1}$$

FASTQ is the standard format for raw sequencing data and is commonly used in single-end and paired-end sequencing. [22] The Figure 1.6 shows an example of a FASTQ file, with the read identifier, nucleotide sequence, and quality score (highlighting the score for one T base and the respective P error).

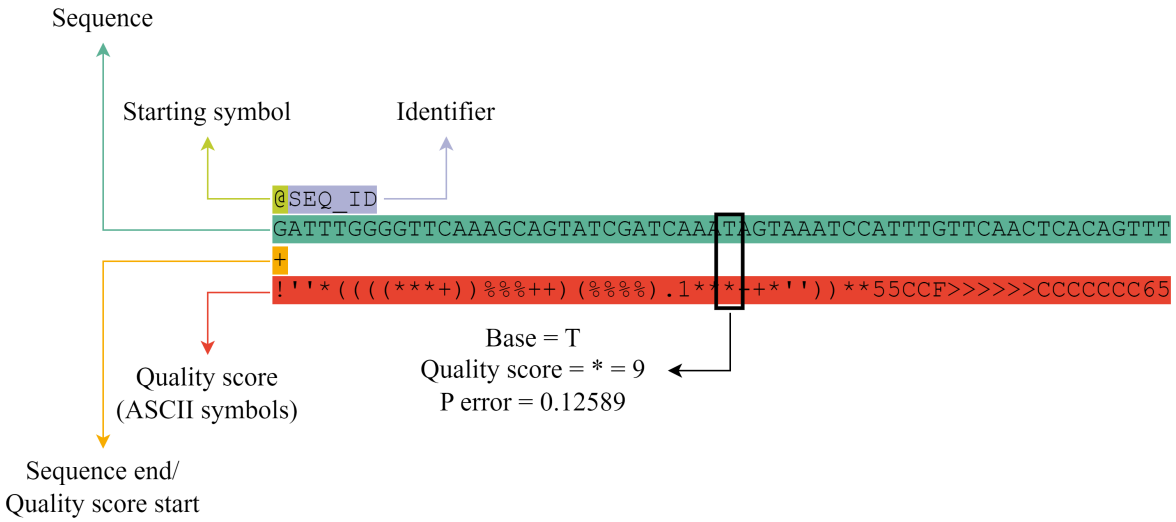


Figure 1.6: FASTQ file format example. The image shows a read identifier, nucleotide sequence, and quality score string in a FASTQ file. The P error calculation was performed based on the Phred quality score equation and Quality Score Encoding from Table A.2. This file example was adapted from [23].

Once the raw sequencing data was stored in a FASTQ file, the next step is **Read Alignment/Mapping**. [22] In this stage, the small DNA sequences (50–400 base pairs), reads, are positioned in relation to a reference genome, such as Reference Consortium Human Build 38 (GRCh38/hg38). [17] To improve alignment accuracy, low-quality bases and adapter sequences are also trimmed and the results are saved in either Sequence Alignment Map (SAM) or Binary Alignment Map (BAM) formats, with BAM being more efficient due to compression and indexing capabilities. A more compressed format, Compressed Reference-oriented Alignment Map (CRAM), further reduces file sizes by storing only differences from the reference genome, though it uses lossy compression in some cases (meaning that some

CRAM files may not be fully convertible back to BAM) CRAM is becoming a popular choice for long-term data storage due to its space efficiency - a CRAM file could be 50%-80% smaller than a BAM file. [22]

Variant Calling

Variant Calling identifies genetic differences between a sample's sequencing reads and a reference genome using specialized algorithms known as variant callers. Common variants, include single-nucleotide polymorphisms (SNPs), single-nucleotide variants (SNVs), small insertions and deletions (INDELs), and copy number variants or alterations (CNVs and CNAs). The first tree are related to Sequence Variants, while the last one is related to Structural Variants. [24], [25], [26], [27] SNPs refer to single-base changes in germline DNA, often biallelic, while SNVs cover any point mutation. CNVs involve larger DNA segments that are amplified or deleted, and CNAs specifically refer to somatic changes, often observed in cancer. [22] There are also other types of variants, such as translocations, inversions, and complex rearrangements, which are less common but can have significant clinical implications. [24]

The confidence with which variants are called depends heavily on sequencing depth and the variant allele frequency (VAF), which measures the proportion of DNA molecules in a sample that contain the variant allele. For germline heterozygous variants, where the variant is expected to appear in roughly 50% of reads, reliable detection can typically be achieved at sequencing depths of 20X to 30X. However, somatic variants, especially in cancer samples, tend to have lower VAFs due to tumor heterogeneity and sample contamination, requiring much higher coverage for reliable detection. [22]

Preprocessing of BAM files is essential before variant calling. Duplicates, originating from the same DNA fragment, must be marked to avoid skewing VAF estimates. Deduplication is recommended for whole-genome or exome sequencing but is usually skipped in PCR-based methods. Base quality scores are also recalibrated to correct systematic errors. After preprocessing, the BAM files are ready for variant detection. [22]

Germline Variant Calling for SNVs can be done using tools like Samtools or more advanced programs like GATK HaplotypeCaller, which improve accuracy in complex regions by applying local realignments. For biallelic SNPs, the reference allele (REF) is compared to the alternate allele (ALT). Humans, being diploid, have three possible SNP genotypes: homozygous reference (REF, REF), heterozygous (REF, ALT), and homozygous alternate (ALT, ALT), corresponding to variant allele frequencies (VAFs) of 0%, 50%, and 100%. Genotype quality (GQ), similar to base quality scores, reflects the confidence in a genotype call, measured on a Phred scale. [22]

When detecting copy number variants (CNVs) using NGS data, algorithms primarily rely on sequencing coverage patterns, although they may also take into account the VAFs of overlapping variants. CNV detection in targeted sequencing approaches like whole-exome sequencing (WES) or gene panels can be challenging due to coverage gaps and inconsistencies caused by capture bias. As a result, CNV identification within individual samples is difficult, and many tools designed for this purpose compare the data against a reference set of normal

samples. [22]

Somatic Variant Detection is enhanced by comparing tumor and matched normal tissue sequencing data, allowing inherited germline variants to be filtered out. When matched normal samples are unavailable, a "panel of normals" can serve as a reference. Additionally, classifiers trained on databases of somatic and germline variants can help predict the somatic status of variants found in tumor tissue. Despite these approaches, somatic variant detection still faces challenges, particularly due to the Euro-centric bias of many population allele frequency databases, which may reduce the accuracy of variant calls in underrepresented populations. [22]

Variants are typically stored in variant call format (VCF) files, which can include data from multiple samples. Genomic VCF (gVCF) files capture both variant and non-variant regions, allowing for easier data merging. Both VCF and gVCF files can be indexed for efficient access. Somatic mutations may also be saved in Mutation Annotation Format (MAF) files, which aggregate variant data from multiple VCFs. [22]

Variant Annotation

Once variants are called, **Variant Annotation** is carried out. This involves adding contextual information to each detected variant, such as determining its location within or near a gene and predicting its potential impact on protein function. Clinical interpretation of variants often includes data from variant databases, evolutionary conservation studies, and in silico predictions of pathogenicity. [17]

NGS often generates numerous variants, including many Variants With Unknown Significance (VUS), making it difficult to determine which are clinically relevant. Variant annotation helps prioritize and interpret these findings. For protein-coding regions, tools like REVEL [28] predict the functional impact of missense variants, while noncoding variants are assessed using resources such as CADD [29], FunSeq2 [30], and RegulomeDB [31]. These are complemented by data from projects like ENCODE [32] and Roadmap Epigenomics [33], and external databases like gnomAD [34], ClinVar [35], HGMD [36], and COSMIC [37]. Annotation software like ANNOVAR [38] integrates these resources to enrich VCF files with variant details. [22]

Overall, accurate and efficient data analysis in NGS requires significant bioinformatics support and robust computational infrastructure. This aspect of NGS is crucial for translating raw sequencing data into clinically relevant insights.

Figure 1.7 illustrates the general bioinformatics analysis pipeline for NGS, highlighting the key steps mentioned previously.

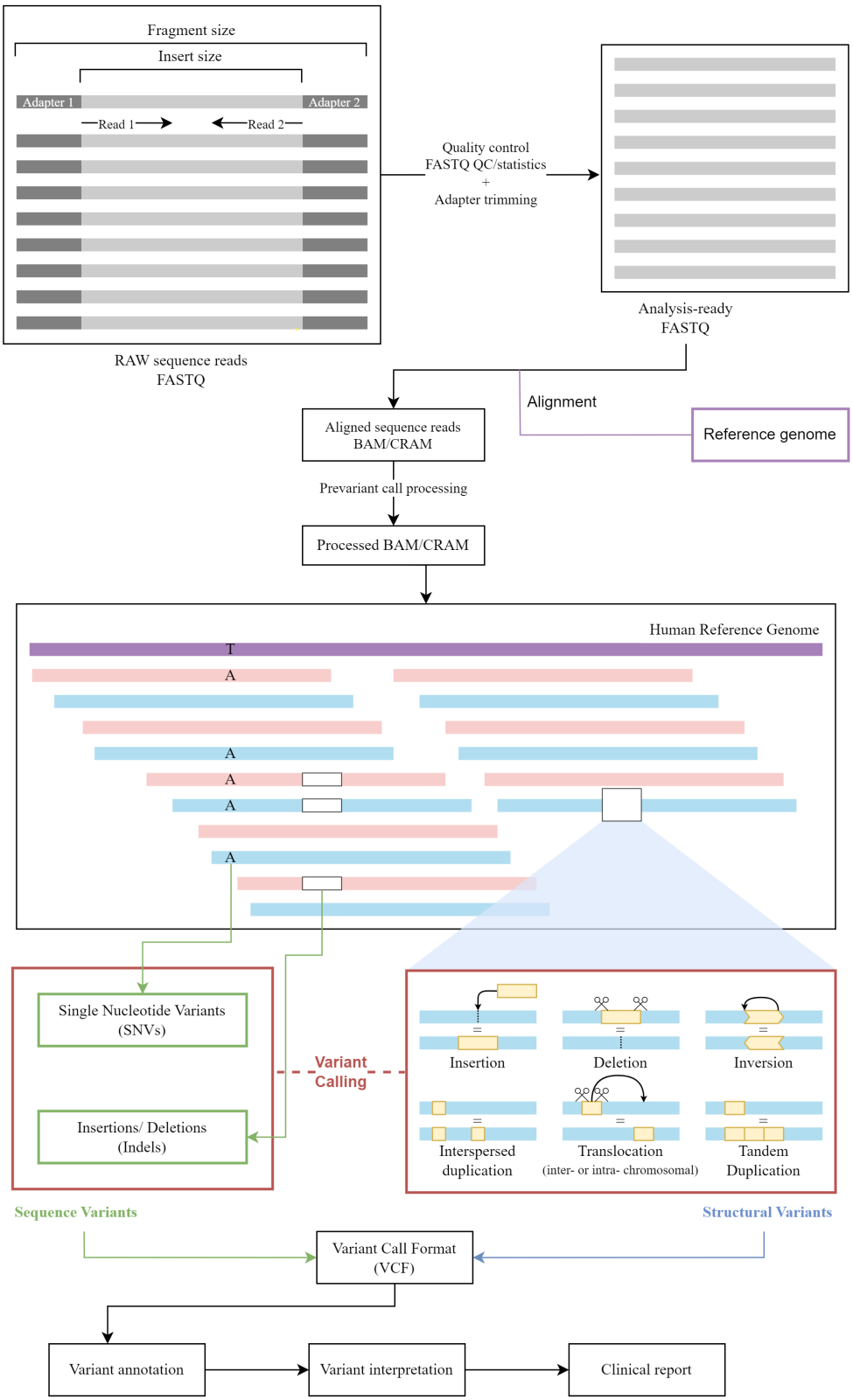


Figure 1.7: NGS data analysis pipeline. Adapted from [24], [25], [26], [27]

1.7.2 Next Generation Sequencing - Validation

In genetic and genomic analysis, namely in NGS, a key focus is often placed on the comparison between coverage and sequencing depth. While these terms are related, they offer distinct insights into the quality and reliability of sequencing data. A thorough understanding of both concepts is essential to ensure accurate and meaningful results in genetic studies.

Coverage or Breadth of Coverage

Several elements could influence the total sequencing output, including the efficiency of the libraries, the extent of sample multiplexing, and the number of sequencing cycles. Given that the number of lanes in the flow cell is fixed, adjusting these factors is key to maximizing throughput. This throughput is commonly described in terms of coverage, which measures the proportion of the genome or target region that has been sequenced at least once. In technical documentation, coverage is often represented by a number followed by "X" such as 30X coverage. [22]

Coverage focuses on ensuring that a substantial portion of the genome, or a specific target region such as the exome or a gene panel, has been adequately sequenced. It is typically expressed as a percentage. For instance, if 95% of the intended target region has been sequenced 30X, the coverage at 30X would be expressed as "95% coverage." However, several factors can influence coverage, including DNA sample quality, sequencing biases, and complexities within the genome, such as regions with high GC content or repetitive sequences, which can be challenging to sequence effectively. [39]

Coverage is a fundamental metric in bioinformatics, as it directly impacts the reliability of detecting genetic variants. Higher coverage means that more sequencing reads overlap each base, leading to improved sensitivity and accuracy in identifying variations. In contrast, lower coverage can increase the capacity to sequence more samples or cover larger genomic areas at similar costs, although it might compromise the confidence in variant detection. [22]

Different types of sequencing require varying levels of coverage. For Whole Genome Sequencing (WES), coverage between 30X and 50X is generally recommended, while Whole Exome Sequencing (WES) typically calls for 100X coverage. In the case of targeted gene panels, the coverage is often much greater-exceeding 500X in some cases-to ensure a high level of confidence in variant detection, particularly for somatic mutations, where greater read depth is crucial for accurate identification. [22]

Sequencing/Read Depth or Depth of Coverage

Sequencing depth, also referred to as read depth, denotes the number of times a specific nucleotide in the DNA sequence is read during the sequencing process. It provides an indication of how thoroughly each base has been examined. The more times a nucleotide is read, the more confidence researchers can have in the accuracy of the base call, as a greater number of reads help mitigate the risk of sequencing errors and minimize noise. For instance, if a particular nucleotide is read 30 times, the sequencing depth at that location would be expressed as 30X. [39]

The primary advantage of higher sequencing depth lies in its ability to enhance the precision of variant detection at specific genomic positions. This is particularly valuable when investigating rare genetic variants or when analyzing samples with significant heterogeneity, such as tumor tissues. [39]

The Figure 1.8 illustrates the relationship between coverage and depth of coverage in a gene. In this example, approximately 10% of the gene is depicted as not covered, and the depth reaches up to 9X. This visualization highlights the importance of both coverage and read depth in ensuring the reliability of sequencing data.

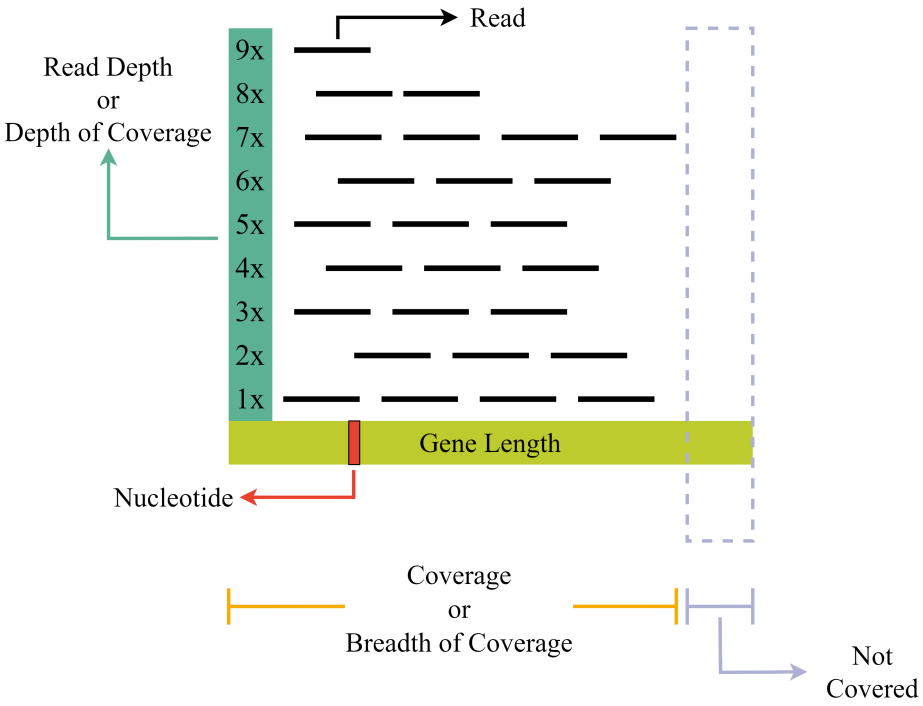


Figure 1.8: Scheme related to Coverage/ Breadth of Coverage and Read Depth/ Depth of Coverage in a gene. In this case, approximately 10% of the gene is depicted as not covered, and the depth reaches up to 9x. Adapted from [40]

Given the critical importance of validation in next-generation sequencing (NGS), as previously demonstrated, this thesis focuses on the development of the following user-friendly software solution designed to calculate and ensure the accuracy of key metrics such as average read depth and coverage. These metrics are essential for evaluating the quality and reliability of sequencing data. The software presented here not only automates these calculations but also provides a streamlined and accessible interface, making it easier for users to perform NGS validation processes with consistency and precision.

CHAPTER 2

Software development process

"I've learned a painful lesson, that for small programs dynamic typing is great. For large programs, you have to have a more disciplined approach. And it helps if the language actually gives you that discipline, rather than telling you, 'Well, you can do whatever you want.'" - Guido van Rossum, Python's creator

Falar de Metodologia Agile

2.1 Requirements

This section presents the requirements for the developed software, categorized into functional and non-functional requirements.

2.1.1 Functional Requirements

Functional requirements specify the functionalities that end-users require as essential components of the system. They describe the system's behavior in terms of inputs, operations to be performed, and expected outcomes. These requirements are directly observable by users in the final product. [41] The main functional requirements of the developed software are:

FR1 - Collection of BAM/CRAM Files for Analysis

The software must enable the collection of BAM/CRAM sequencing files stored on the company's servers.

FR2 - Calculation of Depth of Coverage/Read Depth and Coverage/Breadth of Coverage

The software must facilitate the calculation of Depth of Coverage and Coverage/Breadth of Coverage for the collected BAM/CRAM files. Users should be able to configure analysis parameters, such as selecting regions of interest within the exome. The analysis should be based on a universal BED file containing exon coordinates. The analysis should use

bioinformatics tools like SAMtools, which returns a .depth file with results that must be processed by the software to obtain the desired metrics.

FR3 - Graphical User Interface

The software must have a graphical user interface that allows users to interact with the system in an intuitive and efficient manner. The interface should be simple and easy to use, enabling users to collect BAM/CRAM files, configure analysis parameters, and view the obtained results. The interface should be developed using Streamlit, a Python web development tool that allows the creation of interactive web applications with minimal code. It should support user interaction through widgets such as buttons, text boxes, sliders, and others. The interface should allow filtering of results and exporting results to a CSV file.

2.1.2 Non-Functional Requirements

Non-functional requirements refer to the quality attributes of the software, such as performance, usability, security, and scalability. These requirements are crucial to ensure that the software operates efficiently, is secure, and can be easily used by various user profiles. [41] The main non-functional requirements of the system are described as follows:

NFR1 - Usability

The software should be intuitive and user-friendly, enabling even users with no technical background to interact with the system efficiently. The graphical user interface should be simple and clear, with straightforward instructions on how to use the system. It should allow users to collect BAM/CRAM files, configure analysis parameters, and view results quickly and easily. Clear and informative error messages should be provided in case of task execution failures, along with instructions for resolving issues.

NFR2 - Performance

The software must be optimized to process large volumes of sequencing data, ensuring that the calculation of Depth of Coverage/Read Depth and Coverage/Breadth of Coverage occurs within a reasonable time frame. The possibility of parallelizing calculations should be explored, utilizing multicore or distributed resources to accelerate data processing.

NFR3 - Scalability

The system must be scalable, capable of handling significant increases in data volume or the number of users without compromising performance. This includes the ability to leverage cloud services such as AWS S3. The software should be designed to allow the addition of new modules or functionalities without the need to rewrite the core code, ensuring flexibility for future evolution of the system.

NFR4 - Security and Data Privacy

The software must protect sensitive data, especially patient-related data, in compliance with regulations such as General Data Protection Regulation (GDPR). Temporary data used

during processing must be properly deleted after analysis, ensuring data privacy and security. System access should be controlled through authentication and authorization, ensuring that only authorized users can interact with the system. A logging system should be implemented to record user activities and monitor system usage.

NFR5 - Portability and Compatibility

The software should be implemented on Windows but must ensure access to a Linux environment using WSL. It should guarantee easy integration with tools like Samtools in the WSL environment without requiring advanced configuration by the end user.

NFR6 - Maintainability

The software code must be modular and well-documented, facilitating easy maintenance and extension of the system in the future. Best development practices, such as version control (Git), should be applied, ensuring that the code can be managed efficiently over time. Software updates should be simple to implement, and developer documentation should include clear instructions on how to add new functionalities or adjust existing behaviors.

With clear definitions of functional and non-functional requirements, the development of the software was structured efficiently, ensuring it meets both technical expectations and operational needs of the company and end-users. Considering these requirements throughout the development cycle was crucial for the success of the project.

2.2 System Design and Architecture

The developed software is based on a web interface accessible through a browser, using the Streamlit library to build the application. The interface is composed of several interactive widgets that allow the user to configure and execute different types of genomic analysis: Single Gene, Gene Panel, or Exome.

2.2.1 User Workflow

Initially, the user must select the type of analysis they wish to perform. Depending on the selection, the processing flow adapts to optimize both the user experience and the efficiency of the necessary calculations.

Single Gene Analysis

The user uploads a BAM or CRAM file, containing the sequencing data. Additionally, they automatically receive a universal BED file corresponding to the selected genome assembly. The user then chooses the gene of interest and may specify the exon(s) to be analyzed. SAMTOOLS is invoked to generate a DEPTH file, which contains information about the read depth at each genomic position. This file is processed by a Python script that calculates two key metrics: depth of coverage and breadth of coverage. The results are displayed to the user through a report in the Streamlit interface or can be exported as a CSV file.

Gene Panel Analysis

Similar to the Single Gene analysis, the user uploads a BAM/CRAM file and receives the universal BED file. However, in this case, the user selects the gene panel for the analysis instead of a single gene. The corresponding BED file is processed by SAMTOOLS to generate the DEPTH file, which is then used to calculate the same coverage metrics. The result visualization and export process follow the same steps as in the Single Gene analysis.

Exome Analysis

In Exome analysis, the entire exome is used for the analysis, without requiring the selection of a specific region. The user uploads the BAM/CRAM file, and the universal BED file corresponding to the exome is utilized. As in the other modes, SAMTOOLS generates the DEPTH file, which is processed to calculate the coverage metrics. The results are displayed or exported in the same manner.

The modular design of the software allows for seamless integration between various components, namely Streamlit for the interface, SAMTOOLS for processing sequencing data, and Python scripts for calculating coverage metrics. Each of these steps is interconnected to provide the user with quick, accurate, and real-time results. The software's architecture is designed to be easily scalable and adaptable to different types of genomic analysis, supporting both focused and large-scale analyses (such as whole exome).

The Figure 2.1 shows the scheme of the software architecture, highlighting the main components.

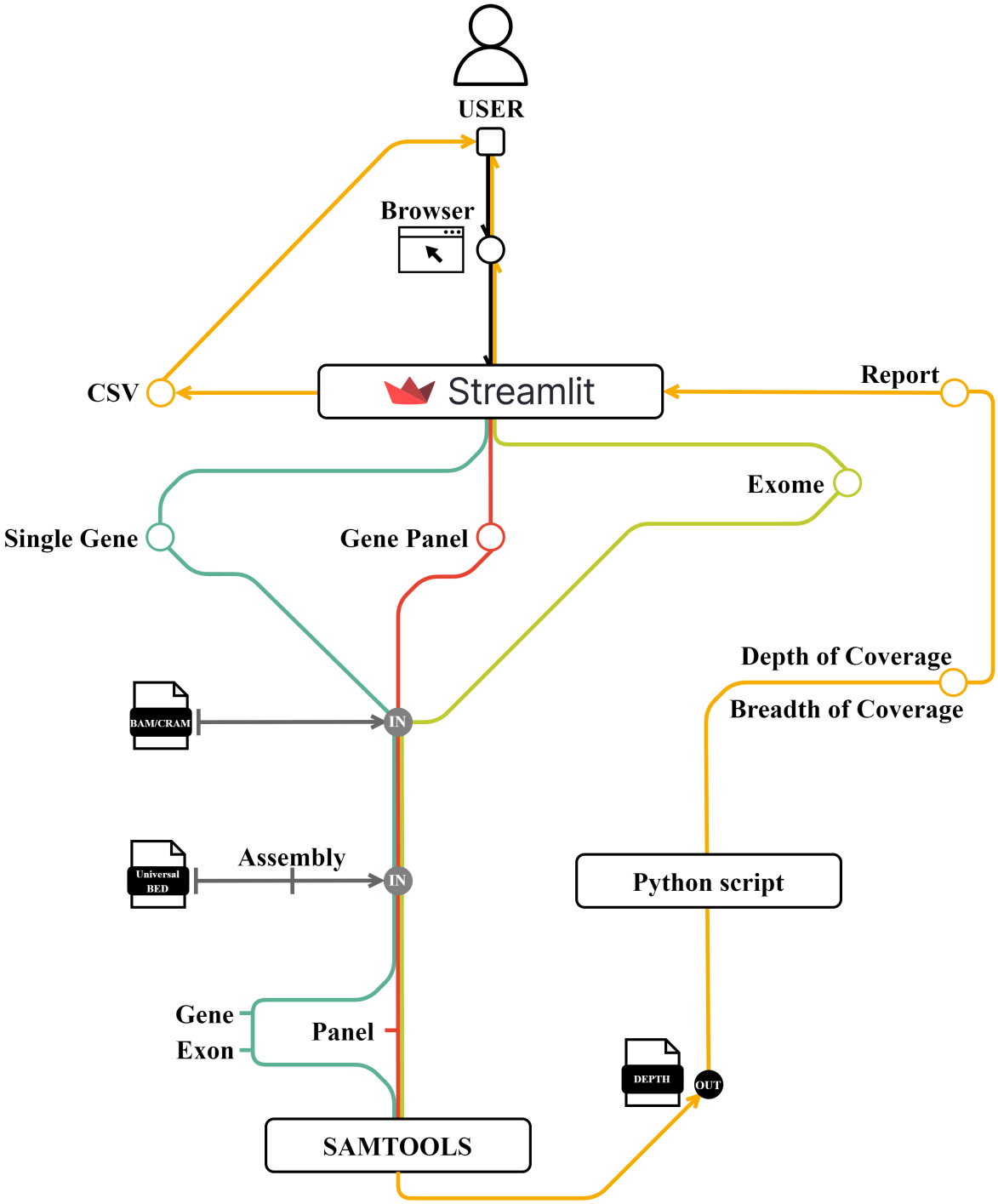


Figure 2.1: Scheme of the software architecture.

2.3 Development

The development of the software followed an iterative and structured approach, with each stage focusing on expanding its functionality while ensuring stability and performance. The process began with the implementation of the Single Gene analysis, which served as the foundation for the project. This initial version was designed to be stable and functional,

enabling users to upload BAM or CRAM files and analyze specific genes and exons with precision.

Once the Single Gene functionality was fully operational, the next phase involved adding support for Gene Panel analysis. This required adjusting the existing framework to handle a broader set of genes while maintaining the same level of accuracy and efficiency. A stable and functional version was again achieved, providing users with the ability to select and analyze predefined gene panels.

Finally, the software was further enhanced to include Exome analysis, allowing for the comprehensive examination of the entire exome. This functionality was integrated without compromising the stability or performance of the system, and once again, a stable and functional version was built.

Throughout the development, various components and tools were employed to ensure the software met the required performance standards and provided an intuitive user experience. The careful progression from Single Gene to Gene Panel, and finally to Exome analysis, highlights the modularity and scalability of the software, as well as the emphasis placed on testing and validation at each stage.

The following sections provide an overview of the key tools and technologies used in the development of the whole software, including all the stages of the project.

2.3.1 Environment preparation

Windows Subsystem for Linux (WSL)

On Windows, developers have access to both the Windows and Linux environments, thanks to the Windows Subsystem for Linux (WSL). With WSL, it is possible to install different Linux distributions, such as Ubuntu, OpenSUSE, Kali, Debian, Arch Linux, among others. This allows Linux applications, utilities and command-line tools to be used directly in Windows, without the need to modify the operating system, resort to virtual machines or dual boot. [42]

In the context of the development of this tool, the need to install WSL was driven mainly due to the scenario in which many essential tools and software for bioinformatics are designed to work in Linux environments. For this reason, we followed the set of steps recommended on the Microsoft website to configure this environment (Build 19041 or higher). [42]

Anaconda and Conda

After installing WSL, the installation step of Anaconda followed, a platform for data science in Python/R that includes conda, a package and environment manager, making it easier for users to manage a collection of more than 7,500 open source packages. [43]

In the case of the creation of the metrics analysis tool, this step was fundamental to allow the installation and maintenance of all the packages and dependencies necessary for the operation of the software. By creating a conda environment, it was possible to ensure that all installed tools work independently without conflicts between versions and packages, thus ensuring the reproducibility of the created software. [44]

Following the documentation provided by Anaconda, the installation and creation of the conda environment was carried out. [45]

Additionally, all the dependencies of the attached x-list were installed within the created environment. This installation was carried out by installing package by package, however, an environment.yaml file was made available that allows the bulk installation [46] of all dependencies on the versions compatible with the software.

Git and GitHub

GitHub is a platform that allows users to store, share, and collaborate on code writing with others. [47]

Its operation is based on repositories managed by Git, a version control system that tracks all changes made by one or more users in a project. [47]

When files are uploaded to GitHub, they become part of the created repository. Any change (commit) to any file is automatically tracked. These changes, made locally, are usually synchronized continuously by pushing the committed changes. Similarly, any changes made locally by another user and synchronized on GitHub can be retrieved by making a pull request. [47]

Thus, by using the documentation of Git and GitHub, this practice was implemented, which not only ensures that each version of the created software is recorded—guaranteeing that the work is not lost and allowing for version rollback in case of bugs—but also ensures that all software produced is reproducible and available for deployment by any user. [47]

2.3.2 Streamlit

2.3.3 SAMtools

SAMtools is an essential tool for the manipulation and analysis of DNA sequencing data. First released in 2009, it allows for converting, manipulating, sorting, querying, calculating statistics, calling variants, and analyzing sequencing data in SAM, BAM, and CRAM formats. [48]

Among the many functionalities of SAMtools, the most notable are its ability to convert formats, manipulate and index files, visualize and export data, and calculate statistics, such as "depth," which served as the basis for the tool created. [48]

In this case, a Python function was developed to generate a .depth file with the desired metrics, using SAMtools depth.

```python
import subprocess

def depth(bam_path, bed_path, depth_path, gene_selection=None, exon_selection=None):
    """
    Calculate the depth of coverage for specific exons of genes in a BAM file using samtools.
s
    Args:
        bam_path (str): Path to the BAM file.
        bed_path (str): Path to the Universal BED file containing exon coordinates.
        depth_path (str): Path to save the depth output.
        gene_selection (list or None): List of gene names to include in the depth calculation.
                                       If None, all genes will be included.
        exon_selection (list or None): List of exon numbers to include in the depth calculation.
                                       If None, all exons will be included.

    Returns:
        None
    """

    gene_filter = ','.join(map(str, gene_selection)) if gene_selection else ''
    exon_filter = ','.join(map(str, exon_selection)) if exon_selection else ''

    # Construct awk command to filter based on gene and exon selection
    awk_command = (f'awk -v gene_filter={gene_filter} -v exon_filter={exon_filter} '
                   f'\'{{split(exon_filter, arr, ","); '
                   f'if (($4 == gene_filter || gene_filter == "") && '
                   f'("" in arr || $5 == arr[1])) {{sub(/^chr/, "", $1); print}}}}\' {bed_path}')

    # Construct samtools command to calculate depth
    samtools_command = f'samtools depth -b - {bam_path} > {depth_path}'

    # Run the commands using subprocess
    try:
        subprocess.run(f'{awk_command} | {samtools_command}', shell=True, check=True)
    except subprocess.CalledProcessError as e:
        print(f"Error occurred: {e}")
```

Code 1: Python function to calculate depth of coverage using samtools and awk.

2.3.4 Python script for metrics calculation

2.4 Test and validation

2.5 Feedback and Iteration???

2.6 Optimization

2.7 Deployment

2.8 Impact on the company

CHAPTER 3

Results

"The only source of knowledge is experience." - Albert Einstein

CHAPTER 4

Additional activities during the internship

"The only source of knowledge is experience." - Albert Einstein

CHAPTER 5

Discussion

"The only source of knowledge is experience." - Albert Einstein

5.1 SWOT Analysis

During the development of the genomic analysis software, a SWOT analysis was performed to assess its strategic positioning. The analysis identified internal strengths, weaknesses, and external opportunities and threats within bioinformatics and genomics. Figure 5.1 summarize the key findings, outlining crucial elements that guided the software's development and deployment.

S
Strengths

Advanced bioinformatics and genomics concepts were applied in the development of an independent software for essential genomic analysis metrics, such as vertical and horizontal sequencing depth, ensuring compliance with NGS guidelines.

The use of advanced technologies like WSL, Anaconda, Conda, Git, GitHub, Streamlit, and Python ensures robust and efficient development. The software features a user-friendly interface, making it accessible to all users, including those without bioinformatics expertise.

Git and GitHub enable code sharing, ensuring traceability and reproducibility, while Anaconda and Conda create an isolated and reproducible environment for managing packages and dependencies.

W
Weaknesses

The requirement for WSL and a Linux environment may limit accessibility for users unfamiliar with these technologies.

While the graphical interface is intuitive, it may not suffice for advanced users who prefer command-line tools for analysis.

The initial setup of the development environment involves multiple steps and requires a solid understanding of the tools, which can be a barrier for less experienced users.

Comparing and validating metrics obtained by the new software with other tools is essential, but ensuring consistent and comparable results can be challenging.

O
Opportunities

The software can be expanded to include other important genomic analysis metrics.

Unilabs could promote internal use, establishing it as the standard tool for metric analysis.

Releasing the software as an open-source solution could attract a community of developers and users to contribute to its continuous improvement.

The rapid evolution of sequencing technologies and data analysis offers opportunities to continuously enhance and update the tool with new advancements and algorithms.

T
Threats

Several established tools and platforms in the market offer similar functionalities, which may hinder the adoption of the new software.

Competing with robust commercial software that offers extensive support can be challenging.

The rapid evolution of sequencing technologies and data analysis could render some features obsolete or require frequent updates.

Dependence on third-party software and libraries that may discontinue or significantly change their APIs presents a risk.

Increasing regulation around genomic data and genetic testing may impose additional challenges for the use and distribution of the software.

Figure 5.1: SWOT Analysis.

CHAPTER 6

Final remarks

"The only source of knowledge is experience." - Albert Einstein

Bibliography

[1] P. Bourque, R. E. Fairley, and I. C. Society, *Guide to the Software Engineering Body of Knowledge (SWEBOK(R)): Version 3.0*, 3rd. Washington, DC, USA: IEEE Computer Society Press, 2014, ISBN: 0769551661.

[2] *Unilabs - sobre*. [Online]. Available: https://www.unilabs.pt/pt/a-unilabs/sobre-nos/unilabs-portugal.

[3] *Unilabs - genética médica*. [Online]. Available: https://www.unilabs.pt/pt/servicos/especialidades-medicos/genetica-medica/sobre.

[4] N. H. G. R. I. (NHGRI), *Genetic timeline*. [Online]. Available: https://www.genome.gov/Pages/Education/GeneticTimeline.pdf.

[5] J. Gayon, «De mendel à l'épigénétique: Histoire de la génétique», *Comptes Rendus - Biologies*, vol. 339, pp. 225–230, 7-8 Jul. 2016, ISSN: 17683238. DOI: 10.1016/j.crvi.2016.05.009.

[6] F. S. Collins and L. ; Fink, *The human genome project*, 1995.

[7] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier, *A programmable dual-rna-guided dna endonuclease in adaptive bacterial immunity*. [Online]. Available: https://www.science.org.

[8] M. A. Gutierrez-Reinoso, P. M. Aponte, and M. Garcia-Herreros, *Genomic analysis, progress and future perspectives in dairy cattle selection: A review*, Mar. 2021. DOI: 10.3390/ani11030599.

[9] N. H. G. R. Institute, *Genetics vs. genomics fact sheet*, Sep. 2018. [Online]. Available: https://www.genome.gov/about-genomics/fact-sheets/Genetics-vs-Genomics.

[10] T. J. Laboratory, *Genetics vs. genomics*, Feb. 2017. [Online]. Available: https://www.jax.org/personalized-medicine/precision-medicine-and-you/genetics-vs-genomics#.

[11] S. Minchin and J. Lodge, *Understanding biochemistry: Structure and function of nucleic acids*, 2019. DOI: 10.1042/EBC20180038.

[12] M. KGaA, *Sanger sequencing steps and method*. [Online]. Available: https://www.sigmaaldrich.com/PT/en/technical-documents/protocol/genomics/sequencing/sanger-sequencing.

[13] S. M. Group, *Crick and watson's dna molecular model*, 1977. [Online]. Available: https://collection.sciencemuseumgroup.org.uk/objects/co146411/crick-and-watsons-dna-molecular-model..

[14] B. Maddox, *The double helix and the 'wronged heroine'*, Jan. 2003. DOI: 10.1038/nature01399.

[15] L. Merrick, A. Campbell, D. Muenchrath, and S. Fei., «Mutations and variation», in W. Suza and K. Lamkey, Eds. Iowa State University Digital Press, Mar. 2016. DOI: 10.31274/isudp.2023.130.

[16] C. A. for Drugs and T. in Health, *Next generation dna sequencing: A review of the cost effectiveness and guidelines*, Feb. 2014.

[17] H. L. Rehm, S. J. Bale, P. Bayrak-Toydemir, *et al.*, «Acmg clinical laboratory standards for next-generation sequencing», *Genetics in Medicine*, vol. 15, pp. 733–747, 9 Sep. 2013, ISSN: 10983600. DOI: 10.1038/gim.2013.92.

[18] J. Majewski, J. Schwartzentruber, E. Lalonde, A. Montpetit, and N. Jabado, «What can exome sequencing do for you?», *Journal of Medical Genetics*, vol. 48, no. 9, pp. 580–589, 2011, ISSN: 0022-2593.

DOI: 10.1136/jmedgenet-2011-100223. eprint: https://jmg.bmj.com/content/48/9/580.full.pdf. [Online]. Available: https://jmg.bmj.com/content/48/9/580.

[19] N. J. Schork, «Genetic parts to a preventive medicine whole», *Genome Medicine*, vol. 5, no. 6, p. 54, Jun. 2013, ISSN: 1756-994X. DOI: 10.1186/gm458. [Online]. Available: https://doi.org/10.1186/gm458.

[20] T. C. GLENN, «Field guide to next-generation dna sequencers», *Molecular Ecology Resources*, vol. 11, no. 5, pp. 759–769, 2011. DOI: https://doi.org/10.1111/j.1755-0998.2011.03024.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1755-0998.2011.03024.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1755-0998.2011.03024.x.

[21] Illumina, *Novaseq 6000 sequencing system guide*, Feb. 2023. [Online]. Available: https://emea.support.illumina.com/downloads/novaseq-6000-system-guide-1000000019358.html.

[22] N. B. Larson, A. L. Oberg, A. A. Adjei, and L. Wang, *A clinician's guide to bioinformatics for next-generation sequencing*, Feb. 2023. DOI: 10.1016/j.jtho.2022.11.006.

[23] Wikipedia, *Fastq format*, Jun. 2024. [Online]. Available: https://en.wikipedia.org/wiki/FASTQ_format.

[24] S. Roy, C. Coldren, A. Karunamurthy, *et al.*, *Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: A joint recommendation of the association for molecular pathology and the college of american pathologists*, Jan. 2018. DOI: 10.1016/j.jmoldx.2017.11.003.

[25] M. Bioinformatics, *Structural variant calling - long read data.* [Online]. Available: https://www.melbournebioinformatics.org.au/tutorials/tutorials/longread_sv_calling/longread_sv_calling/.

[26] R. Somak, *Next-generation sequencing bioinformatics pipelines*, Mar. 2020. [Online]. Available: https://www.myadlm.org/cln/Articles/2020/March/Next-Generation-Sequencing-Bioinformatics-Pipelines.

[27] A. M. Kanzi, J. E. San, B. Chimukangara, *et al.*, «Next generation sequencing and bioinformatics analysis of family genetic inheritance», *Frontiers in Genetics*, vol. 11, 2020, ISSN: 1664-8021. DOI: 10.3389/fgene.2020.544162. [Online]. Available: https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2020.544162.

[28] N. M. Ioannidis, J. H. Rothstein, V. Pejaver, *et al.*, «Revel: An ensemble method for predicting the pathogenicity of rare missense variants», *American Journal of Human Genetics*, vol. 99, pp. 877–885, 4 Oct. 2016, ISSN: 15376605. DOI: 10.1016/j.ajhg.2016.08.016.

[29] M. Schubach, T. Maass, L. Nazaretyan, S. Roner, and M. Kircher, «Cadd v1.7: Using protein language models, regulatory cnns and other nucleotide-level scores to improve genome-wide variant predictions», *Nucleic Acids Research*, vol. 52, pp. D1143–D1154, D1 Jan. 2024, ISSN: 13624962. DOI: 10.1093/nar/gkad989.

[30] Y. Fu, Z. Liu, S. Lou, *et al.*, «Funseq2: A framework for prioritizing noncoding regulatory variants in cancer», *Genome biology*, vol. 15, p. 480, 10 2014, ISSN: 1474760X. DOI: 10.1186/s13059-014-0480-5.

[31] A. P. Boyle, E. L. Hong, M. Hariharan, *et al.*, «Annotation of functional variation in personal genomes using regulomedb», *Genome Research*, vol. 22, pp. 1790–1797, 9 Sep. 2012, ISSN: 10889051. DOI: 10.1101/gr.137323.112.

[32] I. Dunham, A. Kundaje, S. F. Aldred, *et al.*, «An integrated encyclopedia of dna elements in the human genome», *Nature*, vol. 489, pp. 57–74, 7414 Sep. 2012, ISSN: 14764687. DOI: 10.1038/nature11247.

[33] R. E. Consortium, A. Kundaje, W. Meuleman, *et al.*, «Integrative analysis of 111 reference human epigenomes», *Nature*, vol. 518, pp. 317–329, 7539 Feb. 2015, ISSN: 14764687. DOI: 10.1038/nature14248.

[34] S. Chen, L. C. Francioli, J. K. Goodrich, *et al.*, «A genomic mutational constraint map using variation in 76,156 human genomes», *Nature*, vol. 625, pp. 92–100, 7993 Jan. 2024, ISSN: 14764687. DOI: 10.1038/s41586-023-06045-0.

[35] M. J. Landrum, J. M. Lee, M. Benson, *et al.*, «Clinvar: Improving access to variant interpretations and supporting evidence», *Nucleic Acids Research*, vol. 46, pp. D1062–D1067, D1 Jan. 2018, ISSN: 13624962. DOI: 10.1093/nar/gkx1153.

[36] P. D. Stenson, M. Mort, E. V. Ball, *et al.*, *The human gene mutation database (hgmd®): Optimizing its use in a clinical diagnostic or research setting*, Oct. 2020. DOI: 10.1007/s00439-020-02199-3.

[37] J. G. Tate, S. Bamford, H. C. Jubb, *et al.*, «Cosmic: The catalogue of somatic mutations in cancer», *Nucleic Acids Research*, vol. 47, pp. D941–D947, D1 Jan. 2019, ISSN: 13624962. DOI: 10.1093/nar/gky1015.

[38] K. Wang, M. Li, and H. Hakonarson, «Annovar: Functional annotation of genetic variants from high-throughput sequencing data», *Nucleic Acids Research*, vol. 38, 16 Jul. 2010, ISSN: 03051048. DOI: 10.1093/nar/gkq603.

[39] 3billion, *Sequencing depth vs coverage*, Aug. 2023. [Online]. Available: https://3billion.io/blog/sequencing-depth-vs-coverage.

[40] MedGenome, *Understanding gene coverage and read depth*, Apr. 2020.

[41] G. for Geeks, *Functional vs non functional requirements*, Jun. 2024. [Online]. Available: https://www.geeksforgeeks.org/functional-vs-non-functional-requirements/.

[42] M. 2024, *How to install linux on windows with wsl*. [Online]. Available: https://learn.microsoft.com/en-us/windows/wsl/install.

[43] A. Inc, *Anaconda*. [Online]. Available: https://docs.anaconda.com/free/.

[44] J. Leidel, *12 reasons to choose conda*, Sep. 2023. [Online]. Available: https://www.anaconda.com/blog/12-reasons-to-choose-conda.

[45] A. Inc, *Anaconda - installing on windows*. [Online]. Available: https://docs.anaconda.com/free/anaconda/install/windows/.

[46] A. Inc, *Anaconda - managing environments*. [Online]. Available: https://docs.conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html#create-env-file-manually.

[47] G. Inc, «Git and github - get started», [Online]. Available: https://docs.github.com/pt/get-started/start-your-journey.

[48] P. Danecek, J. K. Bonfield, J. Liddle, *et al.*, «Twelve years of samtools and bcftools», *GigaScience*, vol. 10, 2 Feb. 2021, ISSN: 2047217X. DOI: 10.1093/gigascience/giab008.

[49] B. .-. Illumina, *Quality score encoding*, May 2024. [Online]. Available: https://help.basespace.illumina.com/files-used-by-basespace/quality-scores.

APPENDIX A

Additional content

Table A.1: Unilabs test catalog

Test Catalog
WES
Next Generation Sequencing
Sanger Sequencing
Comparative Genomic Hybridization (aCGH)
Karyotyping
Fluorescence In Situ Hybridization (FISH)
QF-PCR, qPCR, RT-PCR
Fragment and Expansion Analysis
Multiplex Ligation-Dependent Probe Amplification (MLPA)
Single Gene Analysis
Variant Analysis
Cytogenetics
NIPT | Tomorrow

Table A.2: Quality score encoding. Adapter from [49]

Symbol	ASCII Code	Q-Score	P-Error
!	33	0	1,00000
"	34	1	0,79433
#	35	2	0,63096
$	36	3	0,50119
%	37	4	0,39811
&	38	5	0,31623
’	39	6	0,25119
(	40	7	0,19953
)	41	8	0,15849
*	42	9	0,12589
+	43	10	0,10000
,	44	11	0,07943
-	45	12	0,06310
.	46	13	0,05012
/	47	14	0,03981
0	48	15	0,03162
1	49	16	0,02512
2	50	17	0,01995
3	51	18	0,01585
4	52	19	0,01259
5	53	20	0,01000
6	54	21	0,00794
7	55	22	0,00631
8	56	23	0,00501
9	57	24	0,00398
:	58	25	0,00316
;	59	26	0,00251
<	60	27	0,00200
=	61	28	0,00158
>	62	29	0,00126
?	63	30	0,00100
@	64	31	0,00079
A	65	32	0,00063
B	66	33	0,00050
C	67	34	0,00040
D	68	35	0,00032
E	69	36	0,00025
F	70	37	0,00020
G	71	38	0,00016
H	72	39	0,00013
I	73	40	0,00010

Table A.3: Samtools - BED file documentation

Column	BED Field	Type	Regex or range	Brief description
1	chrom	String	[[a-zA-Z0-9_]1,255]	Chromosome name
2	chromStart	Int	[0, 2^{64} - 1]	Feature start position
3	chromEnd	Int	[0, 2^{64} - 1]	Feature end position
4	name	String	[\x20-\x7e]1,255	Feature description
5	score	Int	[0, 1000]	A numerical value
6	strand	String	[-+.]	Feature strand
7	thickStart	Int	[0, 2^{64} - 1]	Thick start position
8	thickEnd	Int	[0, 2^{64} - 1]	Thick end position
9	itemRgb	Int,Int,Int	([0, 255], [0, 255], [0, 255]) | 0	Display color
10	blockCount	Int	[0, chromEnd - chromStart]	Number of blocks
11	blockSizes	List[Int]	([[0-9]]+,)blockCount-1[[0-9]]+,?	Block sizes
12	blockStarts	List[Int]	([[0-9]]+,)blockCount-1[[0-9]]+,?	Block start positions